4 VISTA and WFCAM data processing

The Cambridge Astronomy Survey Unit and Edinburgh Wide Field Astronomy Unit have been involved in processing and archiving digital wide–field surveys for some time. We intend to use the experience gained from these surveys (eg. APM/SuperCOSMOS surveys, INT WFS and CIRSI survey) to build specialised pipeline processing schemes, define data products and produce a science archive for all WFCAM data (eg. data from consortium surveys – ie. UKIDSS, PATT and non–PATT programmes).

In developing data processing strategy to date we have had to make provisional assumptions regarding the science and data products requirements. These have been based on the extant VISTA and WFCAM documents together with the UKIDSS and VISTA science cases. Undoubtedly these requirements will be revised in the light of further consultation with the WFCAM/UKIDSS and VISTA communities, and with this in mind we have deliberately designed the pipeline(s) around a selectable modular scheme. Further joint development of the science and data product requirements between WFAU, CASU, WFCAM/UKIDSS and VISTA is essential to the success of the projects and is part of the development effort contained in our bid.

The synergy between the WFCAM and VISTA projects and the commonality of their respective requirements, strongly suggests that similar strategies should be implemented for the VISTA data flow system (VDFS). Building on the WFCAM experience and using it as a prototyping system for the VDFS, will result in considerable cost savings and minimise the risk associated with VISTA data processing deliverables.

We also note that VISTA-related activities carried out within the UK wide field groups programme will directly involve external management oversight and progress will also be monitored through an oversight committe appointed by the Grid Steering Committee (GSC).

4.1 Background

The Visible and Infrared Survey Telescope for Astronomy (VISTA) is a UK Universities initiative to produce the World's leading survey facility for near-IR (and eventually optical) astronomy. VISTA, scheduled to be operational on Paranal by the end of 2006, will consist initially of a 4-m diameter telescope with a wide field NIR (1.65 degree diameter field of view) imager containing 16 detectors each with 2k×2k pixels.

The UKIRT Wide Field Camera (WFCAM) scheduled to be operational three years before VISTA, by the end of 2003, shares many of the science aims of VISTA, and equally, many of the data processing and archiving problems. With its array of 4 NIR detectors of 2k×2k pixels, WFCAM, notwithstanding its own highly significant science programme and data handling issues, can be thought of as a trialling system for an integrated VISTA data flow system (VDFS).

The Joint Astronomy Centre (JAC) in Hawaii have formally asked the UK wide field astronomy units to devise a suitable scheme for coping with the data stream from WFCAM. Given the similarity in the data processing and archiving requirements of VISTA and WFCAM and the timescales, it is clear that WFCAM can be considered as the prototype for VISTA thereby allowing characterisation of the real behaviour of the system two to three years before the VISTA DFS has to be complete.

CASU and WFAU members have already held a series of meetings with the VISTA and WFCAM project teams to facilitate cooperative development of a DFS capable of dealing with the output and the science requirements of these facilities. The commonality between the projects has lead to the identification of several workpackages that share almost identical aims. In particular the need for: summit pipelines that assess data quality on-the-fly and provide feedback to survey planning and progress tools; basic calibration and science product pipelines to be run in the UK and a subset of these, in the case of VISTA, to be run at ESO HQ; advanced survey product pipeline development to deal with the complex issues of asynchronous image processing including deep stacking; and an active GRID-enabled science archive capable of dealing with and enabling sophisticated datamining queries and of being a cornerstone of the UK Virtual Observatory.

Specifically the agreement with ESO over VISTA depends on the following deliverables to ESO:

- providing modules to ESO for the ESO-run Quality Control pipeline at the telescope;
- providing modules to ESO for the ESO-run data calibration pipeline in Garching;
- delivering agreed WFCAM survey products to the general ESO community as part of the UK contributions in kind;

and also requires the collaborative development of optimised survey planning tools for operation by ESO using innovative feedback of derived quality control information.

These aims are almost identical to WFCAM requirements for: survey planning and progress; a summit pipeline; and a basic data calibration pipeline. Delivering WFCAM science data products in an agreed form has considerable leverage but involves negligible extra development costs.

Furthermore both WFCAM and VISTA have additional overall science-driven requirements for the common development of a system capable of:

- producing generic agreed science survey products from the calibrated frames;
- ingesting the calibrated frames and science survey products into a UK database archive;
- development of a dynamic advanced processing system for combinatorial operations on basic survey products and calibrated frames;
- ensuring the tools to mine the archive of survey products are available through the AstroGrid project.

The end–to–end WFCAM project is overseen by JAC at Hawaii, and in broad outline consists of:

- building and installing the camera and designing instrumentation software (including mountain–top quick–look pipeline) – responsibility ATC/JAC/CASU;

- producing the UK–side data processing pipeline – responsibility CASU;
- providing the end–user science archive – responsibility WFAU.

For the equivalent tasks for VISTA, the Vista Consortium Director has asked CASU and WFAU to be be responsible for at least the Phase A deliverables outlined in the VISTA GSC application. The anticipation is that the considerable overlap in specification and interests between the projects will correspondingly lead to common use of developed software and hardware solutions.

The remainder of this section of the document is concerned with an overview of the specification requirements for pipeline processing and science data products (science archiving for WFCAM is the responsibility of WFAU and will be dealt with in their submission) both for WFCAM and insofar as we know them for VISTA. For the sake of brevity many of the technical details have been ommitted from this RG2 but are available as part of the WFCAM preliminary design review documentation submitted for the December 2001 UKIRT board meeting, UKIRTB(01)31, and as part of the VISTA DFS documentation, GSC(02)03, submitted to the Grid Steering Committee in March 2002.

4.2 Pipeline Processing Requirements

As already indicated there is an operational and a practical requirement for three distinct data processing pipelines for both VISTA and WFCAM: a summit pipeline; a basic pipeline back in the UK (of which the ESO version for running at Garching will be a subset); and an advanced pipeline also in the UK. All WFCAM data, including non-UKIDSS data, is to be processed through the summit and basic pipelines and the calibrated science data made securely available through the on-line WFCAM archive. The same is taken to be the case for the VISTA pipeline.

An on-line database describing the observations, and the summit-derived Data Quality Control (DQC) information will be maintained by the JAC, Hawaii, and mirrored in Cambridge, as is done currently with all UKIRT observations. These may require updating when final DQC information is available from the UK-run basic pipeline. This database will form the knowledge interface that helps define and plan the overall observing strategy. Again, equivalent requirements arise for VISTA, and ESO, Garching, will maintain a similar on-line observation database. In particular the survey planning and progress tools currently being devised for the WFCAM surveys through the ORAC Survey Definition (SD), Observation Management Package (OMP) and definition of Minimum Schedulable Blocks (MSBs), have direct equivalents/needs for the ESO/VISTA operation.

4.3 Overall strategy

The ESO and ORAC control systems will produce standard Multi Extension Files (MEFs) in FITS format for export to UK pipelines. Stored within will be sufficient ancilliary information to fully describe the data taken and to invoke the appropriate pipeline processing actions.

On Hawaii the summit pipeline will run on the 'native' NDF format used at UKIRT, but will otherwise provide a subset of the operations carried out in the basic pipeline. CASU will provide

PERL-based scripts for converting from NDF to MEF format at the summit and for undertaking data and header integrity checks prior to export to the UK. CASU are also currently negotiating with the JAC and ATC over the feasibility of using CASU modules for the underlying summit processing and DQC checks.

At Paranal the summit pipeline will consist of a subset of the UK-supplied calibration pipeline operations, in particular those tasks necessary to provide near real time QC information.

Data will be taken according to a fixed menu of pre–defined observing protocols, in order to guarantee pipeline processing both at the summit, (at ESO HQ) and in the UK. These will include single shot, interleaving, dither sequences, tile sequences and so on – details TBD (see glossary in Appendix for definition of terms).

The general philosophy will be that all fundamental data products are FITS MEFs with headers describing the data taking protocols in sufficient detail to trigger the appropriate pipeline processing components, including the generated catalogue binary tables, and that all derived information, DQC, photometric and astrometric calibration and processing details, are also incorporated within the FITS headers.

The FITS headers thereby provide the basis for ingest into databases for archiving and databases for real time monitoring of survey progress and survey planning.

4.4 Summit Pipelines

The Data Aquisition System (DAS) will do coadds of multiple reads/consecutive exposures on the same field, apply non-linearity corrections, as necessary, and assemble each multiquadrant read into a coherent whole. For WFCAM this will minimise the DFS requirements to $\approx$100 Gbytes per night. VISTA will generate roughly a factor $\times 4$ more data each night.

The main purpose of the summit pipelines is to generate near real-time DQC information by processing data on-the-fly using either library master calibration frames or simple locally derived (from each MSB or OB) calibration data.

For WFCAM the summit pipeline will run under the control of the ORAC data aquisition and processing system and together with the CASU-supplied modules, will produce the required DQC information, the export MEF FITS files and instigate a FITS header and data verification program.

The summit pipelines will have to maintain and update a "best guess" series of master calibration frames (dark, flats, skies, hot/bad pixel masks) and to monitor and measure the instrumental non-linearities to provide basic 2D instrumental signature removal. The DQC feedback will probably require some form of catalogue generation software which can be used to produce measures for DQC monitoring, including: astrometric pointing accuracy - requires telescope to point to better than a few arcsec rms; photometric throughput; image shape measures – trailing; sky brightness levels; sky noise levels (details TBD). This latter software should be a subset of the basic pipeline.

As already noted, to minimise data volumes the data aquisition system must carry out basic operations such as reset-correction and stacking multiple sequential exposures (an 'integration') in situ before forwarding data to the rest of the system.

4.5 Basic pipelines

The basic pipeline will operate on single pointings or other "well-defined" observing sequences (protocols specified in MSBs/OBs) taken during a single night to produce instrument signature–free images and to derive object catalogues from them. The object catalogues will be used to provide the first pass astrometric and photometric calibration and together with the images will define the basic science survey products.

The two-dimensional image data products at this stage will be instrumental signature-corrected: single frames; lossless interleaved super frames; stacked dither (super)frames; tiled contiguous mosaics, combining four pointings with any of previous options; and confidence maps for all pipeline output products (details TBD).

The catalogue data products will be lists of detected images with an agreed set of parameters (details TBD) summarising useful astronomical information and providing the necessary DQC information.[1]

In more somewhat more detail the requirements are:

- instrumental signature removal - continuous updating of flatfielding and sky background tracking and correction, if necessary, during interleaving/dither sequence stacking, possible need to also remove other 2D background variations from imperfect multi-quadrant operation of detectors;
- define and produce a strategy for dealing with image retention from preceding exposures if present;
- combine frames if part of an observed dither sequence or tile pattern and develop a consistent internal calibration scheme to put observations on an approximately uniform system to simplify latter external calibration;
- basic catalogue generation including astrometric and photometric calibration, shape and DQC information. This should make use of an appropriate World Coordinate System (WCS) which will enable the basic photometric calibration from the catalogues using suitable pre-selected standard areas covering the entire field-of-view to monitor and control systematics;
- for VISTA, a UK-supplied calibration pipeline will be also be run at ESO HQ in Garching by ESO's DMD, to provide the essential calibration and QC information for their raw Data Archive. ESO's data processing responsibility for their community stops at this point;
- each frame and catalogue supplied with provisional calibration information and overall morphological classification embedded in FITS files and include propagation of error arrays and use of confidence maps;
- processing simulated data and federating with other catalogues to assess the completeness, systematics and accuracy of the science product catalogues;

[1] Parameters to include: image location, various flux and shape measures, and error estimates for them. For a starting point we will base this on the INT WFC pipeline catalogue products, with selective addition from the SDSS and 2MASS pipelines.

- the ability to reprocess the entire dataset after upgrades/bug fixes.

Although many of the basic processing modules are in place, we are still evaluating alternative high level control and user interface strategies for overall pipeline flow control. A generic flowchart outlining the main processing tasks required for the basic pipeline is shown in figure 1.

4.6 Survey progess Databases

There are clear requirements for automatic real time monitoring of survey progress with minimal user input via online databases and the provision of simple visualisation of overall survey progress. This could either be from CASU's normal SYBASE repository, a POSTGRESQL database similar to the experimental one used for monitoring the INT WFS, or an interface to the proposed SX archive database provided by WFAU. For VISTA mirroring the ESO Data Archive in Garching is a further option.

This should include interactive GUIs and CLI for querying survey DQC, state of processing, progress of survey, and should also be capable of feedback to observing strategy.

This database could also provide the driver for many of the more advanced processing options that require predefined (asynchronous) observations to have occurred in for example, multiple passbands or muliple visits to the same field for deeper stacking.

4.7 More advanced processing

Most of the likely requirements for an advanced dynamic processing system stem from the necessary combinatorial operations that will be required on images/catalogues derived from basic calibrated science survey products.

The more important and obvious of these have been grouped into clearly defined interrelated packages and are outlined below.

- Many advanced survey products are predicated on the ability to automatically generate detailed well-sampled PSFs as a function of position over each basic product image. Example uses of accurate PSFs are: optional automated PSF fitting to improve s:n of point sources particularly in crowded stellar regions; and adaptive kernel matching of PSFs for difference imaging for transient event detection (SNe, planetary transits, solar system objects).

- A related, but functionally separate issue, is how to optimally stack images taken at different times in (necessarily) different seeing conditions. This will involve assessing various stacking techniques where optimality may well depend on the science aims for the stacked data and may well delve deeply into (CPU intensive) image restoration techniques.

- Producing contiguous tiled regions and seamless overlap-calibrated large area catalogues. These and the preceding item could be initiated automatically from a DataBase-driven list of survey fields and survey progress eg. is it worth (re)stacking (re)mosaicing (re)calibrating this region ?

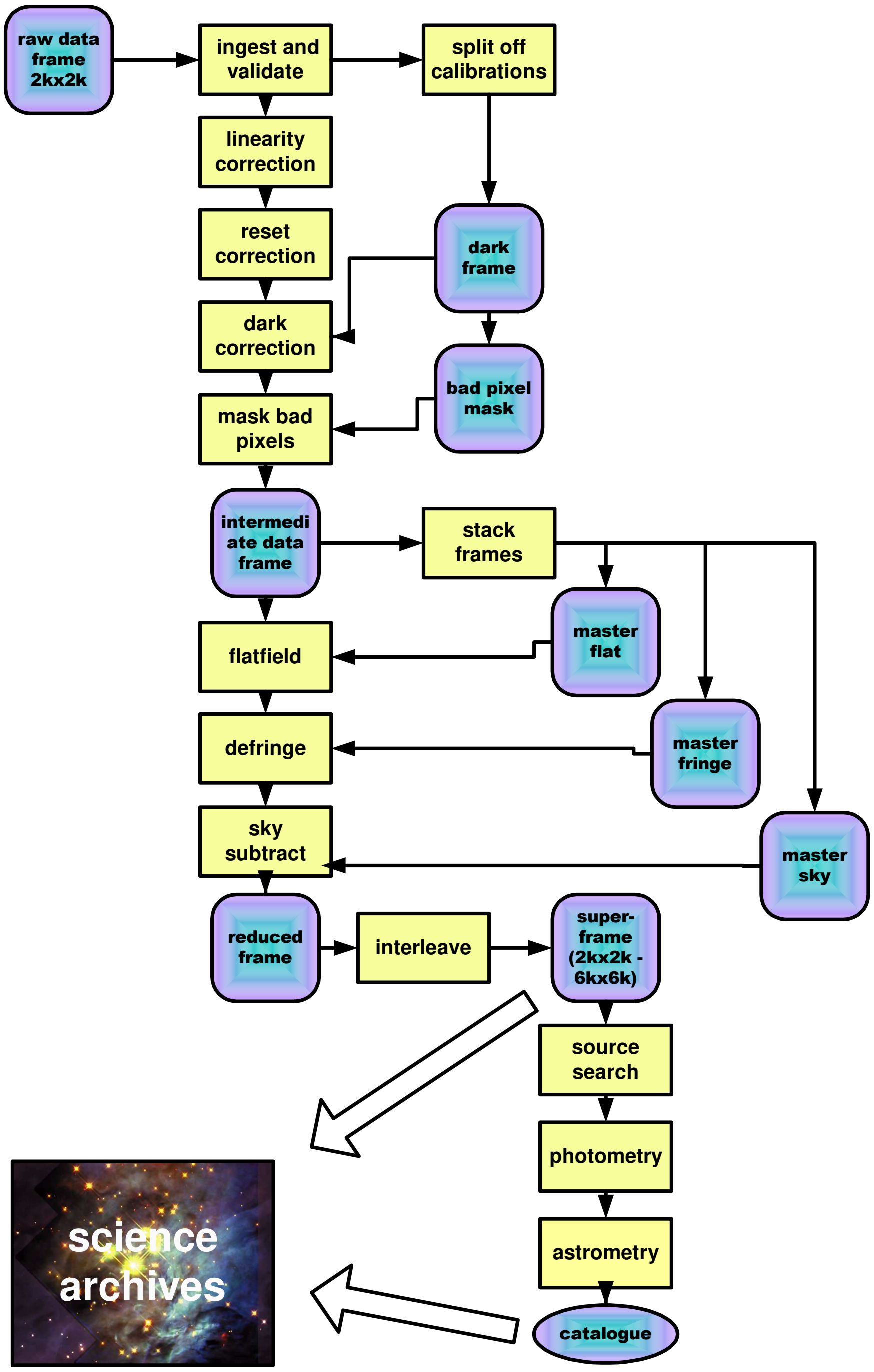

Figure 1: Overview of processing operations required to deliver basic science survey products from WFCAM/VISTA data.

- Detection and parameterisation of Low Surface Brightness objects and an investigation of more general purpose image detection methodology that is not just optimised to find faint PSF-like images.
- Matching/merging frames in different passbands and for use in defining master catalogues for any given field for list-driven co-located photometry (cf. SDSS and their ability to measure fluxes and limits to fluxes in a consistent way).

Several of these items particularly PSF-based operations involving optimal stacking are research areas in their own right and could profitably be pursued as collaborative research projects. We have asked for a further 1 FTE/yr of research effort in our bid mainly for progressing these and related issues. As a further example of this we have already held fruitful meetings with the Subaru Supercomputing Group in Hilo who are keen to investigate alternative (and necessarily compute intensive) methods for optimal stacking of images taken in varying seeing conditions.

4.8 Hardware/software requirements

In order to provide a backup for the main on-line science archive at ROE and also to facilitate several of the advanced processing/reprocessing stages needed, we propose to store on-line the raw science frames and master calibration data of WFCAM in Cambridge, essentially an extension of the rest of the UKIRT raw data archive. Costs of suitable scale commodity RAID arrays ($\approx$10 Tbytes capacity for each full year of operation) have already fallen sufficiently (currently £5k per Tbyte but likley to be half that price by 2004 when the first tranch of bulk storage will need to be in place) to make this a cost effective and viable (scalable) solution.[2]

With on-line storage in place at two sites we do not foresee the need for further (expensive) near-line "tape" storage systems providing that a high density transfer medium is used from Hawaii, whereby the $\approx$100 Gbytes of data expected per night is stored on one "tape". If recovery from the raw tape data (archive) is required this would be feasible, even for on-the-shelf storage.

Precisely which system to use for data transfer between the JAC and Cambridge is still under discussion. If tapes, CASU will need to purchase two drives to both guarantee the transfer speed and provide a backup in the event of unit failure. Transfer of processed data between Cambridge and Edinburgh will be via the internet. CASU currently share a 1 Gbit/s link into the LAN. Assuming WFAU have a similar level of access by 2003–4, 10% of that bandwidth would suffice to transfer a days processed data in a few hours.

We expect to have sufficient computing power in place as part of the Cambridge Data Processing Centre activities to handle the development issues of the pipeline for WFCAM, including trialling different processing methods and benchmarking simulated WFCAM data processing. However, sometime during 2003 we will need to purchase and setup a modest multi-node PC cluster to deal

[2] The annual growth rate of the science storage requirements is estimated as follows: a typical WFCAM science frame will be 4 devices of $\times$ 2048 $\times$ 2048 pixels, $\times 2 \times 2$ interlaced, $\times$4 bytes per pixel $\approx$ 250 Mbyte. Assuming 75% UKIRT time is devoted to WFCAM with 70% fill factor (weather, faults, other science programmes) we will have to deal annually with $\approx$ 200 nights of science data. Assuming $\sim$ 200 frames per night (an average of one every 3 mins over a 10hr night) then the pixel archive grows by $\approx$10 Tbyte/yr. All other data (derived from pixels or elsewhere) will be $<<$ 10% the size of the pixel archive.

with the real data flow expected to start toward the end of that year and a modest high performance PC-based system to begin serious investigation of advanced processing options and equally heavy duty simulations to assess catalogue completeness issues and the reliability of parameter error estimates.

4.9 Human Resource Estimates and Timescale

We intend to appoint a UKIRT WFCAM data processing manager from within our group who will be responsible for overseeing the operation of the UKIRT pipeline and organising appropriate software and hardware effort as necessary. Specialist system-level support will be complemented by the IoA computer manager group. The data processing manager will also be responsible for ensuring that the pipeline software and science data products are appropriately documented.

Since WFCAM can be thought of as a prototpying system for VISTA, for the duration of this RG2 grant there should be no additional VISTA hardware or software maintenance costs. There is, however, an additional need not covered by the WFCAM work for a CASU-VISTA-ESO liaison person, the VDFS technical manager, whose job it will be to aid in the implementation and to help negotiate the requirements for the VISTA QC and calibration modules to be delivered to ESO.

The VDFS technical manager will be responsible for liasing between CASU, VISTA project office and ESO and helping to define the protocols the delivered products must conform to. They will also be expected to contribute directly to the development and help in the deployment of the software modules in a form required by ESO. This includes participating in producing the appropriate docmentation and tools as specified in Appendix B of the VDFS document GSC(02)03 and will include close liaison with the ESO DFS team in Garching. We have identified a suitable individual to undertake this task, initially at the level of 0.5 FTE/yr and have included this in the grant bid under the VISTA–ESO heading.

We estimate that further generic software development and pipeline management procedures for WFCAM/VISTA over the next year will require approximately 3 FTE of effort.

Most of this effort will be expended on designing and coding of the UK-side pipelines, a subset of which may be required for the summit pipeline, and helping to collaboratively devise with the JAC a suitable survey progress and planning interface. This could form the basis for the VISTA/ESO survey planning tool.

Part of the pre-survey responsibility of the WFCAM data processing manager will be community liaison via the UKIDSS Survey Project Scientist to ensure that the projected science survey products and the data taking and processing strategy satfisy user expectations.

We have already taken part in test observing runs at UKIRT to assess the observer interface and begin investigations on the optimal strategy for dealing will issues relating to flatfielding stability, image persistence [3], possible intra-pixel sensitivity problems, non-linearity monitoring and so on. These will continue over the next year.

[3] Including how to deal with bright stars, ghostimages, trails – one possiblity it to use input predictions for positions of bright stars, galaxies, globular clusters, and solar system objects.

Thereafter, we estimate that for CASU to maintain and run the WFCAM pipeline, produce science data products and the survey progress archive, plus further develop the processing tools and to help advertise and ensure access to the products will require a total of 3.5–4 FTE/yr, from the last quarter of 2003, when the first real science data is expected to arrive. The greater fraction of this will be expended in day-to-day maintaining, developing and running the pipeline.

We are working on the assumption that the first science observations for WFCAM will occur in the final quarter of 2003. The pipeline system must therefore be in place by the end of 2003.