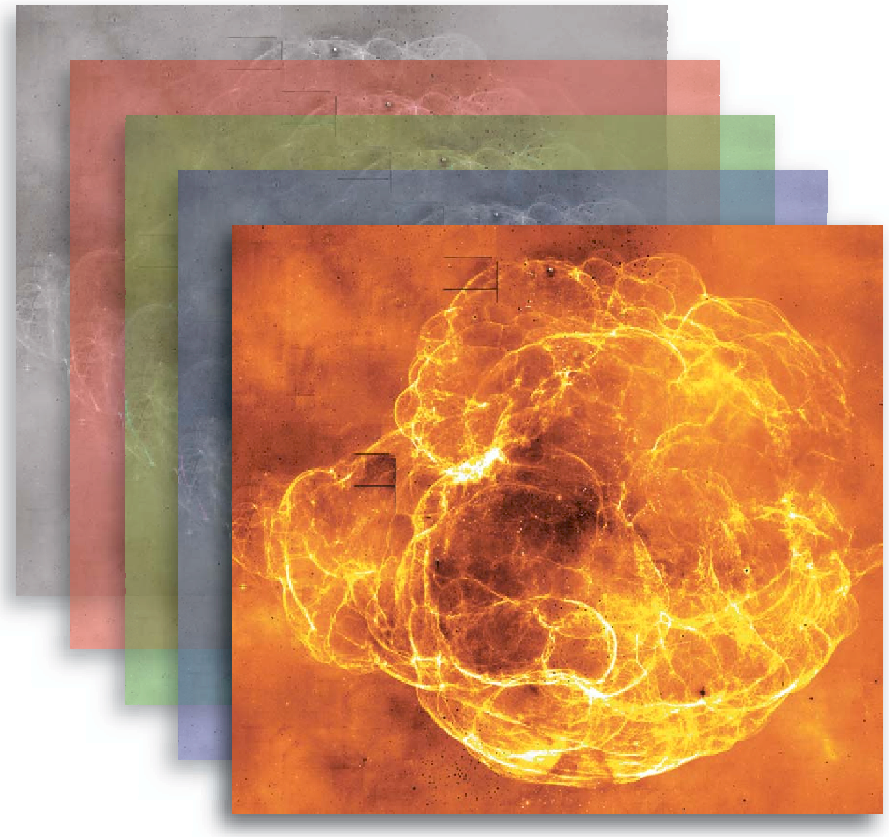
Cambridge Astronomical Survey Unit



Report and Future Programme (2008-2013)

Cambridge Astronomical Survey Unit (CASU) Report and Future Program (2008–2013)

June, 2007

Abstract

Observational astronomy in the UK is in a strong position with the advent of the pioneering near-infrared (NIR) survey facilities of WFCAM and VISTA and the opportunities for detailed followup provided by membership of ESO. Significant advances in astronomy have always relied heavily on surveys of the sky from radio, through optical to X-ray wavelengths and the new era of deep NIR surveys is the latest stage in this progression. It is also one where the UK has both a substantial lead and the requisite expertise in the Cambridge and Edinburgh survey units to exploit this advance in wide field astronomy. Recognising the benefits of this, ESO are now also heavily committed to public surveys through the VST and VISTA telescopes and to wider exploitation of expertly processed and archived science data products.

Members of the Cambridge Astronomy Survey Unit (CASU) have played a leading role in survey astronomy, not only by pioneering techniques to optimally extract knowledge from survey data, but also by taking a proactive role in exploiting this information to produce world-leading research. This synergy and feedback between data processing and exploitation is crucial. It has been a deliberate strategy that has provided the main strength and motivation of the group.

In the modern era wide-field digital survey cameras produce enormous volumes of data that are way beyond the resource capacity and analysis skills of non-specialist astronomers. Systematic pipeline processing, calibration and legacy curation of observational data are a fundamental requirement of an end-to-end integrated observing strategy and a crucial component of a global Virtual Observatory. The CASU facility has been developed to allow an optimal ergonomic solution to this avalanche of data, through access to multi-TB data storage systems and expert pipeline processing systems. Continuing development of the CASU processing and analysis pipelines will not only benefit the UK astronomy community now, but will also be relevant in the era of Extremely Large Telescopes and the radio Square Kilometer Array, by developing the infrastructure to analyse the data from these facilities.

This rolling grant proposal builds on the tremendous advances already made and requests funding for CASU for the period 2008-2013 for the following activities: continued operation and maintenance of the WFCAM data processing and calibration pipeline; advanced development, enhancement, and operation of the UK VISTA pipelines; maintenance and upgrades for the VDFS ESO pipeline deliverables; operational support and pipeline processing for the UK-led VST public surveys in the southern hemisphere and the equivalent precursor surveys in the northern hemisphere; support to provide a range of science data products and services (utilising the AstroGrid infrastructure) as a UK contribution to the Virtual Observatory initiative; continuation of support for maintaining and developing the UK ground-based telescope archives; and a modest research component for the research-active members of the group.

For completeness and context we also summarise the various CASU activities over the period of the current grant, which includes more than two years of operating a highly successful WFCAM pipeline. We also give a brief description of the progress made with the development of the VISTA data processing pipelines since these are also central to this grant application.

Contents

1 Introduction

The Cambridge Astronomical Survey Unit (CASU) is at the forefront of UK efforts to maximally exploit existing and upcoming survey-astronomy facilities such as WFCAM on UKIRT and the ESO VST and VISTA survey telescopes. These facilities dominate the landscape of ground-based survey work, offering almost limitless astronomical opportunities both in their own right and as essential components for the proper exploitation of the UK's access to current 8–10m class facilities such as ESO, Gemini, Subaru and Keck, and to the future design of surveys to exploit the next generation ELT facilities.

The requirements to fully capitalise on WFCAM and VISTA, in particular, are extremely challenging due to the huge volume of data (≈100 TB/year) that must be ingested, processed, calibrated, curated and served automatically to the science community (see figure 1). However, the potential benefit to UK astronomy is enormous.

Processing NIR survey data is technically much more demanding than the equivalent optical survey data. This stems from the fact that NIR detectors are inherently more unstable than their optical counterparts; the sky emission, roughly 100 times brighter than most objects of interest, varies in a complex spatial and temporal manner; and all of this is compounded by the large data volume that arises as a consequence of the bright background.

Over the past few years the survey aspects of frontier astronomy have changed dramatically and have become dominated by ground-based wide field mosaic CCD and NIR imaging systems. In combination with systematic surveys in other wavebands ranging from the radio *e.g.* HIPASS to the x-ray *e.g.* XMM this has lead to a paradigm shift in the way survey astronomy is perceived.

Few users can handle either the data volumes and complexities of processing and calibrating optical and NIR mosaic camera data at their home institutions, or the related effort required to cross-federate data taken in different wavebands. With the realisation that specialist expertise focussed at a few institutions is an effective and powerful ergonomic solution to this problem, user expectations have now changed. The benefits of end-to-end data flow systems that cater for the vast majority of astronomical science requirements by providing expertly processed, calibrated and quality-controlled science products are obvious. Expectations change quickly and now focus on science exploitation, effectively taking the processing and calibration system for granted.

Related to this, is the shift toward queue-scheduled observing which is an effective way of designing and controlling the data taking process. This all helps to make semi-automatic end-to-end data processing architectures and data curation possible and helps ensure efficient and optimal exploitation. Coupled with the cross-disciplinary nature of modern astronomy this has lead to ever increasing use of archival federated multiwaveband material, and the push, through GRID-enabled interlinked databases, toward a Global Virtual Observatory (VO).

In the VO context CASU works closely with the active AstroGrid VO group at the IoA, feeding requirements from the resource provider community to that project, and benefiting from early exposure to the AstroGrid VO infrastructure. Thus the Wide Field survey images at CASU were the first significant data collection to be made available through AstroGrid. CASU has also led the way in utilising AstroGrid acess technology to significantly reduce the cost of publishing new data sets.

The Cambridge Astronomy Survey Unit has been at the forefront of processing and archiving digital wide-field surveys for some time. We have used the experience gained from these earlier surveys (*e.g.* APM/SuperCOSMOS surveys, INT WFS and CIRSI) to design and build end-to-end processing systems for dealing with WFCAM and VISTA data as part of the VISTA data flow system (VDFS) project. In parallel with this development, we have also been developing and running pipeline processing systems for several other existing wide field mosaic cameras to enhance the pipeline components, to assess the problems of day-to-day running of operational pipelines and to explore the practicalities of controlling optional further processing stages driven from a survey quality control database.[1]

[1] see, for example, the WFC processed data interface at http://apm2.ast.cam.ac.uk/cgi-bin/wfs/dqc.cgi and the WFCAM pipeline

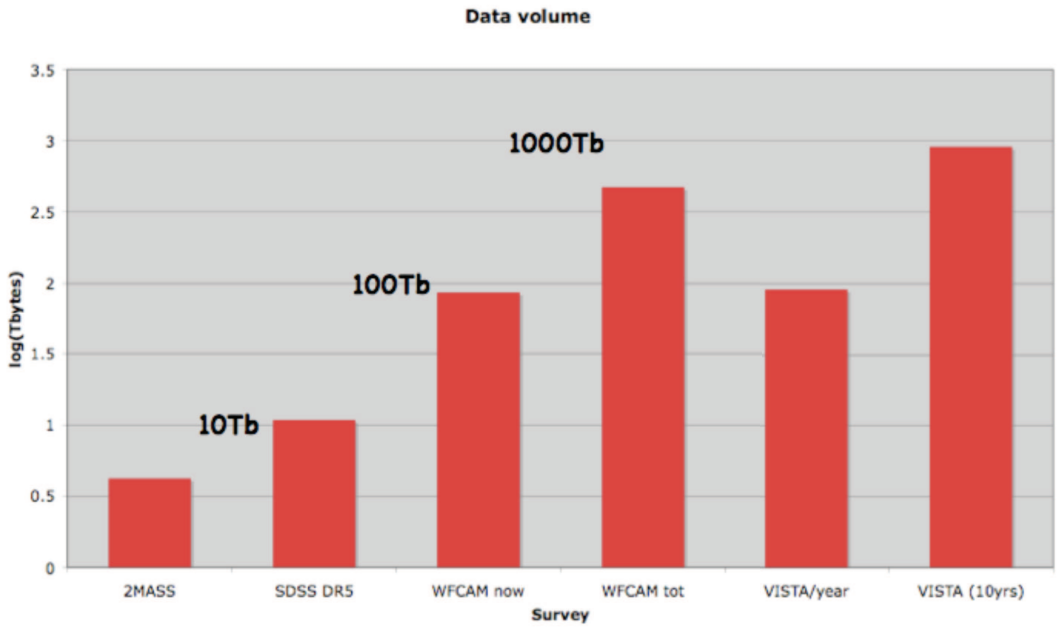


Figure 1: The huge growth in data volume from the WFCAM and VISTA surveys compared to current state-of-the-art surveys like 2MASS and SDSS.

2 Key Programme Aims

CASU was formed in October 1998 following negotiations between the University of Cambridge and PPARC during the closure of the Royal Greenwich Observatory. The original core activities of CASU were funded from a rolling grant and were concerned with continuing to support and develop ground- and space-based UK astronomy wide-fi eld survey activities that had existed within the RGO. Over the intervening years various aspects of these core activities are now based on project-specifi c funding such as the development of the VISTA and GAIA data flow systems (VDFS and GDFS). GAIA is outside the scope of this grant application[2] but the operational aspects of the VDFS are a central component.

The VDFS development grants (including a 6 month extension) cover the period to the end of March 2008 when it is anticipated that VISTA will begin on-sky survey science operations. The VDFS project includes various ESO deliverables in addition to prototyping the VISTA data flow development using WFCAM data. However, it does not cover operational costs, the inevitable further development and maintenance that will be necessary when real on-sky data becomes available, or the necessary feedback and interaction with external users.

In this grant we are seeking funding for the period 2008–2013 for the following key activities:

- **VISTA:** this activity includes two major complementary strands:
  - **VISTA Science Products:** advanced development, enhancement, maintenance and operation of the UK VISTA science processing pipelines for provision of calibrated science-ready data products to the UK community via the VISTA science archive including user liaison;
  - **VISTA Pipeline:** effort for maintenance and upgrades to software and documentation for the VDFS ESO pipeline deliverables, this includes the VISTA calibration pipeline software modules to be run

progress status and quality control web pages at http://casu.ast.cam.ac.uk/surveys-projects/wfcam/data-processing/.

[2]We note that there are strong synergies at CASU between the GAIA activities and those presented here, with a number of staff devoting expertise to, and partly funded by, the Gaia Data Flow System project. This cross transfer of skills and knowledge works to the benefi t of both activities.

at Garching within the CPL environment and the Paranal summit pipeline for near-time quality control monitoring run within the ESO VLT operating environment;

- **WFCAM:** the current CASU rolling grant includes management, development, operation and maintenance of all WFCAM-related data processing. We are seeking continued funding at the same level to cover further operational development as needed, upgrades in response to changing demands, maintenance and enhancements to the data processing and calibration system and pipelines, and general user liaison;

- **Science Service Centre:** the provision of key science data products to the UK community, representing a significant resource provider of content to the UK VO, operation and further development of pipeline processing and global calibration of data for science exploitation from optical wide field survey imaging systems, including UK-led VST public surveys in the southern hemisphere and the corresponding optical surveys (*e.g.* IPHAS[3]) in the northern hemisphere;

- **Data Archive Centre:** management, operation, maintenance and further development of ING, UKIRT and AAT on-line archives, in particular developing more efficient calibration schemes, progressing demand-driven on-the-fly processing of data from these archives, and liaising closely with the AstroGrid project to ensure grid-enabled deliverables;

- **Research** allocation for research-active staff within the group to ensure rapid exploitation of survey products and to facilitate collaborative ventures with external groups;

Progress on each of the main activity areas, including future plans where appropriate, are detailed in the following sections. We note that, unlike the typical 'research only' rolling grant programmes where PDRAs are allocated to single research themes, here the key CASU staff typically allocate their expertise across a number of the activity areas outlined here. The contribution of staff effort to each activity is noted in the relevant sections and summarised in the attached Appendices. Management of the programme is led by Irwin, assisted by Walton (science services and the Gaia/VO interfaces) and Bunclark (Vista operations).

3 Science Research: Highlights & Future Plans

Several staff involved with CASU are active research workers. We therefore request an annual allocation of 1.5 FTE/yr of research effort to ensure the continuing synergy between active research workers and the collective national support facilities and survey projects we are operating. This effort will be split among five named individuals (Irwin, Gonzalez-Solares, Riello, Hodgkin & Greimel – 0.3 FTE/yr each) for science research capitalising on survey science in the areas presented in the following sections.[4]

3.1 Near-field Cosmology in the Local Group

The discovery of the Sagittarius dwarf (Ibata, Gilmore, & Irwin, 1994, Nature, 370, 194), a satellite galaxy of the Milky Way caught in the act of tidal destruction, irrevocably changed our opinion of the local universe and brought modern cosmological predictions to the forefront of our Galactic backyard.

Since then Irwin has been involved in the discovery and analysis of numerous examples of new satellite galaxies and tidal streams and debris, in and around Local Group galaxies, including: great circles of carbon stars ripped from the Sagittarius Dwarf tracing out its previous orbits (*e.g.* Totten, Irwin & Whitelock 2000, MNRAS 314, 630); the discovery of another disrupting dwarf galaxy in Canis Major, this time embedded in the outer disk

[3]IPHAS is the INT WFC Photometric Hα Survey of the Northern Galactic Plane, $-5^{o} < b < +5^{o}$ imaged in narrow-band Hα, and in Sloan r' and i', at a spatial resolution of $\sim$1 arcsec to a 10σ magnitude limit $r' = 20$ (Drew et al. 2005, MNRAS 362, 753). This survey is due for completion during the 2007/2008 observing season.

[4]For this rolling grant submission we have been given permission to include a research component, as in previous CASU grants.

(Martin *et al.* 2004 MNRAS 348, 12) and quite plausibly the progenitor of the rings of material first seen in the Galactic anticentre direction in Monoceros (*e.g.* Ibata *et al.* 2003 MNRAS 340, 21); the discovery and analysis of giant streams and copius stellar substructure and new satellites galaxies in and around M31 (*e.g.* Ibata *et al.* 2007 astro-ph 0704-1318); and the discovery and analysis of several new satellite systems around the Milky Way (*e.g.* Belokurov *et al.* 2006 ApJ 654, 897).

Statistical studies of the ensemble properties are now also revealing for the first time the true extent and nature of the global surface brightness distributions, focussing on the major and minor axes and the giant stellar stream in M31, and the entire disk in M33. By combining directly measured surface brightness profiles of the inner regions with deep star counts of the outer parts we have access to a dynamic range in excess of 15 magnitudes. This has enabled us to trace the entire profile and its colour variation out to unprecedented depths (33rd magnitude) using a range of tracer populations (*e.g.* Irwin *et al.* 2005 ApJLett 628, 105).

Building on the INT WFC data we are half-way though a CFHT MegaCam survey to trace the outer halo of M31 all the way down the minor axis to M33, to further probe the region past the tip of the giant stream and to place limits on the "lumpiness" of outer galaxy halos, directly confronting CDM predictions. An interesting byproduct of these imaging surveys has been the discovery of many new outer halo globular clusters in M31 and the discovery of several representatives of a new population of extended luminous halo clusters which have no clear counterpart in the Milky Way Halo (Huxor *et al.* 2005 MNRAS 360, 1007).

Our recent discoveries in the outer regions of the Milky Way and M31 suggest that galaxy-galaxy interactions and the scars they leave are a ubiquitous feature of these regions. Ongoing projects include: extending these type of imaging surveys and studies on Subaru and the VLT, outward to nearby groups such as members of the Sculptor, Centaurus and M81 groups; and carrying out a detailed kinematic survey of tens of thousands of RGB and AGB stars in M31 (an Andromedan RAVE) using DEIMOS on Keck (Chapman *et al.* 2006 ApJ 653 255).

In the NIR, WFCAM, and soon VISTA, present a unique opportunity to use JHK photometry to study the AGB and brighter RGB populations for all the Local Group galaxies. These stars are excellent tracers of the structural, chemical and kinematic evolution of their intermediate-age stellar populations. The goals of complete maps of these populations across the entirety of M31, M33 and the Local Group dwarfs will allow an unprecedented census and study of the structure of their intermediate age components, essentially free of the (optically) crippling effects of internal extinction and crowding in their inner regions. This ongoing project has been the top-ranked WFCAM survey PATT-time proposal since data taking in earnest began in 2005.

In collaboration with Belokurov and Zucker at the IoA, Irwin has been actively and successfully engaged in using SDSS data to find new examples of faint nearby satellite stellar systems ranging in luminosity from the fainter end of the "classical" dSphs to absolute luminosities $M_V \approx$ -3, equivalent to one RGB star. This has resulted in the discovery and analysis of several new satellite galaxies around the Milky Way, culminating in the discovery of a rare example of a faint isolated stellar system with a significant amount of HI gas and evidence of recent star formation (*e.g.* Irwin *et al.* 2007 ApJLett 656, 13). The totality of these new discoveries has lead to an interesting dichotomy in the comparative properties of star cluster and dwarf galaxies, despite the almost complete overlap in luminosities (see figure 2 adapted from Belokurov *et al.* 2007 ApJ 654, 897).

Dwarf spheroidal galaxies (dSphs) play a central role in modern cosmological theories of galaxy formation, as the objects most likely to form first after reionisation and as the surviving building blocks of large galaxies such as our own. Understanding their formation, structure and evolution is therefore of vital importance to furthering our understanding of galaxy evolution.

In collaboration with Tolstoy and Helmi (Kaptyen Institute), Irwin has used CASU-processed ESO Wide Field Imager (WFI) data in conjunction with the VLT FLAMES spectrograph, to obtain abundance and kinematic measures for several hundred stars in each of the dSphs Fornax, Sculptor and Sextans and Carina. With this we have shown that there is direct evidence for the existence of kinematically and photometrically distinct "cores" in the Sculptor and Sextans dSphs (Tolstoy *et al.* 2004; Kleyna *et al.* 2004 MNRAS 354, 66). The origin of two spatially distinct components in these galaxies is still unknown, but suggests that star formation at early times

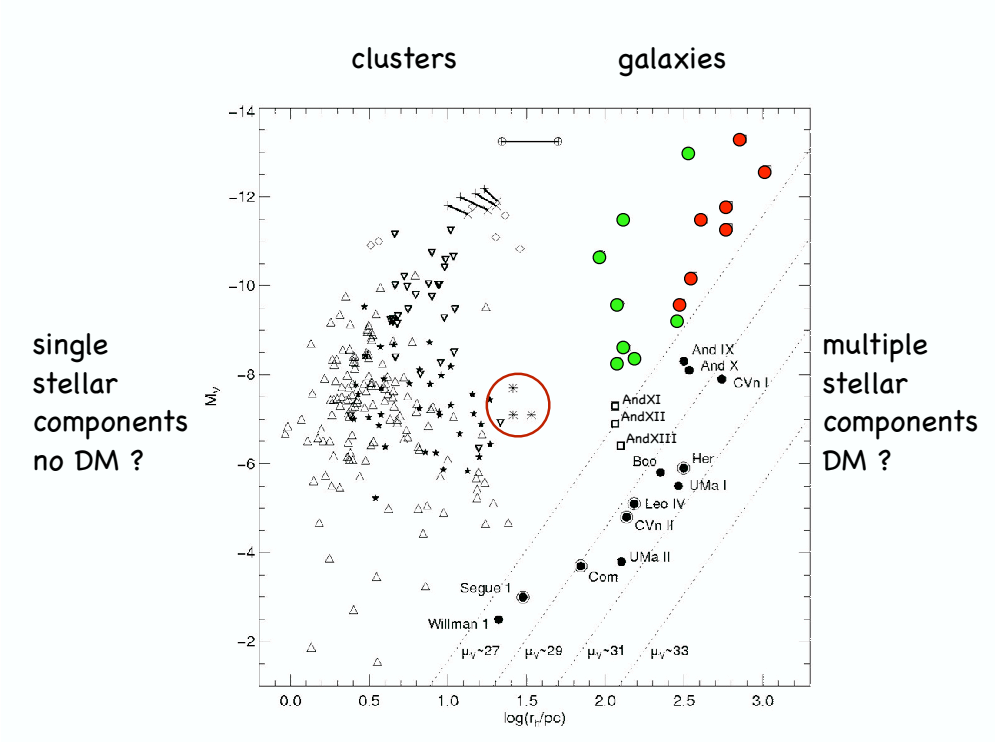


Figure 2: The apparent dichotomy between stellar clusters–no dark matter and dwarf galaxies–with dark matter in the luminosity -v- half-light radius domain. Classical Galactic dSphs are green filled symbols; M31 dSphs are red filled symbols. Some recent satellite discoveries are labelled together with the location of the new extended clusters (red circle) found around M31 by Huxor *et al.* (see Belokurov *et al.* 2007 for more details).

in these systems was more complex than originally thought leading to interesting complexities for interpreting their underlying mass distribution. As most dSphs are additionally found as satellites of larger galaxies, tidal effects will also play an important role. We have also recently found interesting substructure right in the centre of the Fornax galaxy (Battaglia *et al.* 2006 A&A 459, 423) suggestive of recent merging events, these are in addition to the more distant shells discovered by Coleman *et al.* (2004, AJ 127, 832; 2005 PASA 22, 162).

With an enlarged sample of dSph stars other statistical studies are possible. In particular, one of great interest is in searching for the presence of very metal poor (first generation ?) stars with [Fe/H] $<$ -3.0 for more detailed higher resolution spectroscopic followup. Our first results from four dSphs reveals a surprising dearth of even candidate very metal poor stars in these galaxies (Helmi *et al.* 2006 ApJLett 651, 121). This is in contrast to the surveys for very metal poors stars in the Halo, notably by Christlieb and Beers (*e.g.* Christleib 2003 *Reviews in Modern Astronomy* 16, 191) who have found several hundred examples from their still incomplete surveys. We are currently acquiring further data on the VLT to address this issue and will then undertaking a more detailed intercomparison of the [Fe/H] distributions to quantify the significance of the differences we are seeing.

3.2 Galaxy Evolution

The evolution of galaxies and AGN from their origin to the present time is a key question in astronomy. Ideally what is needed is a volume-limited sample of galaxies and AGN to study their evolution as a function of their intrinsic properties, *i.e.* luminosity, type, and distance. Wide-area surveys provide us with large samples of different types of objects. Moreover, since these are generally carried out in several regions of the sky, they allow us to overcome the effects of cosmic variance. While surveys have been done typically in optical wavebands, astronomers have realised the need to carry out surveys at longer wavelengths as well, *e.g.* in the infrared and sub-mm since the presence of dust in galaxies and AGN makes them clearly visible at those wavelengths, while otherwise hidden or very faint in the optical. This multiwavelength approach, requires that

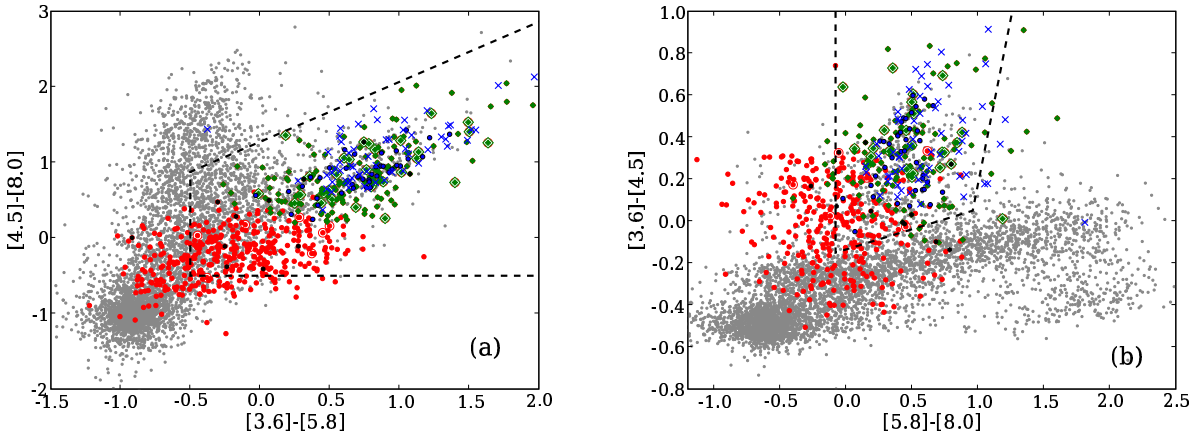


Figure 3: Colour-colour plots for optically faint objects in the ELAIS N1 region. Objects marked with a green symbol are those whose infrared spectral energy distribution is well fitted by a power law $f \propto \nu^{\alpha}$ with $\alpha < -0.5$ (i.e. indicative of AGN dominance). Blue dots are those spectroscopically classified as type I AGNs while blue crosses are type II AGNs. The dashed lines indicate the areas defined by Lacy et al. 2005, ApJS 154 166 (left) and Stern et al. 2005, ApJ 631, 163 (right) used to select AGNs.

the depths of the different surveys are well-matched.

The Spitzer Space Telescope launched on 25 August 2003, has obtained infrared images and spectra in the wavelength range 3μm to 180μm and the SWIRE consortium (Lonsdale et al. 2003, PASP 115, 897) has used this facility to carry out a large infrared survey over 60 square degrees. Surveys in the optical and near-IR (VST, WFCAM, VISTA) will enable us to model the spectral energy distribution of the objects and determine their redshifts and types. This will allow us to study the clustering of infrared galaxies at larger redshifts than previously done. In fact, with the large number of objects detected by SWIRE we are carrying out studies of clustering of objects by type, luminosity as well as clustering between different types of objects (Oliver et al. 2004, ApJS 154, 30). Traditionally IR infrared surveys rely on the optical for source identification. For the SWIRE survey the data obtained as part of the Wide Field Survey (WFS) with the INT was used. The optical data has proven essential for the study of detected sources in the mid-IR (*e.g.* Rowan-Robinson et al. 2005, AJ 129, 1183) as well as in X-rays by Chandra (Franceschini et al. 2005, AJ 129, 2074) . The optical catalogues have also been used to select U band drop-outs and study the z=3 QSO luminosity function (Siana et al. 2006, astro-ph/0604373). Due to their importance, these optical products have been included as legacy products in the SWIRE data release.

Various authors (*e.g.* Lacy et al. 2004, ApJS 154, 166; Sajina et al. 2005, ApJ 621, 256) have shown that the mid-IR colours from Spitzer are efficient to select reliable and complete samples of AGN, independently of their dust obscuration. This is due to the fact that the UV photons from the central engine are absorbed by the dust torus and re-emitted in the IR. For objects detected in the optical, their morphology and colours improve the selection of type I AGNs. However little is known about faint optical objects which usually show powerful IR emission consistent with massive galaxies at $z \sim 1-2$ (e.g. Yan et al. 2004, ApJS 154, 75). Indeed measurements in the IR range provide an opportunity to look for obsecured AGNs not detected in the optical surveys. Figure 3 shows two colour-colour plots used to select AGNs from mid-IR fluxes. The objects in colour are those without an optical counterpart down to r=23.5 (Vega). Furthermore, AGN dominated objects (even if they are obscured and very faint in the optical) do show a power law emission ($f_{\nu} \propto \nu^{\alpha}$) in the mid-IR with a variety of slopes (*e.g.* Alonso-Herrero et al. 2006, ApJ 640, 167). Those objects for which the spectral energy distribution can be well fit by a power law with $\alpha < -0.5$ are marked in green in the figure 3. Near-IR fluxes measured for these objects can overcome the lack of optical information. The Deep eXtragalactic Survey (DXS) is one of the WFCAM UKIDSS surveys which is observing areas previously observed by Spitzer in the northern hemisphere (while the VIDEO survey will do the same with VISTA in the south). Figure 4 shows two examples of colour diagrams using both near-IR from UKIDSS and mid-IR from SWIRE. Type I AGNs (blue dots) are clearly separated from type II and obscured AGNs in the J-K colour. In fact a carefull SED modelling

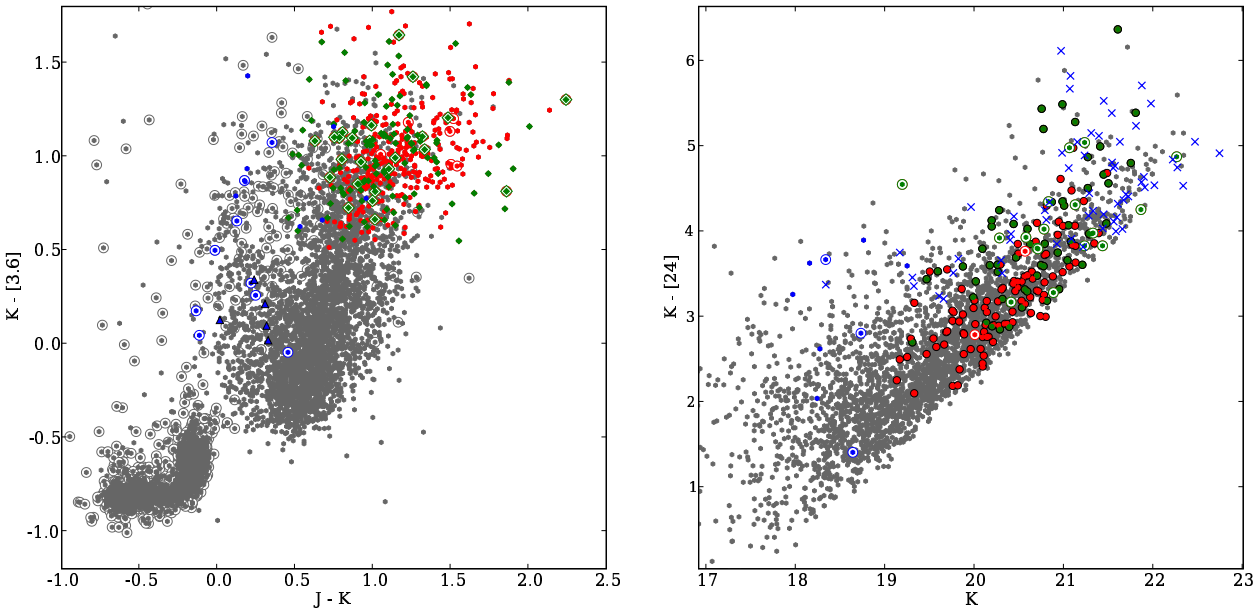


Figure 4: Near- and mid-IR colours of optically faint objects in the ELAIS N1 region. Symbols as in the previous figure. While mid-IR colours are able to select reliable samples of AGNs, near-IR can be used to distinguish type I and II in those samples.

suggests that optically blank fields are a mixture of low redshift AGNs (those showing a power law SED) and $1 < z < 2$ elliptical and starburst galaxies.

The UKIDSS Large Area Survey (LAS) is observing the areas previously observed by SDSS. The comparison of these two large surveys provides a powerful tool to discover high-z QSOs as demonstrated with the first $z \sim 6$ QSO detected using WFCAM (Venemans et al. 2006, MNRAS 376, L76). This work will continue in the south with the advent of the VISTA Hemisphere Survey, which when combined with equivalent VST optical surveys, will, among many other things, enable searches for similar rare objects in the south.

Herschel (to be launched in July 2008 with a nominal operational lifetime of three years) will cover the far infrared to sub-millimetre parts of the spectrum (from 60μm to 670μm), opening up an almost unexplored regime which cannot be observed well from the ground. Surveys carried out with Herschel will be relying on Spitzer observations for their identifications. The combination of Spitzer and Herschel will allow us to detect very rare, luminous objects for which we will need to obtain redshifts (*e.g.* using ALMA). Obviously, in order to detect those rare objects we need to sample the biggest volume possible.

3.3 Supernova Rates

In the recent years the rate of occurrence of Supernovae (SNe) as a function of cosmic time has been recognised as an important topic as testified by the number of new publications on the subject. The SNe rate is critically linked to some of the most basic ingredients of galaxy evolution, such as mass, star formation history (SFH), metallicity and environment. For example, the rate of core collapse (CC) SNe, including type II and type Ib/c SNe, provides, for a given initial mass function, a direct measurement of the on-going star formation rate (SFR) because of the short time scale that leads massive stars ($M > 8M_{\odot}$) to a core collapse-led explosion. On the other hand, the rate of SNe Ia, echoes the long term SFH because these SNe originate from low mass stars in binary systems and show a wide range of delay times between the progenitor formation and the explosion. In recent years the rate of type Ia SNe as a function of redshift and galaxy type has been used to constrain the progenitor scenario (*e.g.* Strolger et al. 2004, ApJ 613, 200; Sullivan et al. 2006, ApJ 648, 868), however, the current studies have still not provided conclusive evidence in favor of any of the different progenitor scenarios.

Riello's PhD thesis (Jan. 2005) was focussed on a project which made use of ESO facilities to perform a SNe search aimed at discovering both type Ia and type II+Ib/c SNe to estimate the corresponding SNe rates at an

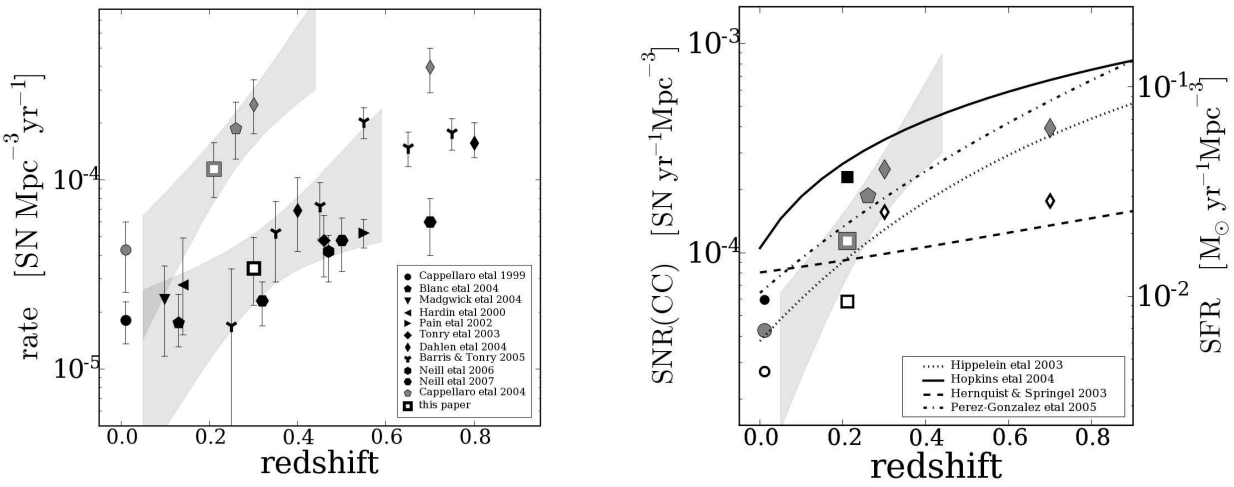


Figure 5: *Left panel* : Observed SNe rates as function of redshift. The black (grey) symbols indicate SNe Ia (SN CC) rate measurements. The shaded area shows the 1σ confi dence level of our rate evolution estimate. *Right panel* : Comparison between SN CC and SF rate evolution. Symbols are as in the left panel with in addition open symbols showing measurements not corrected for extinction and fi lled black symbols estimates for a high extinction correction. Lines are selected SFR evolution from the literature. All have been scaled to the adopted SalA IMF.

average redshift $z \simeq 0.3$. Preliminary results based on a sub-sample of the collected data had already been published (Cappellaro, Riello, Altavilla et al. 2004 A&A 430, 83). The observing program was completed in Spring 2005 and now Riello is co-authoring a publication presenting the fi nal results of the project (Botticella, Riello, Cappellaro et al. 2007, A&A submitted). This work is not just an extension of the previous results to a wider sample but introduces some major novel concepts.

During the fi rst half of 2006 Riello extended the SNe host galaxy extinction models (Riello & Patat 2005, MNRAS 362, 671), which were initially developed for Type Ia SNe only, to estimate the extinction correction also for core collapse SNe. Using these models he has been able to account for the fraction of SNe that were not detected due to the high extinction from their parent galaxy. Although the correction can only be applied statistically, it provides a signifi cant improvement with respect to other recent works, where the effects of extinction were either ignored or simply treated as a ”fudge-factor” with no physical meaning.

These results indicate that, with respect to the local value, the CC SNe rate per unit *B*-band luminosity increases by a factor of ~ 2 already at $z \simeq 0.2$ whereas the Type Ia SNe rate is almost constant up to redshift $z \simeq 0.3$. The observed trend of SNe rates from red to blue galaxies at intermediate redshifts is similar to that in the local Universe. These results are in good agreement with published measurements of the evolution of the rates of both SNe types (see Fig. 5). Finally, comparing the observed evolution of the Type Ia rate with predictions of different progenitor models Riello found that, although the models are not constrained by measurements at intermediate redshifts, they reproduce very well the observed shallow evolution of the type Ia SNe rate.

This and other recent works (e.g. Mannucci et al. 2006, MNRAS 370, 773; Neill et al. 2007 astro-ph/0701161) have highlighted the crucial role played by dust extinction, and the corresponding correction, in measuring the rate of SNe. There are two main lines of research that we propose to carry out on SNe rates. The fi rst line of research will be dedicated to extend the dust correction models (Riello & Patat 2005), which currently only deal with ”normal” spiral galaxies, to treat also Starbursts and (U)LIRGs which are extremely dust-rich environments and therefore can introduce a signifi cant bias in the SNe detection effi ciency. The second line of research will tackle the problem from an observational point of view: we plan to extend SNe searches to the near infrared (NIR) domain where the effect of dust extinction is mitigated compared to optical bands (*e.g.* V, R). In particular, we plan to investigate the feasibility of a ”piggy-back” SNe search using WFCAM data

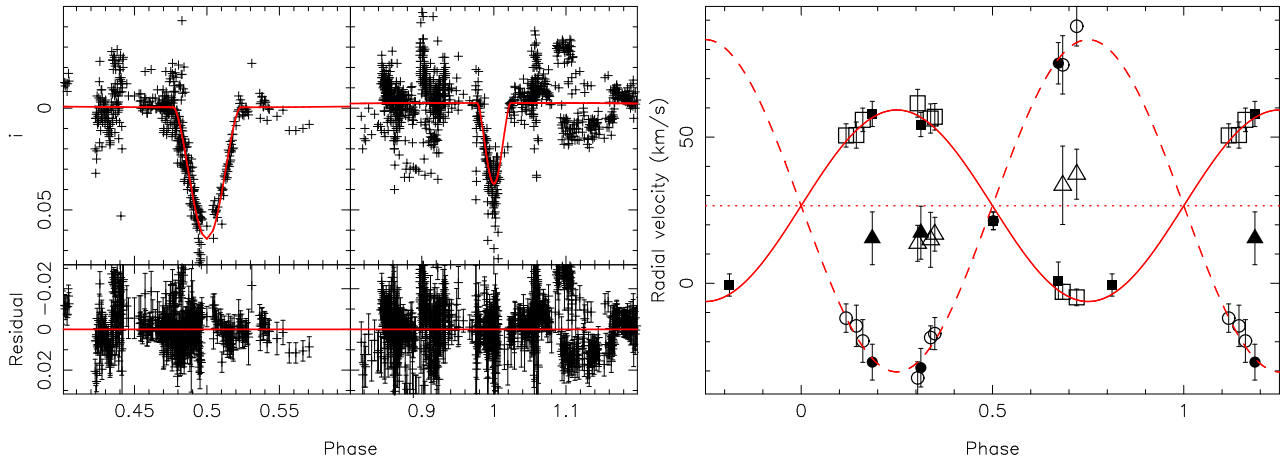


Figure 6: Left panel: phase-folded *i*-band lightcurve for JW380 (phase 0 is defined to be mid secondary eclipse) magnified around the eclipses. The upper panel also shows the best fit JKTEBOP model (Southworth, Maxted & Smalley 2004, MNRAS 351, 1277). Residuals are shown in the lower plots. Right panel: radial velocity curve for JW380 folded on a period of 5.2991 days. The curves show the best fit circular orbit to the primary (solid) and secondary (dashed) radial velocities. The dotted line indicates the systemic radial velocity. We find 3 components in the cross-correlation functions (primary: squares, secondary: circle, tertiary: triangles). Filled symbols denote data from VLT/FLAMES and open symbols are from Gemini/Phoenix.

collected in the framework of the UKIDSS ultra-deep survey (UDS), based on the CASU-pioneered list-driven photometry approach that is currently used by other members of the CASU group to search for transiting planets and low-mass eclipsing binary systems. If this pilot program proves effective we plan to extend it using data from the deep and ultra-deep VISTA public surveys and if successful obtain dedicated telescope time to carry out a spectroscopic follow-up.

3.4 Time domain astronomy

These projects make extensive use of VDFS software development for optimising knowledge extraction from time series photometry. In particular they make extensive use of the CASU-developed list-driven photometry and also have employed the CASU adaptive kernel difference image analysis (DIA) software designed for work in complex nebulosity or severely crowded regions.

The Monitor project is a photometric monitoring survey of nine young (1-200 Myr) open clusters to search for eclipses and transits by very-low-mass (and brown dwarf) companions and planets. Monitor uses 2m and 4m class telescopes equipped with wide-field imaging CCD mosaics for the initial survey, to a total of 100 hours of coverage in each cluster and a typical observing cadence of a few minutes. The key goals of the survey are:

- to calibrate the relation between age, mass, radius and where possible luminosity, from the K dwarf to the planet regime, in an age range where constraints on evolutionary models are currently very scarce;
- to detect an exoplanet in one of our youngest targets ($\leq$ 10 Myr) and provide important constraints on planet formation and migration time-scales and their relation to protoplanetary disc lifetimes;
- to measure large numbers of rotation periods for low mass stars in each of our target clusters, and to investigate the evolution of angular momentum for stars with and without radiative cores.

The scope and sensitivity of Monitor is discussed in detail in Aigrain et al. (2006, MNRAS 375, 29). The initial imaging, processing and lightcurve generation (Irwin et al. 2006, MNRAS 370, 945) is now complete for 6 of our target clusters. We have identified in excess of 50 high-quality candidates which show deep and repeatable eclipse events and show photometry consistant with cluster membership. Spectroscopic follow-up is now underway for the candidate eclipsing binaries with the aim of: (1) Securing cluster membership via measurement of gravity and youth diagnostics: Lithium, Hα, Na and K line strengths and equivalent widths; and (2) Multi-epoch radial velocity measurements to determine the initial velocity curve and systemic velocity (also

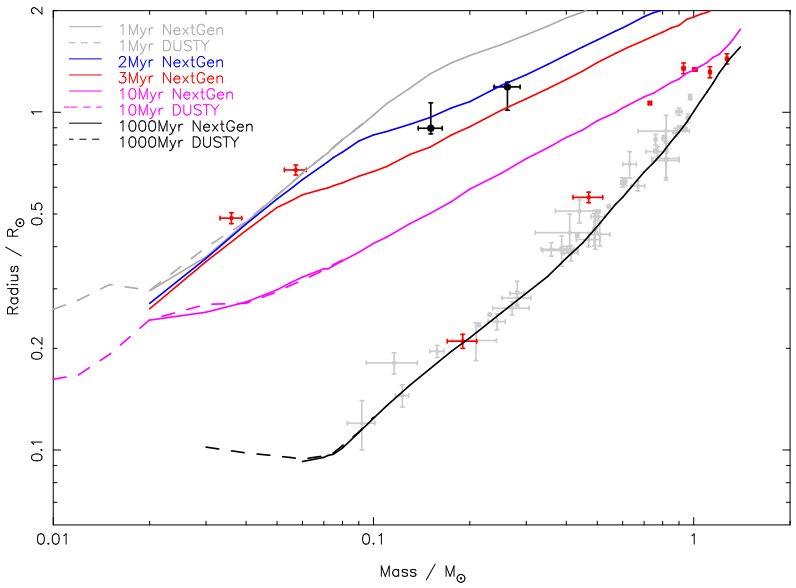


Figure 7: The Mass-radius relaton for low-mass stars and eclipsing binaries. JW380 is shown with black points and error bars, and the lines show pre-main-sequence NextGen (Baraffe et al. 1998, A&A 337, 403) and DUST (Chabrier et al. 2000, ApJ 542, 464) at ages of 1 Myr, 2 Myr, 3 Myr, 10 Myr and 1 Gyr (from top to bottom). Systems shown in red are the few known pre-main sequence eclipsing binaries, while systems in grey are a compilation of fi eld eclipsing binaries.

---

an important criteria for cluster membership). We have made use of Phoenix on Gemini and Flames+Giraffe on the ESO VLT and, where our targets are bright enough, EMMI on the ESO NTT to follow-up an initial sample of ∼15 candidates. We have applied for further observing time with Flames (UVES+Giraffe) for four clusters in ESO Period 80.

Our fi rst published eclipsing binary is the ONC cluster member JW380 (Irwin et al. 2007, MNRAS submitted). In Figures 6 we show the lightcurve and radial velocity data for the system together wth best fi t models. We derive primary and secondary masses of $0.26M_{\odot}$ and $0.15M_{\odot}$ for the components which are in a 5.3 day orbital period. Our measurements of the masses and radii of these very young stars are shown in Figure 7, compared to model predictions and the handful of other known young systems with dynamically measured parameters. We fi nd evidence for a third component in the system. Further spectroscopy is planned to reduce the errors on our current measurements (We are dominated by uncertainties on the relative effective temperatures of the components), and to establish whether the third object is physically connected to the system.

Detailed analysis of the Monitor data is now underway to search for shallower transit events, by fi ltering out intrinsic stellar variability caused by accretion, rotation, spots and activity (Miller et al. in prep). Rotation periods and angular momentum are discussed in detail in Irwin et al. (2006, MNRAS 370, 945; 2007 MNRAS 377, 741) and the role of disk locking in the young cluster N2362 is explored in Hodgkin et al. (in prep).

The WFCAM transit survey: Searches for extra-solar planets have, for the most part, targeted solar analogues via precise spectroscopy of individual stars using large aperture telescopes (*e.g.* Butler et al. 2006, ApJ 646, 505), or wide-angle photometry searching for transits in large numbers of objects in the fi eld using very modest telescopes (*e.g.* Horne 2003, ASPC 294, 361). In excess of 200 exo-planets are now known, the vast majority of which are gas giants with masses close to that of Jupiter.

Cool stars are up to 10 times smaller than solar-type stars, and signifi cantly fainter. The available sample for precision spectroscopic observation is consequently rather small (*e.g.* Bonfi ls et al. 2005, A&A 443, L15). The transit method is potentially more effi cient and, thanks to the smaller stellar radii, capable of detecting smaller planets around such stars. A transiting earth-like planet for example, would cause a detectable ∼1% transit of a $0.1M_{\odot}$ star. However, cool stars are much fainter at optical wavelengths, and existing optical transit surveys

lack the sensitivity to find such planets. In the infrared, where cool stars are brightest, the search for transiting cool star planets becomes viable.

The theory of planet formation, via core accretion, predicts that gas giants should be extremely rare around M dwarfs, while rocky planets would form just as frequently as around higher mass stars (Laughlin et al. 2004, ApJ 612, L73; Ida & Lin 2005, ApJ 626, 1045). However, a handful of results seem to be at odds with this (e.g. Delfosse et al. 1998, A&A 338, L67; Rivera et al. 2005, ApJ 634, 625) and suggest that at least some M dwarfs have Jupiter mass planets.

The WFCAM Transit Survey (PIs Hodgkin & Pinfield) has been awarded 200 nights of UKIRT time to perform the first ever systematic near infrared search for transiting exo-planets around cool dwarfs. Hodgkin is responsible for the processing and analysis of the lightcurves following the procedures outlined in Irwin et al. (2007, MNRAS 375, 1449). Hodgkin will work with Aigrain (Exeter) to search the lightcurves for transits and with Pinfield (Herts) to design the follow-up strategy for candidates.

3.5 Exploiting optical surveys

Greimel's research in the last 3 years has focussed on multi-disciplinary data mining of optical surveys including: finding and studying planetary nebulae (PNe), AGB stars and variables in nearby galaxies using the INT WFC Local Group Census in collaboration with with Laura Magrini (U. of Florence), R. Corradi and P. Leisy (ING), H. Habing (Leiden Observatory), and M.R. Cioni (U. of Hertfordshire); using SDSS catalogues and spectroscopy in collaboration with T. Augusteijn (NOT), E. van den Besselar and P. Groot (U. of Nijmegen) to define and analyse a catalogue of high proper motion red dwarf-white dwarf binaries; detailed exploitation and support for the IPHAS Galactic Plane survey; working on precision radial velocity measurements in collaboration with S. Yang (U. of Victoria) using telluric lines to correct the wavelength scale and improve the measured radial velocities of 12 year baseline dataset of a few hundred stars taken with the DAO 1.2m telescope to an accuracy of 20m/s.

The majority of his recent research is linked to exploiting the IPHAS survey. Apart from being responsible for the observing scripts and survey management he is also heavily involved in follow up source selection and spectroscopy. This includes searches for extended PNe in collaboration with L. Sabin, A. Mampaso (IAC) and R. Corradi (ING) using large scale (2x2 deg) mosaics created by J. Irwin at CASU. Although still ongoing this has already lead to the discovery of nearly one hundred new PNe. A companion collaborative search for compact PNe with K. Viironen (IAC) based on CASU-generated source catalogues so far has found 30 new spectroscopically confirmed compact PNe. Searches for symbiotic stars in collaboration with E. Flores (IAC) and R. Corradi (ING) from federated IPHAS and 2MASS photometric information have found 3 new symbiotic stars in the IPHAS area (only 11 symbiotics are known to date). An emission line star survey is in progress and several thousand candidates have been selected and follow up spectroscopy acquired by D. Steeghs (SAO) using Hectospec on the 6m MMT telescope. This search has uncovered one fascinating source - IPHASJ021448.45+622622.6 - only the second object known of its type. Work to understand this star which might be in a so far unknown short-lived phase of stellar evolution is being done in collaboration with D. Lennon (ING). Greimel has also analysed the IPHAS data for variable stars and from the first two years of data found 17000 variable star candidates. Follow up photometry in collaboration with R. Robb (U. Victoria) has shown that most systems picked up by the IPHAS survey are short period eclipsing stars.

The main science direction of his research for the grant period will be the continuing exploitation of the vast amount of data produced by the IPHAS and UVEX (a related INT WFC UV extension survey of IPHAS – PI Groot) surveys. This will be extended to the south once VPHAS+[5] becomes available. Follow up spectroscopy and photometry is currently being acquired (*e.g.* through the ITP program on La Palma) and this will continue over the next few years. The focus of the work for the areas mentioned above will shift from the general

[5] At the end of 2005, the ESO Public Surveys Panel identified two UK-led public surveys VPHAS+ (PI Drew), the southern equivalent of IPHAS, and ATLAS (PI Shanks), a large area high latitude survey as key components of the VST survey programme (the only other supported VST public survey is KIDS – PI Kuijken).

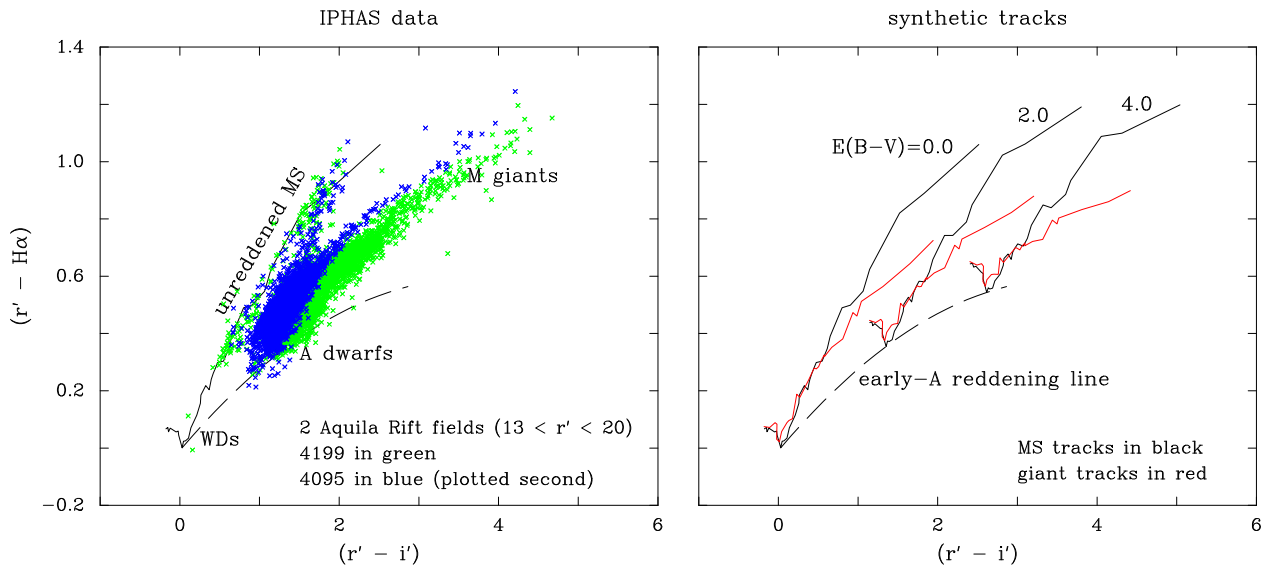


Figure 8: The left panel shows the $(r' - H\alpha, r' - i')$ plane for data from two differently reddened IPHAS fi elds in the Aquila Rift region, annotated to point out where key object types fall. The right hand panel shows synthetic tracks in the same colour-colour plane to demonstrate how the main sequence (black), and giants sequence (red) shift with increasing reddening (fi gure courtesy of J. Drew).

selection of sources to the analysis of the most interesting objects. Cross correlation of the IPHAS catalogue with existing surveys from X-ray to radio will also be an important source of rare objects that IPHAS will discover. Especially interesting will be the cross correlation of the IPHAS and WFCAM UKIDSS GPS and upcoming VISTA catalogues. This is a challenging problem simply due to the vast amount of objects ($\sim$1 billion) produced by the surveys.

With the IPHAS survey nearing completion use of the data for galactic structure research will be possible by looking at the galactic distribution of stellar population tracers (see fi gure 8) in conjunction with constructing a 3D extinction map of the northern galactic plane in collaboration with S. Sale and J. Drew (Imperial). Information from other spectral bands, *e.g.* g-band observations from UVEX and J,H,K from 2MASS/UKIDDS GPS will be factored in to improve the accuracy.

The IPHAS data can also be used to look for high proper motion objects in the galactic plane where surveys for these objects are traditionally incomplete. A search for bright high proper motion objects is possible by correlating the IPHAS and 2MASS surveys. The work on the SDSS selected sample of red-dwarf white-dwarf binaries will also continue with follow up spectroscopy of interesting objects. Additionally the selection that has been confi ned to a high proper motion sample to avoid contamination by Quasars will be extended to cover the complete SDSS dataset.

3.6 Summary of resources requested

Staff: 1.5 FTE/yr from Q2 2008
Costs: included in overall formula funding per FTE per annum

3.7 Key deliverables and milestones

2008–2013 Research papers, conference proceedings, seminars
2008–2013 Regular feedback informing design and operation of CASU pipelines and analysis systems based on research use of CASU data products

4 VISTA Data Flow System

The UKIRT Wide Field Camera (WFCAM) on Mauna Kea and the VISTA IR mosaic camera at ESO, Paranal, with respectively 4 Rockwell 2kx2k and 16 Raytheon 2kx2k NIR arrays on 4m-class telescopes, represent an enormous leap in deep NIR survey capability. With combined nightly data-rates of up to 0.5-1TB, automated pipeline processing and data management requirements are paramount.

In response to this challenge over the last few years CASU have developed a pipeline architecture capable of processing NIR imaging data from WFCAM and VISTA which produces well-calibrated quality-assessed science products. The end-to-end system robustly removes instrument and night sky signatures; monitors data quality and system integrity; provides astrometric and photometric calibration; and generates photon noise-limited images and astronomical catalogues.

With average nightly data rates of 0.2-0.3TB, these systems will generate PB scale volumes of raw data over a 5-10 year period. This takes the pipeline processing, data management and archive requirements into a new regime in terms of processing power, storage and automated operation. Further constraints are imposed by progressively more demanding science requirements coupled with a need for low operating and maintenance costs.

The VISTA Data Flow System (VDFS) has been designed and developed to meet these aspirations and has been thoroughly trialled and tested on WFCAM data for the last two years. By the time VISTA comes on-line for survey science in 2008 the majority of the main development work will have been completed. However, there will still be a strong need for further development, both operational and science-driven, and a requirement to maintain and upgrade all the VDFS pipelines as a result of tuning of the processing system, including feedback from science users, during the first years of the VISTA surveys. The synergy between the WFCAM and VISTA projects and the commonality of their respective requirements with other survey cameras, strongly suggested that similar strategies should be implemented built around a modular system driven by processing recipes. This has now proven an efficient and cost-effective solution and minimises future risk with, for example, VISTA data processing deliverables.

The development work for the VDFS has been separately funded through two rounds of e-science funding and covers the development and delivery of the following software items:

- summit quality control data processing pipelines for both WFCAM and VISTA;
- a VISTA calibration pipeline to be run by ESO in Garching including processing software and detailed documentation deliverables;[6]
- WFCAM and the UK VISTA data processing pipelines to be run in Cambridge that have to meet defined WFCAM and VISTA science requirements (see http://casu.ast.cam.ac.uk/ for more details);
- advanced processing software to be run in Edinburgh to be driven from the WFCAM and VISTA science archive databases.[7]

The operational aspects of running the UK part of these pipelines is the main subject of this section of the grant bid. For context we also briefly discuss details of schedule and the progress made to date in the CASU part of the VDFS project. Further details of the VDFS planning and pipeline development are summarised in the papers in Irwin *et al.* 2004, SPIE 5493, 411 and Emerson *et al.* 2004, SPIE 5493, 401.

4.1 VISTA data processing

The Visible and Infrared Survey Telescope for Astronomy (VISTA) is a new 4-m telescope designed specifically for imaging survey work at visible and The VISTA infrared camera (VIRCAM) has a field of view of 1.5°

[6] These deliverables are defined in "Data Flow for VLT instruments Requirement Specification VLT-SPE-ESO-19000-1618 Issue 1.0 and in "Data Flow for VLT/VLTI Instruments Deliverables Specification VLT-SPE-ESO-19000-1618 Issue 2.0."

[7] Additional archive end database-driven processing includes deeper stacking and mosaicing in specified fields and provision for list-driven matched-aperture photometry using CASU-supplied software.

diameter with a pawprint that will cover 0.59 degree^2 in ZYJHK$_s$ passbands, using a 4×4 array of 2k×2k non-buttable Raytheon detectors with 0.34 arcsec pixels resulting in a 269MB FITS file per exposure. Surveys using this camera are expected to produce an average of 200GB of data a night with high volume nights producing up to ∼650GB.

The majority ($\sim 75-80\%$) of the time on VISTA will be dedicated to 6 UK-led public surveys: VHS - a VISTA hemisphere survey, PI McMahon; Viking - a deep NIR equivalent of the optical 1500 sq deg KIDS survey, PI - Sutherland; VMC - a systematic survey of the Magellanic Clouds, PI - Cioni; VIDEO - a deep targetted study of galaxy/cluster structure and evolution in selected fields, PI Jarvis, UltraVISTA - an ultra deep survey of the COSMOS field, PIs - Dunlop, Le Fevre, Franx, Fynbo; VVV - a Bulge variability and Galactic plane survey, PIs - Minniti and Lucas

Each of these surveys have many common requirements but also given their different scientific aspirations require different science products and different levels of sophistication of the pipeline processing. Both CASU and WFAU have been interacting regularly with the survey PIs and have been closely involved in negotiating the science requirements and in the preparation and agreement of the design of the survey and the delivery of the survey management plans to ESO. The VDFS UK pipeline will be used to process all data taken on VISTA when when it begins science observations, this includes the UK-led public surveys and also all other PI data.

The development of the VDFS has been closely monitored through regular VDMT and VDUC meetings and recently underwent a detailed external review in October 2006. The review panel subsequently produced a detailed report from which the following quotation is taken:

"*The VDFS project team has done a great job at hitting big problems head on first. They have built a state-of-the-art processing/archiving system that is better than the SDSS system in many respects.*"

– UK Review Panel Report 28 November 2006

The data flow system now been thoroughly trialled and tested on WFCAM data for the last two years at data rates of up to 250GB/night. Versions of this pipeline have also been used to process ESO ISAAC data e.g. the FIRES survey data and data from a wide range of optical CCD mosaic cameras. By the time VISTA comes on-line for survey science in 2008 the majority of the main development work will have been completed. However, there will still be a strong need for further development, both operational and science-driven, and a requirement to maintain and upgrade all the VDFS pipelines as a result of tuning of the processing system, including feedback from science users, particularly during the first years of the VISTA surveys.

The VISTA Data Flow System has been designed and developed by CASU to meet a range of scientific and operational requirements. The first are the requirements set out by ESO and which have all been thoroughly reviewed and passed by ESO. The second are the additional requirements of the UK community[8] which are more focussed on the final quality of the science data products that come from the VDFS. As a result VIRCAM data will pass through at least three pipelines from the time of observation. We discuss these in turn in the following sections since all requirement some continuing committment of staff effort in this rolling grant application.

Considerable effort has gone into automated quality control (QC) parameter generation in the pipeline design, (see the Data Reduction Library Design v1.6 available at http://www.vista.ac.uk/vdfs/esoqc1/) for further information). The most basic version of the QC process occurs in near-time on Paranal, while more sophisticated versions will be run in Garching and later in Cambridge. All of the Cambridge pipeline QC information will be available to PIs via a QC database in Cambridge[9] and is also recorded in the data product FITS headers.

4.2 ESO deliverables for VISTA pipeline

CASU are responsible for delivering to ESO software modules for the following two pipelines.

[8] see documentation set for UK review referred to at the end of this section

[9] for an example see http://casu.ast.cam.ac.uk/surveys-projects/wfcam/data-processing/

On Paranal a causal QC pipeline will run at the telescope in near real time. It will use purpose-written recipes to compute a range of QC parameters which will enable early assessment of the quality of the data. Calibration data (such as flat fields) will be drawn from a library which will be updated only infrequently. Because of the causal nature of pipeline execution, each recipe can only make use of data taken up to that point.

In Garching a calibration and quality control pipeline will be run at ESO headquarters, to remove instrumental signatures, produce catalogues and enable astrometric and photometric calibration in addition to extra quality control checks to be made. This will use the same software modules as both the Paranal and UK pipelines.

ESO also want a longer term committment from CASU to update and maintain these pipelines over the duration of the VISTA surveys.

At the time of writing, the ESO components of the VISTA Data Flow System development have progressed on schedule and have passed several important milestones. Central to this is that the relationship between CASU and ESO DFG has been an extremely productive two-way process, with a close professional relationship having been built up in which the UK deliverables to ESO for the VDFS project are being designed, built and tested in the manner of a partnership between the principals.

Following the successful Preliminary Design Review at ESO in April 2004, many constructive suggestions were incorporated into the design for the Final Design Review, also at ESO, in January 2005. After some minor iterations, the design was finalised and serious code building started in Summer 2005. The first code release to ESO, version 0.1, was in December 2005. Minor releases followed, up to v0.5 in March 2007, a milestone release fulfilling the ESO standard of "Preliminary Acceptance Europe".

At the current time, the VISTA pipeline has been tested on near-realistic simulated data and on a certain amount of laboratory–test data from the actual VIRCAM camera produced at RAL. Testing on data from the camera in-situ at Paranal will commence shortly.

In addition to these software modules CASU have also delivered extensive sets of documentation ranging from a detailed specification and prototype exposure time calculator to documents describing the design and detailed interface of the data reduction library and a detailed discussion of the processing and calibration procedures required for calibration of VIRCAM data. A list of the relevant documents and related papers is available at http://www.ast.cam.ac.uk/ukreview, these formed part of the recent documentation set for the UK review of the VDFS project.

4.3 UK deliverables for VISTA pipeline

The UK pipeline will be responsible for producing most of the science-grade reductions of VIRCAM data. Many of the algorithms used in the QC pipelines will be re-employed, the difference being that a more flexible approach can be taken in the production of master calibration data. Processing steps that are not possible in the QC pipelines will be run in the UK, thus improving the quality of the science data products.

The UK processing will adopt the same philosophy as the WFCAM pipeline and will be based around a standard pipeline to remove instrumental signature, generate astrometric and photometric calibration, provide assorted quality control measures, and produce a standard set of science products. In essence the standard pipeline in Cambridge will be a superset of the Garching calibration pipeline such that if the same processing sequence is used along with the same master calibration files, identical output products will result. It will operate on the Observing Blocks (OBs) taken during each night to produce the final individual instrument signature-free images, including mosaiced tiles, and to derive astronomical object catalogues from them. The object catalogues will be used to generate the astrometric and photometric calibration, again in a similar manner to that developed for WFCAM.

A further processing stage will produce detailed estimates of the Point Spread Function (PSF) and its variation as a function of position within the array. This information will then be used to drive the PSF fitting software which adds extra entries to the catalogues derived from the PSF fitting.

The UK pipeline will be implemented to keep up readily with the average expected VIRCAM data rate of ≈200 Gbytes/night, which is comparable to the current WFCAM average ≈150 Gbytes/night. Hence the availability of science products will be mainly driven by the frequency of ESO data delivery to CASU and the need to control master calibration updates. We anticipate that, as for WFCAM, a steady-state processing schedule that delivers science products roughly one month after the data are taken is achievable.

Data products will be made securely available for WFAU to transfer to Edinburgh from Cambridge directly via the Internet (or private UKLight circuit) with an end-to-end minimum bandwidth of 1 Gbits/s, again in a similar manner to that developed for WFCAM.

Asynchronous processing operations such as time-series generation (and analysis), deep stacking and general-purpose list-driven photometry occur more naturally at the archive end and VDFS will offer a choice of methods to do this, including use of software developed by CASU.

4.3.1 UK pipeline extra functionality

Because the summit and Garching pipelines only work causally with observation time, the best choice of calibration information which is relevant to a particular observation may not be available when reducing a given OB. The Cambridge pipeline will allow an entire night (or more) to be treated as a single entity and thus allow OBs to be reduced in a different order from when they were taken, ensuring that the most up-to-date calibration information is used. This is particularly advantageous at the sky-correction stage.

As an example, laboratory VIRCAM frames suffer from low level striping (similar to WFCAM curtaining) perpendicular to the detector readout axis. This is a random electronic effect that is common to each set of 4 detectors attached to their IRACE controllers. Because the Summit and Garching pipelines operate at the detector level this effect is difficult to remove without affecting the astronomical objects on the frame. However, the duplicity of the effect over 4 detectors means that by processing the entire observation as a single entity it is possible to use the 4-fold redundancy of the patterns to make a robust estimate of the problem and to reduce it to negligible level.

The layout of the detectors in the VIRCAM focal plane of VIRCAM and the science requirements for nearly uniform observational coverage of contiguous areas of sky has lead to a tiling strategy involving six pawprints. A full tile observation will ensure nearly uniform coverage and that each area of sky will be observed at least twice. A result of this strategy is that in order to extract objects to the depth specified by a given survey, it is necessary to combine the resulting pawprints into a filled tile image. The significant distortion of the focal plane by the camera optics also requires that in order to create such a tile, input pawprints will have to be resampled onto a new grid. Only the Cambridge pipeline will combine pawprints into a tiled images in addition to extracting tile-level catalogues. These are expected to be the main science data products for VIRCAM.

Many of the extra processing stages described here rely on information from the observers about the nature of the observations. Information regarding the type of background correction to be done or the interpolation routine to be used in stacking needs to be specified in order for the pipeline to run without manual intervention. This will be done using a RECIPE keyword, which has been specified for the primary header unit of each of the VIRCAM FITS files. The ESO pipelines will not be able to deal with this.

4.4 VIRCAM data transport

Transmission of around 200 Gbyte a day from Paranal to Europe is currently not possible using the Internet. ESO write data to hard disks using a Linux operating system and the ext3 file system. The disks are physically located in a Chenbro chassis and are permanently mated to the corresponding mounting bracket. Currently, a capacity of 250GB per disk is the ESO standard (sic!). These disks are periodically to sent Garching, where they are inserted into an identical Chenbro chassis and the data read off.

Subsequently, because the network connection to ESO headquarters is also insufficient to support bulk data transfers, ESO plans to ship the same disks to Cambridge for insertion into another similar chassis and read again onto the CASU online storage system for use in the UK pipeline. It is expected that there will be a three-week cycle before a particular disk is returned for re-use. A standalone ingestion system (plus spare) is desirable because the constant physical insertion/removal of disks will inevitably result in frequent system reboots. After the raw data has been verified an LTO tape backup copy will be created for useas an off-line raw data archive. We note that data ingestion and verification is usually one of the most manually-intensive tasks in a pipeline, taking non-trivial operator effort. A substantial percentage of this time is often spent in recovering lost data.

Transporting processed data from Cambridge to Edinburgh is much simpler. With gigabit connectivity, either over JANET or the reserved UKLight connection, reduced science products will be flagged as ready-to-transfer and copied up to the WFAU archive automatically, using high-performance networking protocols.

4.5 VIRCAM Pipeline requirements

4.5.1 Software

CASU data reduction pipelines have three basic components: a data access layer to allows I/O of FITS images/tables and their headers; a core of data reduction modules to perform the actual data manipulation; and an intelligent "glue" that is used to bolt the reduction modules together into a pipeline. These three components are not necessarily separate since there is often overlap between the functions provided by them (*e.g.* there is often a need to access information from a FITS file within the pipeline glue).

For the data access layer CASU use the "industry standard" package CFITSIO [10] which provides both low and high level interfaces to the data and also seemlessly to the lossless Rice tile compression that we make use of. CFITSIO is also used extensively in the high level scripting components that provides the glue to operate the pipeline.

ESO, on the other hand, have their own infrastructure called CPL which consists of libraries of routines for image and table I/O as well as some basic data manipulation algorithms. There are also higher level facilities within CPL which are used to create plugins for ESO user interfaces. A decision was reached in conjunction with ESO that the summit and Garching pipelines should be written in the ESO-specific CPL environment. This meant that the CASU modules that had been developed over many years had to be recast to use the CPL infrastructure (though they are still functionally identical to the CASU versions).

Although the CASU infrastructure is far more flexible than CPL, to use it in the UK pipeline would mean having to maintain two separate sets of data reduction modules. This would clearly not be a cost effective or efficient approach and given the requirement to update and maintain the delivered ESO pipeline components it was decided that the bulk of the Cambridge pipeline will reuse the modules delivered to ESO.

The high level Data Organiser used at ESO to sort data files and present them to the pipeline will not however be used in Cambridge. It is too restrictive and changes to it would be difficult or impossible. Instead a set of Perl scripts, following the CASU WFCAM infrastructure will be written to create the reduction menu to present data in a more flexible way to the pipeline. This will allow for rapid development of the organisational procedures and will not affect the status of the CPL-based reduction modules. Most of the development effort for this is included in the existing VDFS grants but there will inevitably be further development work required in Cambridge when the UK pipeline is operated on real VIRCAM survey data.

[10] see http://heasarc.gsfc.nasa.gov/docs/software/fitsio/fitsio.html

4.5.2 Hardware

The requirements for processing power are based on experience with WFCAM and from benchmarking results from early versions of the VIRCAM reduction recipes. Enough processing power must be available to meet the peak sustained data rate of 650Gb/day stated above. Operational experience with over a year of WFCAM reductions have shown that a 12-cpu (6×3GHz twin processor) array can handle both regular data deliveries and re-processing with improved techniques. The VISTA data rate is expected to be (only) twice that of WFCAM on average and so it is expected that a 2x16-cpu system will cope with the demands.

The ESO DFS group have benchmarked the QC pipeline and derived a requirement of 16 CPUs to cope with near-time data analysis. The base model tested is a 2.4GHz Opteron with 4Gb memory. While that system must be able to run at peak data rates, and the UK one only at mean data rates, the UK pipeline will be doing extra processing compared with the QC pipeline. This suggests that a system with 2×16 CPUs is a good compromise between over specification and the desire to have some redundancy in the processing chain.

At any one time enough new on-line storage must be available to hold the likely mean data acquisition and reduced data products for at least a semester in advance. Given the rapidly changing commodity market for bulk storage a sensible purchasing model is to acquire enough storage to cope with the anticipated data flow on a roughly yearly basis.

Unlike WFCAM, where microstepping is common, we anticipate that the majority of the VIRCAM reduced data frames will occupy a similar volume to that of the raw data. With Rice compression saving a factor of 3–4 in disk space requirements, the average raw data rate of 200Gbytes/day is equivalent to about 40–50 TB of on-line storage for raw + processed data per year. At the time of writing our current solution would be in the form of 10TB SATA RAID6 fibre channel arrays with a total cost of approximately 1k/TB of physical storage, but as usual the cost per TB is expected to reduce in the long run. Having duplicate copies of the reduced data products at CASU and WFAU is probably the most effective way to have a much desired backup. This is tantamount to making a disk copy and storing it off site.

4.5.3 Personnel

The pipeline will operate essentially continuously and although it is heavily automated, at any one time at least one person will need to monitor progress and perform interactive quality control including updating and verifying the pipeline produced calibration frames. This involves, for example, iteratively weeding out unusable dark and twilight flat field frames and visually inspecting the latest calibration frames to ensure their quality.

The data ingestion cycle takes about 3 weeks, and in a given week there is about a day of disk handling. Data-transfer disk problem-shooting is expected take a day a month. Hardware commissioning and maintenance will be required and so will operating-system maintenance and upgrades. Upgrading, maintenance and testing of new releases of VDFS software will also take significant time and effort.

At CASU VDFS development is currently funded at the 3 FTE level until the end of September 2007. Because of the delay in VISTA commissioning we are negotiating with STFC for a 6 month extension of the development effort, since some of it is contingent on early commissioning results and more detailed understanding of the properties of the VIRCAM detectors. Further operational development and the running costs of the VISTA pipeline in Cambridge are being sought as part of this CASU rolling grant renewal, with a start date of 1 April 2008. From previous experience with WFCAM operations and VDFS work so far we anticipate VDFS operating requirements to be at the 3 FTE/yr level. This includes management, normal processing, reprocessing after major bug fixes and/or enhancements, system maintenance and upgrades, liaison with external users, attendance at meetings and conferences, and exploiting training and development opportunities.

4.6 Summary of resources requested

Staff:	3.0 FTE/yr from Q2 2008
Travel and subsistence:	guideline amount plus £1k/yr for extra ESO visits
Equipment:	CPU processors, storage, curation and transfer hardware (£219k total)
Consumables:	£7.5k/yr for tapes for off-line curation, shipping and internet costs (£37.5k total)
Maintenance:	£3k/yr for CPU processors for the whole period (£15k total)

4.7 Key deliverables and milestones

Q2 2008	Begin VISTA surveys
2008–2009	VISTA pipeline optimisation phase, system configuration, CPU/Storage/Network
2009–2013	Full pipeline operations, data ingestion, verification, fault finding, science verification, onward delivery of data products to ESO and WFAU
2008–2013	Annual pipeline s/w releases to ESO (Paranal + QC pipeline)
2010–2013	Full Iterative PSF photometry deployment
2008–2013	QC database deployment and continual QC monitoring and feedback
2009–2013	Annual major VISTA data release, product packaging, QA image access

5 WFCAM data processing

The Joint Astronomy Centre (JAC) in Hawaii formally asked CASU and WFAU to devise a suitable scheme for processing and archiving all of the data taken with WFCAM. Although the proposed increase of WFCAM usage to $\approx$75% of the total time on UKIRT from 07B represents a much larger fraction of time than originally intended, in practice since WFCAM became operational on 1st April 2005, it has actually been on the telescope 68% of the time and has averaged $\sim$150 Gbytes of data per night of observations. This corresponds to around ten thousand 2k$\times$2k images per night, and more than 3 million science images since WFCAM began survey operations, highlighting both the scale of the challenge and the need for as much automation as possible.

CASU are responsible for processing all of the data taken when WFCAM is mounted on UKIRT, this includes UKIDSS[11] (*e.g.* Dye *et al.* 2006 MNRAS 372, 1227; Warren *et al.* 2007 MNRAS 375, 213; Lawrence *et al.* 2007 MNRAS in press), other survey campaign proposals and also all PI data. The current total of raw WFCAM data taken in semesters 05A, 05B, 06A and 06B is well over 50 TB.[12] This is equivalent to a total of around 16 TB of actual on-line disk storage as the lossless Rice tile compression used achieves a factor of 3.5 saving in disk space, whilst still maintaining the FITS file format and rapid access. Processed data stored on-line is $>$150% of the raw data in native volume, due to saving both the individual processed files and the interleaved products, intermediate and stacked. Here the average saving on disk space using Rice tile compression is a factor of 4 and hence only requires about 24 TB of on-line disk storage space to date. All of the ingested raw and processed is stored on-line to enable rapid reprocessing and recalibration, and rapid access to the raw and processed data for science verification and fault diagnosis. This also provides an off-site backup of the processed data products stored in Edinburgh. A live web interface, updated hourly, allows external users to see the overall progress of data reductions from raw data ingest through to processed data transfers to WFAU. An example snapshot of this is shown in figure 9.

The raw data, which arrives on tape roughly 2–4 weeks after the data is taken, is verified and ingested into the raw WFCAM data archive at CASU within a few days of arrival (see http://archive.ast.cam.ac.uk/wfcam/) and is available to registered users on request. Transfers of raw WFCAM UKIDSS and calibration data to the ESO

[11] UKIDSS requires 1000 nights of UKIRT time over 7 years, currently scheduled to complete in 2012 with final data release 2013.
[12] Semester 07A is still in progress at the time of writing and is not included in the statistics in this section.

archive are closely monitored and are handled by an ESO-supplied script and are generally completed a month or so after the raw data has been ingested in Cambridge.

Over the course of the first 4 semesters of operations various improvements have been made to the pipeline processing system that are detailed in 6-monthly reports, and annual presentations, to the UKIRT Board. These reports are available at http://casu.ast.cam.ac.uk/documents/other-meetings/ukirt/. Some of these improvements, as expected, necessitated complete reprocessing, particularly of data from the first two semesters (05A, 05B). However, the need for bulk reprocessing has lessened and reprocessing is now focussed on specific projects or specific nights of data when users feedback problems and have requested reprocessing if relevant. This interaction with end-users occurs through either web page contacts or via a dedicated casuhelp email account. Problems are assessed by the operations and development team and if feasible, and necessary, an improved and updated version of the pipeline products are created and made available.

In parallel with pipeline tuning as an image processing problem, as part of the WFCAM operational development we undertook a detailed study of the astrometric and photometric calibration based on the all-sky 2MASS point source catalogues (*e.g.* Skrutskie *et al.* 2006 AJ 131, 1163). This was only fully possible after accumulating data from more than one semester of observations to enable decoupling of instrumental and seasonal effects. The global quality of the 2MASS catalogues exceeded our expectations and we have demonstrated that it can be used to routinely deliver astrometric calibration to better than 30mas across the entire array. We have also demonstrated by repeated comparison with WFCAM measurements of external faint standards that is feasible to use 2MASS to provide an independent photometric calibration for every science product frame and that broadband photometric calibration at the 1–2% level is globally achievable. The nature of the photometric measurements bypasses the usual problems of explicitly monitoring extinction correction since the extinction in any passband in any science frame is automatically measured and accounted for. Furthermore global trends in system performance can also be readily monitored as shown in section 5.1.

Both the astrometric and photometric calibration of the pipeline products are well within the original science requirements for the processed data. Indeed the calibration is now so reliable that the frequency of standard star observations has been reduced to a much lower level.[13] We are now exploring more subtle improvements to the calibration aspects of WFCAM data by examining position-dependent illumination effects across the arrays using stacked 2MASS-WFCAM residuals from normal science data. The rms variation over the arrays is already within 2% but there are patches where the residuals are noticeably worse than this. This requires further investigation but we anticipate releasing correction tables for the next major data release milestone DR3.

5.1 WFCAM Quality Control Database

An assortment of quality control (QC) parameters including seeing, stellar ellipticity, sky properties, limiting magnitude etc.) are generated by the pipeline and are used to assess instrumental integrity, data quality and the astronomical utility of the products, some examples are shown in figure 9). This information together with other relevant information from the FITS headers, *e.g.* WCS calibration, weather conditions, temperature measures, etc., is ingested and recorded in a WFCAM PostgreSQL QC database enabling investigation of systematic effects and trends in data quality like *e.g.* the dependence of sky brightness on time, temperature, humidity and moon phase (Riello 2006). The ingestion process is also very effective in spotting problems that may have occurred during the reduction.

The QC database has a web-based front-end that is used internally by CASU to rapidly search for images at a specific position on the sky, optionally satisying user-specified constraints for any QC parameter. The front-end is also complemented with an image cut-out service that is used to create on-the-fly object postage stamps and full-chip previews. Among other things these facilities are required for efficiently tracking down and assessing problem datasets reported back by users.

[13] Further details can be found in the WFCAM web pages at http://casu.ast.cam.ac.uk/surveys-projects/wfcam/

This page displays the reduction progress of WFCAM data. Information is automatically updated every hour (you need to reload the page).

Back to WFCAM start page.

Night	Status	N_{raw}	N_{ESO}	Checked	Transferred by WFAU	Version	Summary Plots	Photometry Plots	Summary Info	Observation Log	Size raw [Gb]	Size red [Gb]	N stacks All	N stacks UKIDSS	N stacks LAS	N stacks DXS	N stacks UDS	N stacks GPS	N stacks GCS
2006/04/28	reduced	476	361	23 Aug 2006	06 Sep 2006	1	GIF1 GIF2	GIF	summary.list	obs_index	8.86	16.13	39	30	0	0	0	30	0
2006/04/29	reduced	411	366	23 Aug 2006	06 Sep 2006	1	GIF1 GIF2	GIF	summary.list	obs_index	8.60	2.58	22	0	0	0	0	0	0
2006/04/30	reduced	145	133	23 Aug 2006	06 Sep 2006	1	GIF1 GIF2	GIF	summary.list	obs_index	2.01	1.68	10	0	0	0	0	0	0
2006/05/01	reduced	1055	794	23 Aug 2006	06 Sep 2006	1	GIF1 GIF2	GIF	summary.list	obs_index	20.46	35.09	126	69	42	0	0	27	0
2006/05/02	reduced	1290	139	23 Aug 2006	06 Sep 2006	1	GIF1 GIF2	GIF	summary.list	obs_index	24.99	50.84	47	0	0	0	0	0	0
2006/05/03	reduced	2402	640	23 Aug 2006	06 Sep 2006	1	GIF1 GIF2	GIF	summary.list	obs_index	32.26	47.46	44	0	0	0	0	0	0
2006/05/04	reduced	1564	182	23 Aug 2006	06 Sep 2006	1	GIF1 GIF2	GIF	summary.list	obs_index	29.42	60.45	60	0	0	0	0	0	0
2006/05/05	reduced	1639	116	23 Aug 2006	06 Sep 2006	1	GIF1 GIF2	GIF	summary.list	obs_index	25.64	44.36	37	0	0	0	0	0	0
2006/05/06	nodata	66	55							obs_index	0.69								
2006/05/07	nodata	0																	
2006/05/08	nodata	554								obs_index	5.57								
2006/05/09	nodata	768	84							obs_index	7.09								
2006/05/10	reduced	2516	1918	25 Aug 2006	27 Sep 2006	1	GIF1 GIF2	GIF	summary.list	obs_index	42.30	76.66	257	226	0	0	0	0	226
2006/05/11	reduced	2603	2553	25 Aug 2006	23 Sep 2006	1	GIF1 GIF2	GIF	summary.list	obs_index	49.60	128.29	357	330	2	0	0	107	221
2006/05/12	reduced	1841	1823	25 Aug 2006	27 Sep 2006	1	GIF1 GIF2	GIF	summary.list	obs_index	35.96	57.71	229	178	89	0	0	52	37
2006/05/13	reduced	1129	563	19 Sep 2006	23 Sep 2006	1	GIF1 GIF2	GIF	summary.list	obs_index	18.18	60.40	60	0	0	0	0	0	0
2006/05/14	reduced	2402	1424	19 Sep 2006	23 Sep 2006	1	GIF1 GIF2	GIF	summary.list	obs_index	46.56	95.81	111	30	1	17	0	12	0
2006/05/15	reduced	2674	1540	19 Sep 2006	23 Sep 2006	1	GIF1 GIF2	GIF	summary.list	obs_index	53.59	105.73	103	20	0	20	0	0	0
2006/05/16	reduced	148	145	19 Sep 2006	23 Sep 2006	1	GIF1 GIF2	GIF	summary.list	obs_index	1.85	1.93	15	0	0	0	0	0	0

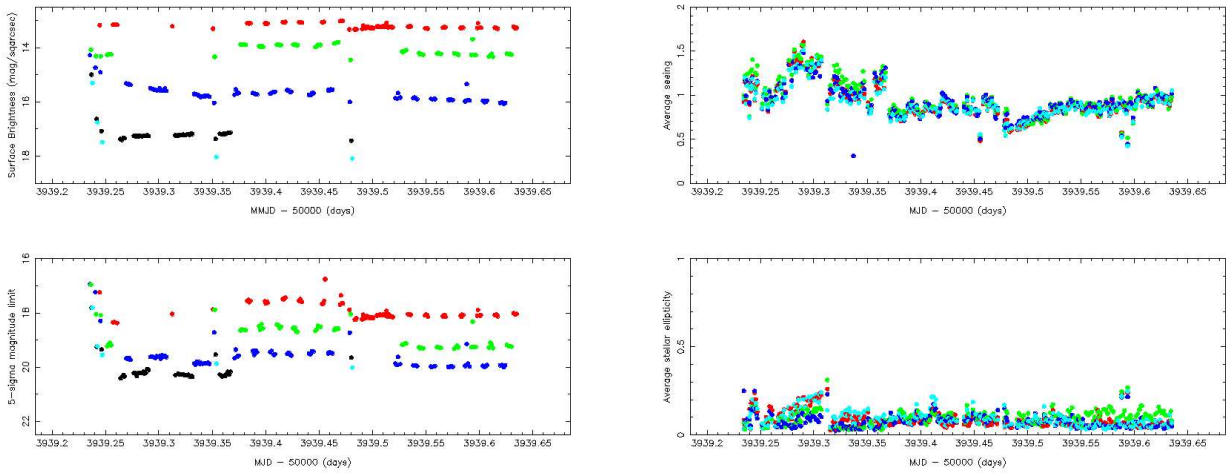


Figure 9: A snapshot of the reduction progress page for 06A and examples of the QC plots showing the average sky brightness, magnitude limit, seeing (FWHM) and stellar object ellipticity for each science product frame of a night.

The QC database is used also to keep track of the night status within the processing data flow, to flag missing fi les when the tapes from JACH are ingested and the MEF fi les created, to record every single fi le that was transferred to ESO to be ingested in their archive, and to automatically generate and update (hourly) user-friendly publicly-accessible web processing status pages grouped by semester.[14]

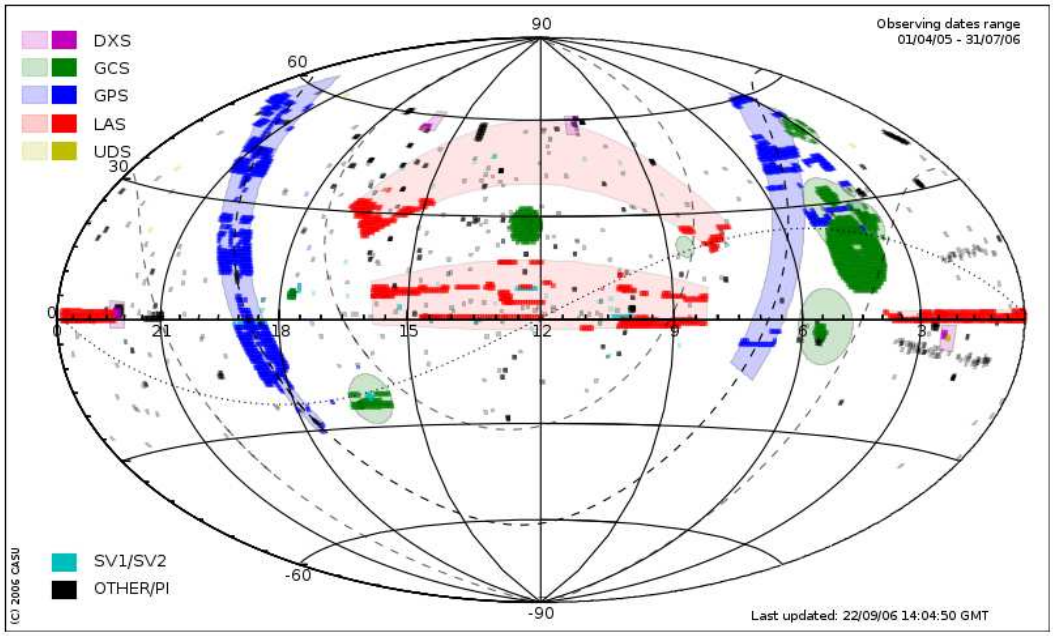
5.2 Survey Progress Tool

The WFCAM Survey Progress Tool makes use of the QC database to monitor the progress of the observations carried out by WFCAM and in particular is designed to highlight UKIDSS survey progress. The entry page[15] provides a whole sky overview of the planned UKIDSS observations and of the observations actually made, with relevant statistics like the number of MSBs completed, area covered and observing time elapsed (see fi gure 10). This page provides links to more detailed views for each UKIDSS survey composed basically of zoomed areas displaying only the information of that particular survey. The user can select several constraints on the frames to display and how to combine mutliple observations: *e.g.* selecting only those fi elds observed in the J, H and K bands and observed more than once in each of them; display frames with seeing within a range; display frames observed within a specifi c date range; or a combination of all these.

Also linked from the main page the QC view provides access to the distribution and evolution with time of different quality control parameters: seeing, ellipticity, magnitude zero point, magnitude limit and sky brightness.

[14] see http://casu.ast.cam.ac.uk/surveys-projects/wfcam/data-processing/.

[15] available at http://casu.ast.cam.ac.uk/survey-progress/wfcam/.



Survey	MSBs completed						Area covered (sq. deg.)						Observing time (hours)						Data Quality
	Z	Y	J	H	K	Total	Z	Y	J	H	K	Total	Z	Y	J	H	K	Total	Summary
DXS	0	0	409	0	837	1246	0.0	0.0	6.1	0.0	16.1	16.1	0.00	0.00	71.31	0.00	140.52	211.83	dqc
LAS	0	1998	2597	2471	2339	9405	0.0	387.8	531.9	488.9	480.7	697.0	0.00	27.73	42.99	32.64	30.86	134.22	dqc
GPS	0	0	1085	1046	3681	6234	0.0	0.0	180.5	177.3	677.0	684.2	0.00	0.00	22.09	21.27	40.42	102.54	dqc
GCS	549	521	516	730	2837	5153	105.0	105.0	105.0	145.4	586.1	597.4	6.10	5.79	5.73	8.11	31.52	57.26	dqc
UDS	8	0	32	5	50	95	0.9	0.0	2.6	0.2	2.8	4.1	0.39	0.00	6.10	0.44	10.01	16.93	dqc

Figure 10: All sky Aitoff projection showing the areas defined in the UKIDSS survey (light colours) and the areas actually observed and processed (dark colours).

As before the user has the option to display a range of dates, zoom in the plots or specify additional constraints using SQL language. [16]

5.3 Software development

PSF fitting

PSF analysis is normally an extremely CPU-intensive task and CASU have spent a lot of effort in minimising CPU overheads by optimising code performance without impacting on the quality of the final products. Many advanced survey products are predicated on the ability to automatically generate detailed well-sampled PSFs including: improving point source photometry and astrometry; adaptive kernel matching of PSFs for difference imaging for transient event detection (SNe, planetary transits, solar system objects); optimally stacking images taken at different times in (necessarily) different seeing conditions and for detailed survey completeness studies.

Tests of the CASU-developed PSF fitting software using UKIDSS GPS data from sample crowded fields and areas of varying nebulosity have been completed. These tests were carried out in conjunction with P. Lucas (University of Hertfordshire) who used DAOPHOT as a control comparison. The tests compared standard pipeline aperture and prototype pipeline PSF photometry with DAOPHOT PSF photometry. Results are mixed. In general the standard pipeline gives marginally better results but in specific cases, *e.g.* in the middle of globular clusters, DAOPHOT is better. In part this is due to how well the standard pipeline already performs using a soft-edged aperture method coupled with an innovative improvement to the astrometry algorithms (described in the UK review documentation set). The conclusion from these tests was that unless CASU can implement full iterative crowded field PSF photometry as a pipeline option, the improvements given by PSF

[16] Both the QC database and the survey progress tool have been designed with VISTA applications in mind and will form the basis of the equivalent tools for VISTA survey status monitoring.

development stages #1 (PSF photometric fits) and #2 (PSF photometric and astrometric fits) are extremely marginal and the only real benefit is more reliable error estimates.

The current development of the CASU PSF software has several innovative features and is being written up as part of a draft paper which will be submitted to MNRAS. Further PSF fitting software development is ongoing and will inevitably overlap the start of real VISTA science survey operations due to the complexity of the problem. Implementing the equivalent of a fully automatic DAOPHOT is challenging but much of the groundwork is already in place. Due to the success of the standard CASU pipeline products we envisage offering full iterative PSF fitting as an optional extra stage since most surveys do not require it.

List-driven photometry and DIA
The deployment of the list-driven photometry and difference image anaylsis (DIA) software is described in section 3.4. This was developed as part of the VDFS project and is essentially complete. Further upgrades and maintenance will be required but the bulk of the development work is finished as evidenced by the singularly successful use of this software for lightcurve generation and analysis in the Monitor project.

Completness estimation
Another area where software development has begun but will also overlap start of VISTA survey science operations is accurate estimation of survey depth and associated completeness of surveys as a function of magnitude and object type. CASU are devising an analytic approximation to this problem, since it is seriously CPU-intensive if attempted in the standard Monte-Carlo fashion. We intend to calibrate the analytic approximation using selected Monte-Carlo studies from the main protagonist surveys from WFCAM, VISTA and VST.

After the completion of the VDFS development funding any remaining aspects of this software development will be undertaken as part of the general operational support required for all CASU-supported pipelines thereby benefiting from their inter-related synergy.

5.4 WFCAM pipeline requirements

5.4.1 Hardware

Operational costs for WFCAM are minimised by basing the operational pipelines and bulk storage on commodity PC and disk RAID array products. For on-line storage, RAID5/6 disk systems are being used with a mixture of external storage devices (made up of a disk controller populated with SATA/IDE disks but accessible through a SCSI interface) and internal ≈1 TB RAID arrays within each processing PC. The raw CPU power is currently provided by an array of twin processor Xeon PCs (3GHz) with sufficient internal memory (currently 2Gbytes) to hold full dither stacks and standard mosaic "tiles" in memory. For both the PC processors and disk storage systems we hold spares (≈5-10%) in lieu of maintenance, since these provide a much cheaper and faster response alternative. For the foreseeable future there is no immediate requirement to further expand the available WFCAM processing CPU.

For off-line backup and storage we currently use, and propose to continue to use, highly cost-effective LTO-II/III tapes (200/400 Gbytes native capacity), and a 17-shot library loader (3+ TB native capacity). These have a well-defined upgrade path and provide a cost-effective means for backups and off-line curation. LTO-I/II's are the current medium for transport to the UK of WFCAM data. By using lossless Rice compression a huge cost saving results in data transport, I/O overhead and storage. WFCAM on-line storage requirements (raw and processed data) are growing at the rate of around 25 TB/yr. Costs of suitable scale commodity RAID systems have steadily fallen over the last few years such that the current best buys work out at ≈£1.0k per TB (all-in) and are expected to continue to improve at a similar rate.

With on-line storage in place at two sites we do not foresee the need for further (expensive) near-line "tape" storage systems due to the use of high density tape transfer media (LTO-I,II) from Hawaii. If recovery from the raw tape data (archive) is required this would be feasible with our proposed backup on-the-shelf tape storage system. We have also begun investigating Internet transfers from JAC to take advantage of the recently

improved (1 Gbit/s) connectivity between Hilo and the summit. However, if this proves successful we will still need to create a local tape off-line copy.

Data transfers between Cambridge and Edinburgh are via the Internet currently over two networks, JANET and a private UKLight network both with bandwidths 1 Gbit/s, more than sufficient to transfer the processed (Rice-compressed) data products. We have investigated alternative use of a private network (at fixed annual cost) because the IoA are charged for pro-rata use of the University network connectivity. We have split the costs for network use since not all data can be shipped using the private network. (annual cost 2k/yr). Over the first two years of WFCAM operation, before our UKLight connection was running, the CASU contribution to the IoA component of University network costs has averaged ≈£10k/year.

5.4.2 Personnel

Operational tasks include: managing and maintaining the computer hardware and system software; data ingestion and verification; running the standard pipeline and monitoring data quality; running the further processing pipeline and monitoring data quality; liaising with external users on science verification issues; fault finding and reprocessing as required; visual inspection of random samples of data products; carrying out statistical tests *e.g.* completeness and comparisons; updating and enhancing the control systems for the pipelines; maintaining a WFCAM raw data archive and ensuring the delivery to ESO of the raw archive and to WFAU of the processed data products; continued development of operational infrastructure and quality control monitoring.

There is also a significant staff overhead involved in liaising with external users and survey PIs (UKIDSS), holding regular minuted meetings, reporting to various committees (VDMT, VDUC) and Boards (UKIRT); reporting and giving presentations to extenral review bodies; and in providing a helpdesk for answering queries. We also regularly carry out our own internal scientific assessment of data products to ensure their quality.

To fulfill these requirements we are requesting continued funding for WFCAM data processing operations from 1st April 2008 at the level of 1.5 FTE/yr. We have found this to be a necessary level to maintain and run a successful WFCAM operations project and to be responsive to the needs of external users and committees.

5.5 Summary of resources requested

Staff:	1.5 FTE/yr from Q2 2008
Travel and subsistence:	guideline amount
Equipment:	processed and raw data storage and curation hardware (£80k total)
Consumables:	£5k/yr internet costs (£25k total)
Maintenance:	£2k/yr for CPU processors for the whole period (£10k total)

5.6 Key deliverables and milestones

2008–2013	Full pipeline operations, data ingestion, verification, fault finding, science verification, onward delivery of data products to ESO and WFAU
2008–2013	Annual pipeline s/w releases to JAC (JAC QC pipeline)
2010–2013	Full Iterative PSF photometry
2008–2013	Continual QC monitoring and feedback
2009–2013	Annual major UKIDSS/WFCAM data releases, product packaging, QA image access

6 CASU Science Services

CASU is the major provider of advanced science data products from key UK ground-based facilities.

Recognising that wide-field optical and near-infrared mosaic camera systems are sufficiently similar in their technical properties and user science requirements to benefit from a common pipeline processing strategy, CASU has been able to devise a general purpose data processing framework and apply it to a variety of survey camera systems. CASU have demonstrated the practicality and cost-effectiveness of this approach (noting that the average astronomer may not possess the technical capability or the requisite computing infrastructure to process large, multi-TB, amounts of imaging data). Thus, for instance, CASU produces the reduced science data products for the ING's INT Wide Field Camera, both survey and PI data. Support is now requested to allow for the provision of science data from an increased range of stratigically important missions, including the creation and distribution of science data products from major UK led ESO VST legacy survey programmes (including the VPHAS and ATLAS programmes).

CASU and the IoA, Cambridge are major partners in the AstroGrid Virtual Observatory (VO) project. AstroGrid has developed the Virtual Observatory service infrastructure which will become operational for the UK community from 2008[17]. CASU is a major partner in the emerging UK Data Centre Alliance, which in turn provide the key data and astronomy services to the UK and the European community through the VO infrastructure. The effectiveness of the VO is maximised with the provision of high quality and relevant science data products and services. The Science Service activities outlined here will ensure the provision of a range of high quality science data products to the UK community, essential in enabling the maximum scientific return from the UK investment in these facilities (such as the ESO telescopes).

In partnership with the AstroGrid activities within the IoA, CASU has been involved in providing the first generation of new science data products specifically tailored for access through VO standard protocols. The VO paradigm offers the opprotunity to dramatically reduce the overheads in publishing science data products. By utilising the infrastructure that AstroGrid provides the archive centre need not develop the normal interface software required to support end user access to the data products. For instance, authentication, data staging, the handling of remote database queries, etc are all handled by the AstroGrid software.

CASU is releasing (in July 2007) the Early Data Release of the IPHAS northern galactic plane survey. The main data product is the unique photometric object catalogue, containing some 200 million objects over the 2000 square degrees of the survey are (the northern galactic plane with $|b| \leq 5$ deg). Ancillary tables cover pointing and quality assurance information. This work has been carried out with input from the IPHAS[18] survey team, and the AstroGrid project. The database is accessible to any user of the AstroGrid system and allows full data queries (SQL) or simple RA, Dec searches. The data is held within a Sybase ASE RDBMS, the table size is some 200GB, and a quad dual core Intel Xeon unit provides the processing power. The publication of this data product has been extremely cost effective compared to using legacy techniques.

CASU hosts a number of key science services for the VO. These include standard image access to the Wide Field Survey and IPHAS image collections as well as access to a range of legacy data archives. In addition a number of science applications are hosted such as Sextractor and the STILTS[19] table manuiplation library and the WFC processed data interface at http://apm2.ast.cam.ac.uk/cgi-bin/wfs/dqc.cgi which allows remote execution of various software tools such as adaptive kernel difference image processing, stacking and mosaicing, and band-merged catalogue generation.

From 2008 we propose to transition CASU from its current role as a traditional Data Archive Centre to a fully, Virtual Observatory compliant, Science Service Centre (CASU-SSC). Four main strands of activity will occur: maintaining and upgrading access to legacy ground-based observatory data archives; providing key science

[17] Note that AstroGrid has been funded to provide the infrastructure, not the provision and support of the data and science services outlined here.

[18] see http://www.iphas.org

[19] www.star.bris.ac.uk/ mbt/stilts/

data services (thus leveraging the science data pipeline outputs discussed earlier); developing relevant core VO science applications; and providing a range of data quality assurance services which enable end users to verify the quality of the data products from the CASU mission sets. These activities are a vital complement to the continuing provision of the long term curation of mission data archives.

We note also that Cambridge is the lead partner in the STFC supported Gaia Data Flow System project. This is developing the photometric pipeline analysis system to handle the large scale Gaia data. From 2012, with the launch of the Gaia mission, the ground segment will move from development to a post-launch operational phase. Cambridge will host the primary Gaia photometric processing centre (CASU-GDPC), for which a fully costed proposal to STFC will be submitted in due course. A number of the activities that CASU will be pursuing over the period of this grant (2008-2013) will have relevance to the forthcoming CASU-GDPC, for instance in establishing the necessary interface components to allow for efficient publication of science alert photometric data in a timely fashion to the community.

6.1 Science Data Services

The core activity here will be to provide seamless end user access to a range of key high quality science data products, primarily those being generated by the CASU pipelining activities. Effective, and economic access will be possible because all data will be published utilising VO protocols, as provided by the AstroGrid VO infrastructure project[20].

The development of the CASU automatic processing pipelines was a major challenge, covering not ony standard image processing tasks but also a large number of essential extra tools including: recognising, monitoring and correcting for inherent non-linearities; automatic defringing and residual background correction; object catalogue generation with optimised object parameter descriptors; automatic incorporation of World Coordinate System (WCS) information; automatic photometric calibration; hands-off object morphological classification; merging catalogue information from different passbands and optionally different instruments to enable simple generation of multicolour information; the development of stacking and mosaicing tools; and the development of automatic quality control (QC) measures that characterise the integrity and usefulness of the observations.

This investment in CASU's unique and world class analysis pipeline can now be leveraged to provide high quality data products for a range of mission data as described now. All data products described also have attached quality control measures and are fully astrometrically and photometrically calibrated. Astronomical object catalogues are generated by default and if requested, data from different passbands and/or epochs are merged for use in, for example, producing colour-colour or colour-magnitude diagrams.

Activities here provide the operational science data and applications which form a vital component of the Virtual Observatory in the UK. Appropriate s/w developed by the AstroGrid project will be deployed, such as the Data Set Access component as a means of publishing large database held resources (such has been demonstrated now with the early IPHAS data release, and planned for by the ALMA regional data).

The main science datasets that will be created and made available to the UK and wider astronomical community include:

VST: pipeline operations and science data product access [2009-2011]: CASU has reached agreement with the key UK led ESO VST legacy survey teams (VPHAS - PI: Drew, ATLAS - PI: Shanks) to provide the pipeline

[20] Briefly the key VO software components which CASU will use in publishing science data and applications are:

1. Data Set Access (DSA): this component allows simple (RA, Dec, size) and complex (SQL) queries of data held within relational databases.
2. VOSpace: allows for authorised data staging, thus allowing end user long running asynchronous queries.
3. DAL (Data Access Layer) toolkit: providing VO compliant SIAP (Simple Image Access Protocol) and SSAP (Simple Spectra Access Protocol) interfaces to expose image and spectra data (respectiveley) held as flat file collections.]

analysis system for the programmes[21]. ESO will take eventual responsibility for the distribution of the reduced data products (thus images, per pointing catalogues). We propose here to provide a UK point of access to the image and flat file tabular data. In addition, building on the successful experience with the IPHAS combined object catalogue, we will provide a combined object catalogue (exposed via AstroGrid/ VO protocols, and managed by a RDBMS) of these VST surveys. Tasks to be undertaken will include:

1. VST science requirments capture
2. transfer of the VST processed data files from the staging area post pipeline to the long term storage area. Ordering the image collection, configuration of the SIAP access to the image collection.
3. generation of the database schemas and ingression of the per tile fits object catalogues to the RDBMS
4. generation of the Hierarchical Triangular Mesh (HTM[22]) index for all objects and generation of table indices
5. configuration of the DSA component accessing the VST catalogue RDBMS, generation of appropriate descriptive metadata describing the data products, and the publication of this to the CASU local VO registry (which is turn is uploaded to the main UK AstroGrid central registry).
6. quality assurance of the key science products and ongoing monitoring of operational usage, responding to faults (e.g. disk failure/ replacement)

Science access to the IPHAS/EGAPS object database and image collection [2008-2010]: The IPHAS survey (PI: Drew, see http://www.iphas.org) is a large public access multiwaveband survey of the entire Northern galactic plane covering some 2000 sq degrees in broad band r, i and the narrow H-alpha band. As described above in section 3, the initial data release will be a catalogue of some 200 million unique objects, this representing a data resource comparable in size to the SDSS catalogues.

Over the period of 2008-2010, the IPHAS consortium will be integrating the UVEX survey of the northern galactic plane which will add u, g, and r data to match the IPHAS data collection. Further the VPHAS survey of the southern galactic plane has been accepted as a ESO VST Legacy Survey programme (PI Drew). This will map the entire southern plane to complement the IPHAS survey. CASU will provide the database for these additional data images and catalogues, providing community access to some 25TB of image files and a 400 million object catalogue in 5 colours (a resource comparable to the entire SDSS). The tasks to be undertaken are as above for the VST.

Quality assurance VO interface to the VISTA image collection [2009-2011]: Simple image access will be provided to the VISTA images. This will complement access available through the VSA, providing additional access capaciity to what will be a primary data resource for the UK community. Additionally it is planned that images will be available in 'Pre' form for early usage by the community, before the regular six monthly full uploads to the VSA. Image staging at CASU will further improve efficiency when accessing the CASU hosted science services (minimising the amount of data movement over the SuperJanet network)

The tasks to be undertaken are as above for the VST, except that there will be no catalogue creation. This work complements the activities of the IfA, Edinburgh Vista Science Archive activities, which focuses on the curation of the VISTA database access to the object catalogues.

Quality assurance VO interface to the UKIDSS image collection: [2008-2010]: Simple image access will be provided to the UKIRT WFCAM images, both UKIDSS survey and PI programmes. This will complement access available through the WSA, providing additional access capaciity to what will be a primary data resource for the UK community. Additionally it is planned that images will be available in 'Pre' form for early usage by the community, before the regular six monthly full uploads to the WSA. Image staging at CASU will further improve efficiency when accessing the CASU hosted science services (minimising the amount of data movement over the SuperJanet network)

[21] Part of the ESO requirements for these proposed surveys is that the various groups have to make available fully processed and calibrated data.

[22] http://skyserver.org/htm/

The tasks to be undertaken are as above for the VST, except that there will be no catalogue creation. This work complements the activities of the IfA, Edinburgh Vista Science Archive activities, which focuses on the curation of the VISTA database access to the object catalogues.

The Wide Field Survey and INT Wide Field Camera PI data products [2008–]: The INT WFC survey is complete, with several TB of raw data processed, and catalogues for several hundred million detected objects over thousands of square degrees of sky generated. Additionally CASU have provided a data processing service for other INT WFC PI data, since few users are able to handle the complexities of processing such mosaic camera data at their home institutions. All data products have attached quality control measures and are fully astrometrically and photometrically calibrated. Astronomical object catalogues are generated by default and if requested, data from different passbands and/or epochs are merged for use in, for example, producing colour-colour or colour-magnitude diagrams. This facility has continued to be heavily used, both for major programmes such as the IPHAS suvery described above, and for smaller programmes.

These science data products will be made available to the wider community through authorised access, with 'public survey' reduced images and catalogues available to all, and other PI products available with PI consent.

Processing for Time-Domain Astronomy: [2009–]: Rapid, robust, near real-time processing of the vast amounts of data that will be generated from dedicated time-domain monitoring telescopes is a challenging task. The requirements on a data processing system are tough: almost complete automation is essential; the algorithms must be both robust and optimal, since photon noise limited photometry is the goal; and the data is often severely undersampled with corresponding intra-pixel sensitivity issues (a very similar problem to Gaia data processing).

Hodgkin and Irwin are involved in the data processing development for two such projects: the Wide Angle Search for Planets (WASP), a UK university consortium telescope designed to enable a census of variability across the optical sky; and the 0.5m Automated Patrol Telescope (APT) in Australia whose primary goal is to search for extra-solar planets via their transits of the parent star. Both projects have almost identical data processing problems in that they are multi-CCD wide-area, heavily undersampled ($\approx$10 arcsec per pixel) data capture systems. Both suffer from intrapixel sensitivity problems, which without special software, currently limits the APT accuracy to 2-3% at the bright end, and both will produce TB of data each year.

Hodgkin and Irwin are also involved in other related ambitious long term monitoring proposals. The first of these is designed to carry out extra-solar planet transit searches using the ESO 2.2m WFI mosaic camera. This complements the classically styled shallower wide area APT/WASP searches with much deeper monitoring over a subset of these fields. The second is a novel proposal to intensively monitor the Orion Nebula Cluster (ONC) and associated star forming region to look for eclipses caused by planets, or planetesimal aggregates, forming in orbit around young stellar objects. The third project (as noted in section 3.4) will use already awarded CFHT and CTIO time to take high cadence photometry of tens of thousands of very low mass stars and brown dwarfs. All of these projects require the application and further developement of VDFS data processing tools.

Existing results from processing APT data through a slightly modified version of the CASU optical imaging pipeline are very encouraging (see *e.g.* Hidas *et al.* 2004). A photometric precision of 0.2% was attained for 9th magnitude objects rising to only 1% error by 13th magnitude, using what are essentially general-purpose CASU pipeline components. Tuning of the pipeline for this particular problem is still under investigation but the early results are very competitive.

There are several inter-related factors, mainly systematic effects, that currently limit the attainable bright-end precision which we propose to further investigate. In addition we are using the transit search data to further trial some of the PSF–combinatorial techniques (*e.g.* difference imaging) that will be required for several of the advanced processing options for VISTA/WFCAM data. We anticipate the amount of software development required for these systems to be at the level of 0.25 FTE/yr for the next few years but as this has a synergy with VISTA/WFCAM requirements have subsumed this within the VDFS VISTA/WFCAM software development.

Future Survey Support [2009–]: Further in the future are the possible processing requirements for the two

suggested dark energy cameras, VISTA Visible and/or the proposed Dark Energy Camera on the CTIO 4m, which have also expressed interest in CASU processing pipelines.

Likewise, preliminary exploratory feasibility planning is underway considering the use of CASU pipeline and processing techologies in meeting the needs of emerging large scale radio data missions. For instance, in the run up to SKA, processing of MeerKAT data. In particular, cross federation of UKIDSS and VST survey data with the MeerKAT HI 2000 sq deg survey will be vital in studying the interstellar medium in the local uiniverse. In conjunction with members of the Cavendish Laboratory and colleagues in South Africa is now being investigated, and may result in an additional funding proposal at a later date.

CASU will actively engage in investigating opportunities to leverage its knowledge and capabilities to support the processing needs of new missions, to the benefi t of the UK community.

6.2 Science Analysis Services

An important element of the proposed work in the period 2008-2013 will be to support the development of VO science services to the community. These represent the deployment of application services that will be provided by CASU and thus available to users of the AstroGrid UK-VO system.

A range of core services required by the astronomical community will be deployed as VO science services, where the service runs and stages data on CASU computational facilities.

The key initial services that will be deployed in the period 2008-2009 are listed below. Work from 2010 will focus on providing extended science services reflecting demand from the user community.

1. *On demand catalogue creation service:* CASU has a long history of the development of high quality image transformation and analysis software. For instance 'imcore' is a highly sophisticated source extraction algorithm which forms the basis of the catalogues creation element of the CASU astronomical pipeline software. Here it is proposed to exploit standard VO application interfaces (in this case utilising a standard IVOA UWS-PA interface) such that the 'imcore' routines are available as general services to any user of the UK or external VOs. 'imcore' will be confi gured though a standard VO appicaltion interface (UWS-PA compliant) which means that it will be available as either a stand alone task, or as an element in a workflow or script.
2. *Mosaic image service:* AstroGrid has currently implemented a pilot mosaicing service, where an interface is provided to the Montage application running on a 64 CPU node cluster at IPAC, Caltech. This allows for a UK VO user to create on demand mosaics of 2MASS image data up to 25 square degrees in size, with the result being automatically computed and transferred to their UK VOSpace storage. This service will be extended to allow on demand mosaic creation of CASU hosted science images. Work in 2008 will extend this service in two areas. One to allow for the creation of mosaics from CASU image survey data fi les (e.g. INT-WFC, VST, UKIRT-WFCAM, VISTA) and combine these with IR data held at IPAC (2MASS, Spitzer) to generate on-demand multicolour mosiacs
3. *Redshift and Cluster Analysis Service*
   This service will expand upon the current AstroGrid prototype. A number of redshift methods will be confi gured on the CASU compute cluster, such as ImpZ, LePhare, ANNz. This will allow a user to have the redshifts of objects in their multi colour photometric survey data from the current 'best in class' methods, in one easy step. Add ons will be available to enable the automatic determination of cluster and two-point correlation functions, of wide use in the intepretation of large scale galaxy surveys.
4. *Science service application provision*
   Continued effort at the level of 0.25 FTE/yr will be devoted to providing further science service applications in the period 2010-2013. These will be selected to support key survey science undertaken by the UK community *e.g.* service required in support of the VISTA Hemisphere Survey (PI: McMahon).

6.3 Science Quality Assurance Services

The scientific value of all science data products is enhanced through clear quality assurance procedures being in place as the raw data is transformed into calibrated and value added science product. CASU plan to make available a number of 'quality assurance' services to the community which will be of benefit, from defining suitable selections of data (via cross link the quality control database to the science products) to planning new observations based on knowledge of fundamental observational parameters such as varying sky brightness.

Survey progress Quality Control database: There are clear requirements for automatic real time monitoring of survey progress via QC parameters with minimal user input using online databases and the provision of simple visualisation of overall survey progress. The CASU-processed INT WFC data has been used to refine and progress this aspect of pipeline processing via a data products database. FITS header information, including QC parameters, are ingested on a regular basis, and a series of flat files point to the processed object catalogues and image data. This has allowed us to assess the problems of day-to-day running of operational pipelines and to explore the practicalities of controlling optional further processing stages driven from a survey QC database[23].

The database interface supports a large range of user options including: on-the-fly multipassband merging of catalogue data and optional federation with other catalogues such as FIRST and 2MASS; production of target "postage-stamps" for candidate selection or for making finding charts; interactive interrogative catalogue overlays on images; uploading of lists of target objects; automatic retrieval and downloading of catalogue and image products; the ability to group and remotely process images using CASU facilities including the CASU VDFS image subtraction, image stacking and image mosaicing software utilities and retrieval of the results *etc*. In particular, these latter advanced image manipulation tools, have seen heavy use by members of the IPHAS consortium.

Work in the period 2008-2010 will enhance the utility of the quality control information available. It will be used for automatically updating survey progress web pages. Through cross linkages with the science data products, the QC information will allow finer grained user selection of appropriate data (based on complex selections of for instance observing conditions, depth of data etc).

Sky brightness from NIR surveys: We have been using the WFCAM quality control database (see section 5.1) to investigate the NIR sky brightness properties using UKIDSS data. The sky level is one of the quality control measures automatically produced by the pipeline during the reduction for each chip image extension. This worked started as a simple data quality control task, but it actually developed into a fully independent research project aimed at monitoring the properties of the NIR sky. Although there are several papers discussing the properties of the atmosphere from a physical point of view, surprisingly enough, there are no published studies on the NIR night sky properties from an astronomical point of view.

For our study we used the WFCAM data collected during the first $\sim$ 18 months of operation in the framework of the UKIDSS surveys. Results from this project have indeed proved extremely useful in the planning phase of the VISTA public surveys. We investigated several aspects of the NIR sky in the five WFCAM broadband filters (YZJHK): (1) dark- vs. bright-time sky brightness; (2) variations of the sky level as a function of the angular separation from the Moon; (3) variations of the sky level as a function of the elapsed time from the twilight (see Fig. 11).

We foresee three lines of research in the study of the NIR sky brightness. First we plan to include more data to increase the time baseline and look for long-term trends and for possible correlation with the Solar cycle. A second line of investigation would be to exploit the fine temporal sampling provided by NIR observations to characterise the sky behavior on short time scales. Finally the third line of research would be to carry out the same analysis using VISTA data and investigate the Paranal NIR sky compared to that on Mauna-Kea.

[23]The WFC interface is accessible through http://apm2.ast.cam.ac.uk/cgi-bin/wfs/dqc.cgi with registration required for more compute intensive use of CASU facilities.

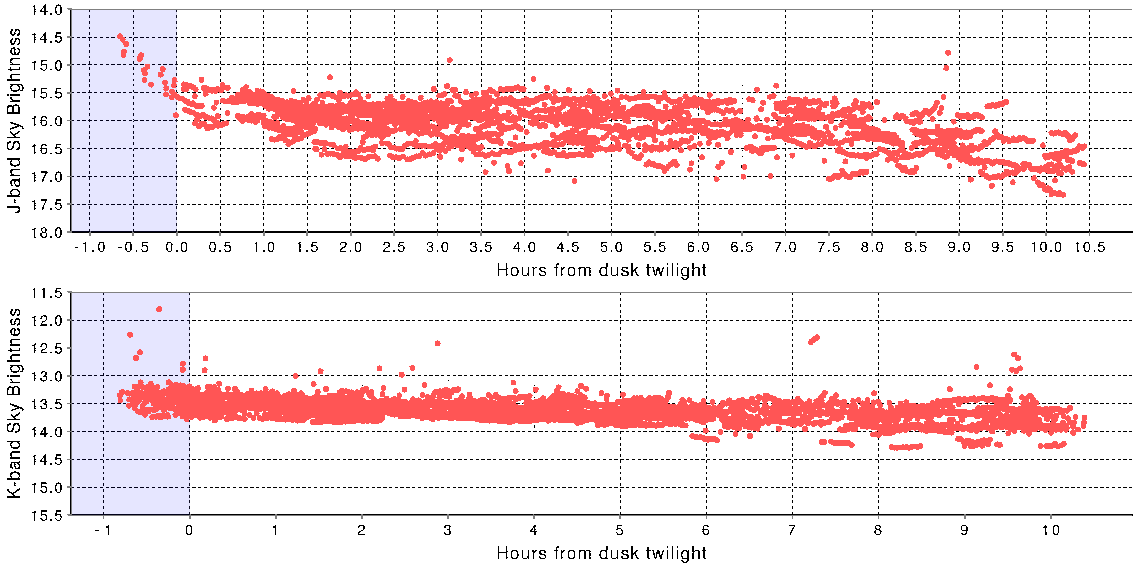


Figure 11: Sky brightness evolution versus elapsed time (hours) from the dusk twilight for J and K. Shaded regions indicate evening twighlight.

6.4 The CASU Data Centre Archives

Science archives underpin VO activities and data curation is a vital and aspect of modern astronomical research. Archive data sets have a long-term legacy value that stretches far beyond the initial operational and exploitation phase. In recognition of this, CASU have been making use of affordable TB-scale on-line storage systems to place much of the UK ground-based archives online. Although the main immediate benefits are automatic and rapid access to all the data, the ability to properly organise and categorise the data opens up opportunities for providing on-demand association of calibration information and basic processing options. The latter is an important element of the VO concept.

The main components of the CASU Data Centre are the UKIRT, ING and AAO archives. We discuss the progress with each of these in turn, together with a summary of current activities and future plans. In addition to these larger archives, there are many smaller object catalogues (*e.g.* APM, UCAC, 2MASS, FIRST *etc*.) incorporated within our Sybase server which, are updated and maintained as they become available.[24]

6.4.1 The AAO archive

An automated mirror of the AAO observation archive was set up some time ago and data were transferred to CASU from native storage at the AAO on DVDs which were added to the CASU DVD tower. The process of transferring the older data onto DVD at the AAO encountered unforeseen difficulties resulting in large chunks of missing data and unacceptable delays in copying newer data. Resolving these issues has taken a great deal of time and delayed opening the archive for general access until recently.

Tests of transferring data directly from the AAO via the Internet have been very encouraging and it is now realistic to copy most, if not all, the AAT data using this approach. Much of the missing data are now being supplied to CASU via the Internet and in future a backlog in supplying current data should no longer occur.

At present most of the raw AAT data from 1990 to early 2007 are on disk in Cambridge (≈1.5 TB) and are directly available through the archive interface. There are still roughly 0.3 Tb of data that have not been transferred to CASU. These are mainly data from the wide field imager (WFI) and are due to be transferred to CASU shortly. Requests for AAT data are made in a similar manner to those for the UKIRT and WFCAM archives. However, as the AAO archive has only just recently come on line there are few statistics currently available on archive usage.

[24]For a complete list of maintained catalogues consult http://archive.ast.cam.ac.uk/.

Of particular interest is the 2dF data from the AAT and the data from the AAOmega spectrograph. Running a combination of on-demand 2dF and AAOmega pipelines and providing access to a "best efforts" spectral libraries in collaboration with the AAO, would form a valuable component of an International Virtual Observatory. We have set up close links with the AAO to investigate this.

6.4.2 UKIRT and WFCAM archives

The UKIRT archive contains all of the observations taken with all UKIRT instruments with the exception of WFCAM.[25] CASU has been responsible for running the UKIRT raw data archive since May 1999. Catalogues of observations are logged locally at JAC on a Sybase server and are mirrored at CASU on a Sybase server. Periodic shipments of raw data are made from JAC to CASU on tape (DLT/LTO). These are ingested, converted from the UKIRT native sdf format to FITS files and stored in online RAID disk systems. The speed of download depends on the amount of data requested, but typical requests are now be filled automatically in a matter of minutes.

The UKIRT raw data archive currently consists of approximately 2.2Tb of data covering the time period July 1994 to October 2006. No data past that point are available as WFCAM has been on the telescope since late October 2006. Since taking over the UKIRT archive CASU has filled 663 requests from 188 users in over 30 countries. So far 1.8 million datasets have been dearchived and sent out. The companion WFCAM raw data archive consists of all the observations taken on UKIRT with WFCAM. As the data volumes are so high and since all data from WFCAM is pipeline processed at CASU, it was anticipated that WFCAM raw data would have a limited audience and so should be kept separately from the rest of the UKIRT raw data.

The WFCAM raw data resides on its own set of RAID hard disks so requests for data can be fulfilled as soon as they are received. The archive currently contains over 820000 observations taken since March 2005. This is equivalent to about 55 Tb of raw storage volume, but with use of the Rice tile compression algorithm we save an average factor of 3-4 in disk space. So far over 22000 datasets have been downloaded for 94 requests. The requests have come from 15 different users, mostly from the UK.

Ingestion and curation is part of the WFCAM operations effort, so no separate funding is required. A web interface for monitoring the ingest of WFCAM raw data is provided as part of the WFCAM quality control database (see section 5.1). The UKIRT archive effort is such a small perturbation on the WFCAM activities that we have absorbed any hardware and software requirements within WFCAM operations.

6.4.3 ING Archive

The ING data archive has been built up now over two decades, and forms an important asset for the UK and wider international astronomical community. The archive contains practically all observations collected with ING telescopes since 1984 and as such is the oldest and most extensive database of ground-based astronomical data available. The collection represents a huge investment over the years, and the data itself, as the observatory's prime "legacy product", must be adequately retained for future use. The scientific value of the archive has been amply demonstrated, and it continues to be heavily exploited by a wide range of users.

A web-based user interface, available since 1996, allowed for rapid scanning of the observation catalogues by users, but still required the requests to be filled by loading the relevant tape or disk volumes by the Archive Manager. Between May 2003 and 2006, the whole of the existing archive was transferred to disk. This involved reading 6177 tapes, 11175 CDs and 2926 DVDs, checking the data integrity and re-structuring the storage from a media-based to a date-based system. This processes included all data up to the end of 2004. From the begining of 2005 all archive data transfers from La Palma have been streamlined using automated Internet transfers. After data integrity checks are complete a monthly off-line LTO tape backup of them is made and the data Rice tile compressed and transfered to the live on-line ING archive RAID disks.

[25] With the advent of WFCAM JAC decided that WFCAM observations should be logged separately from the data observed with the normal cassegrain instruments. The WFCAM raw archive is described separately.

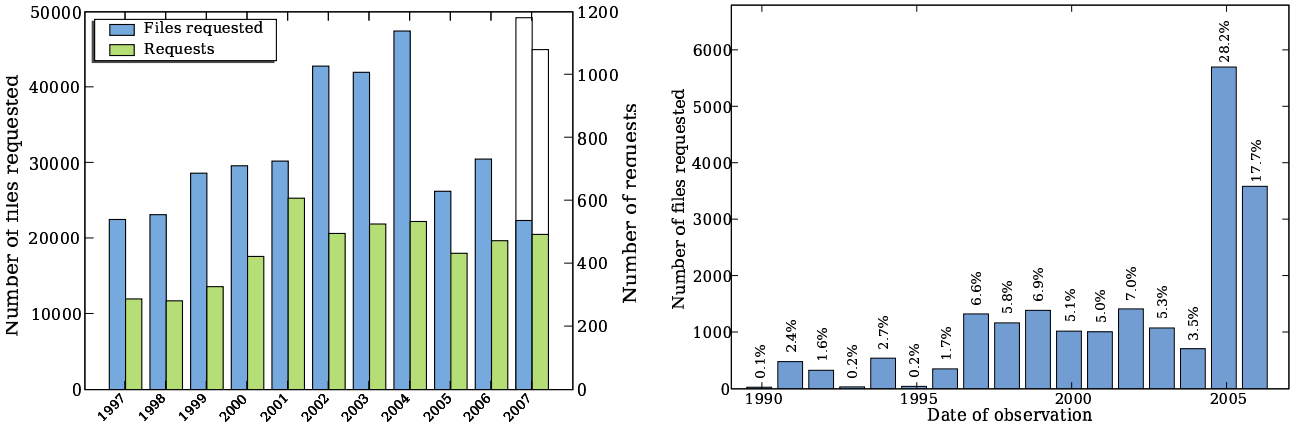


Figure 12: Left: number of files requested and requests per year. Note that the blocked 2007 figures are for the first 5 months only while the outlined totals for 2007 are the pro-rata predicted figures. Right: number of files requested vs. observing date since October 2006.

In compressed form, the complete ING archive currently occupies some 20 TB of disk space, which corresponds to approximately 40 TB of actual raw data. The archive is growing at the rate of about 3–4 TB of raw data per year. On-line storage and the re-factored date directory structuring, combined with the new web interface (see below), has reduced turnaround times for requests from hours, or days, to just a few minutes.

The exploitation of ING archive data continues to grow with a total number of 790 users in 51 countries making 4880 requests since its inception. This required 350000 files to be delivered on-the-fly (see figure 12). These are optionally available as Rice compressed, gzipped, or uncompressed files using either wget over http or interactively via a browser.

The original archive web interface, which had been in use since 1996, was was becoming increasingly out of date and difficult to maintain. With the push toward complete on-line storage nearing completion a new archive interface was designed built and tested to take advantage of the easier access. This interface was officially opened for general use in October 2006 and implements the standard ING proprietary access period for downloads.

As with the previous interface the user has to register with the system once to download data. All information about each request is stored in a database and is available to the user so he/she can view the request history or re-request datasets. Once a request is made the user can follow its status (pending, processing, completed) via the web interface. On completion the user receives a confirmation email with instructions on how to download the data. Since all the datasets are now on disk, the system is completely automatic and response times have been vastly improved. Virtually all requests are completed within a maximum period of 30 minutes but most take less than 5 minutes.

Since data are transfered from La Palma by the Internet, after internal disk transfer from the incoming disk buffer and verification, the files are immediately available on disk and are ingested in the database. Following the one year proprietary period the data are public from the interface (although PIs can request immediate access if required) and as can be seen from figure 12 the requests for recent data form a large part of the demand.

6.4.4 Future enhancements

Successfully managing the ING archive is now a much less labour-intensive task than previously but routine operational effort is still needed. This includes monitoring new data ingest, verification and correction (in conjunction with ING staff), administration tasks including dealing with user queries and procuring, managing and maintaining the hardware required for on-line disk storage. There is still also some work to be done in correcting and updating the FITS headers of the earlier archive entries to make them compatible with the majority of the archive.

All the archives apart from the ING archive have user interfaces that would benefit from an upgrade. We propose to use the ING archive interface as a template for upgrading the UKIRT and AAO archive interfaces. This is a relatively modest one-off upgrade that will provide vastly increased functionality and offer users a more uniform look and feel for the entire CASU raw data archive collection.

Future enhancements that have been requested by users include: provision of access through a VO-compliant interface; an image previewer option to facilitate file selection; and the ability to appropriately link calibration frames to science frames with the further aim of providing standard on-the-fly calibration at the image processing level (*i.e.* bias- and overscan-correcting, trimming, and flatfielding).

We have scoped the effort required to do these based on experience with running operations for WFCAM and will progressively phase in these enhancements to tie in with the increasing demands of the VO.

6.5 Software development

*O*verlap calibration: A common problem for uniform photometric calibration of optical surveys is the paucity of reliable faint optical standards. In contrast this is not an issue for the NIR surveys due to the availability of 2MASS which provides a sufficiently dense and accurate photometric backdrop at the level of 1–2% over the entire sky. In the Northern galactic cap the optical problem is mitigated by the availability of SDSS. Indeed CASU have been using this to help calibrate INT WFC surveys. However, SDSS does not currently cover, for example, the IPHAS survey of the northern galactic plane (nor the upcoming VPHAS+ southern equivalent).

The remarkable discoveries from 2MASS and SDSS have demonstrated the utility and importance of achieving an overall calibration error for surveys at the 1-2% level as a crucial step toward full exploitation of the data products. At heart, this task demands examination of overlaps between adjacent fields in order to establish a network of corrections to a common photometric level. We intend to apply this first to IPHAS and then role out the technique to VST and other optical surveys.

Calibration of narrow band photometry (e.g. H-alpha survey's) will require a bootstrapping technique combined with spectrophotometric flux measurements of standard phtometric field stars (Landolt, Sloan). The final solution will require significant quality assurance in terms of zero-point and gradient matching checks.

6.6 Science service requirements

6.6.1 Hardware

In order to support access delivery to the VST, ING and related data services, and host a small number of application sevices, a modest application cluster is required. 8 CPUs will be configured to support data access whilst 8 CPUs will be reserved for applications and providing computational support for PSF fitting photometry and completeness simulations. Re-balancing may be required depending on eventual usage patterns. If additional computational support is required, for instance through large demand for the mosaic service, then use will be made of community resources such as those provided by the National Grid Service.

Storage demands will equate to some 5TB/yr - this providing for the growth of ING, UKIRT (non WFCAM) and AAO data flows, plus the increasing availability of VST survey data from 2009. Storage blocks will be ordered and configured in common with those for the VISTA and WFCAM thus keeping costs to the minimum. Total hardware storage costs for the next 5 years are modest, access to a further 25 TB of on-line storage and the equivalent amount of off-line LTO tape storage.

6.6.2 Personnel

We have estimated the effort required to support these science service activities by comparison with the WFCAM operations development, estimates for s/w delivery in Gaia, and our experience with AstroGrid. Providing

creation and access to VST survey products will require 1.0 FTE/yr by the end of the period. Science service support will be at the level of 0.35 FTE/yr. For the raw data archive support, 0.5 FTE/yr for the first two years will be sufficient to develop and carry out operations and the major interface enhancements, thereafter 0.2 FTE/yr will be needed to operate and maintain the ING archives. Maintaining and running the other main ground-based archives here (AAO and UKIRT but not including WFCAM and VISTA) is included in the above.

6.7 Summary of resources requested

Staff:	0.8 FTE/yr from Q2 2008, ramping up to 1.4 FTE/yr as VST comes on stream
Travel and subsistence:	guideline amount
Equipment:	for CPUs, rack and storage hardware (£35.5k total)
Consumables:	£2k/yr for internet costs (£10k total)
Maintenance:	£1.5k/yr for CPU processors for the whole period (£6k total)

6.8 Key milestones and deliverables

2008–2010	IPHAS processing & calibration, product hosting; community access, science services
2008–2013	VO science application hosting: source extraction, catalogue matching, PCA analysis
Q1 2009	Begin VST surveys
2009–2013	VPHAS processing & calibration, product hosting; community access, science services
Q1 2010	VST DR1 release
2009–2013	Science services supporting VST survey data, e.g. database ingression, database access, object extraction, image viewing, time domain access
2008–2010	2dF pipeline products, reduced spectra, position-spectra matching
2009–2011	AAOmega pipeline products: reduced spectra, position-spectra matching, line flux database
2008–2009	UKIRT/AAO Legacy Archive interface updates
2009–2011	UKIRT/AAO/ING Legacy Archive: on demand science data creation service
2009–2013	New data handling initiatives, demonstrators for SKA Dark Energy Camera, WFMOS, etc
2011–onwards	Gaia, link through VHS/IPHAS cross match services

Note: appendices removed from this version.