VISTA Data Flow System: Overview

Jim Emerson[a] , Mike Irwin[b], Jim Lewis[b], Simon Hodgkin[b], Dafydd Evans[b], Peter Bunclark[b], Richard McMahon[b], Nigel Hambly[c], Robert Mann[c,d], Ian Bond[c], Eckhard Sutorius[c], Michael Read[c], Peredur Williams[c], Andrew Lawrence[c,] Malcolm Stewart[e]

[a]Queen Mary University of London, Astronomy Unit, Mile End Road, London, E1 4NS, UK
[b]University of Cambridge, Institute of Astronomy, Madingley Road, Cambridge, CB3 0HA, UK
[c]Insititute for Astronomy, School of Physics, University of Edinburgh, Royal Observatory, Blackford Hill, Edinburgh, EH9 3HJ, UK
[d]National e-Science Centre, 15 South College Street, Edinburgh EH8 9AA, UK
[e]UK Astronomy Technology Centre, Royal Observatory, Blackford Hill, Edinburgh EH9 3HJ, UK

ABSTRACT

Data from the two IR survey cameras WFCAM (at UKIRT in the northern hemisphere) and VISTA (at ESO in the southern hemisphere) can arrive at rates approaching 1.4 TB/night for of order 10 years. Handling the data rates on a nightly basis, and the volumes of survey data accumulated over time each present new challenges. The approach adopted by the UK's VISTA Data Flow System (for WFCAM & VISTA data) is outlined, emphasizing how the design will meet the end-to-end requirements of the system, from on-site monitoring of the quality of the data acquired, removal of instrumental artefacts, astrometric and photometric calibration, to accessibility of curated and user-specified data products in the context of the Virtual Observatory. Accompanying papers by Irwin et al[1] and Hambly et al[2] detail the design of the pipeline and science archive aspects of the project.

Keywords: astronomical surveys, data quality control, pipeline, science archives, VISTA, WFCAM, VDFS

1 INTRODUCTION

1.1 Overview

WFCAM (the Wide Field CAMera) and VISTA (the Visible and Infrared Survey Telescope for Astronomy) are wide field survey instruments designed to rapidly build up multi-band surveys of very large areas of the sky using arrays of 2Kx2K IR detectors on 4-m class telescopes in the Northern and Southern skies respectively. Some 80% of WFCAM time and at least 75% of VISTA time will be dedicated to large scale ‘public’ surveys whose data volumes put the data reduction beyond the means of most individual astronomers, or groups of astronomers. Thus it is necessary to have a Data Flow System to remove instrumental artefacts from the images, astrometrically and photometrically calibrate them, and to combine many such images into surveys and resulting catalogues, making the resulting products, and tools to use them, accessible to science users. This is in addition to the facilities required to allow assessment of data quality at the telescopes themselves. The strong UK interest and background in survey astronomy, coupled with its construction of WFCAM and VISTA therefore led us to decide to process this survey data in a uniform and consistent manner with what is now known as the VISTA Data Flow System. The name reflects the long term goal, although in fact the system applies equally to WFCAM data. This paper gives an overview of the requirements and design approach adopted for the VDFS, and is accompanied by two more detailed papers focussing on the (most immediate) case of WFCAM data, one by Irwin et al.[1] on the pipeline processing, and one by Hambly et al.[2] focussing on the science archive.

1.2 WFCAM and VISTA instruments

The fields of view of WFCAM (1.0°), and VISTA (1.65°) are sparsely sampled by arrays of 2Kx2K IR detectors (WFCAM: 2x2 Rockwell detector arrays with 0.40” pixels, VISTA: 4x4 Raytheon VIRGO detector arrays with 0.34” pixels). Both have interchangeable sets of filters which will initially be for WFCAM: Z,Y,J,H,K and narrow band

Correspondence: email j.p.emerson@qmul.ac.uk

$H_2S(1)$, CO, Brγ and for VISTA: Y,J,H,K_s. WFCAM uses the existing UKIRT telescope on Mauna Kea Hawaii from late 2004, whereas VISTA was designed as a single purpose survey facility located at ESO's Cerro Paranal Observatory in Chile from late 2006. Details of the design of WFCAM[3] and VISTA are given elsewhere[4,5,6].

Because the detectors are separated by a significant fraction of a detector width (WFCAM: 0.94 in X&Y, VISTA: 0.90 in X and 0.425 in Y) each exposure generates a non-contiguous image of the sky known as a "pawprint". A filled image of the sky can be constructed efficiently without gaps by combining several of these pawprints made with specific telescope offsets to produce a contiguous image known as a "tile". For WFCAM 4 offsets, two of 0.97 detector widths in the focal plane in X, and two of 0.97 in Y, suffice to cover each piece of sky at least *once*, and for VISTA 6 offsets with 2 steps of 0.95 detector widths in the focal plane in X, and 3 steps of 0.475 in Y, suffice to cover each piece of sky at least *twice*. The VISTA case is shown in Figure 1, and the resulting exposure time map in Figure 2.

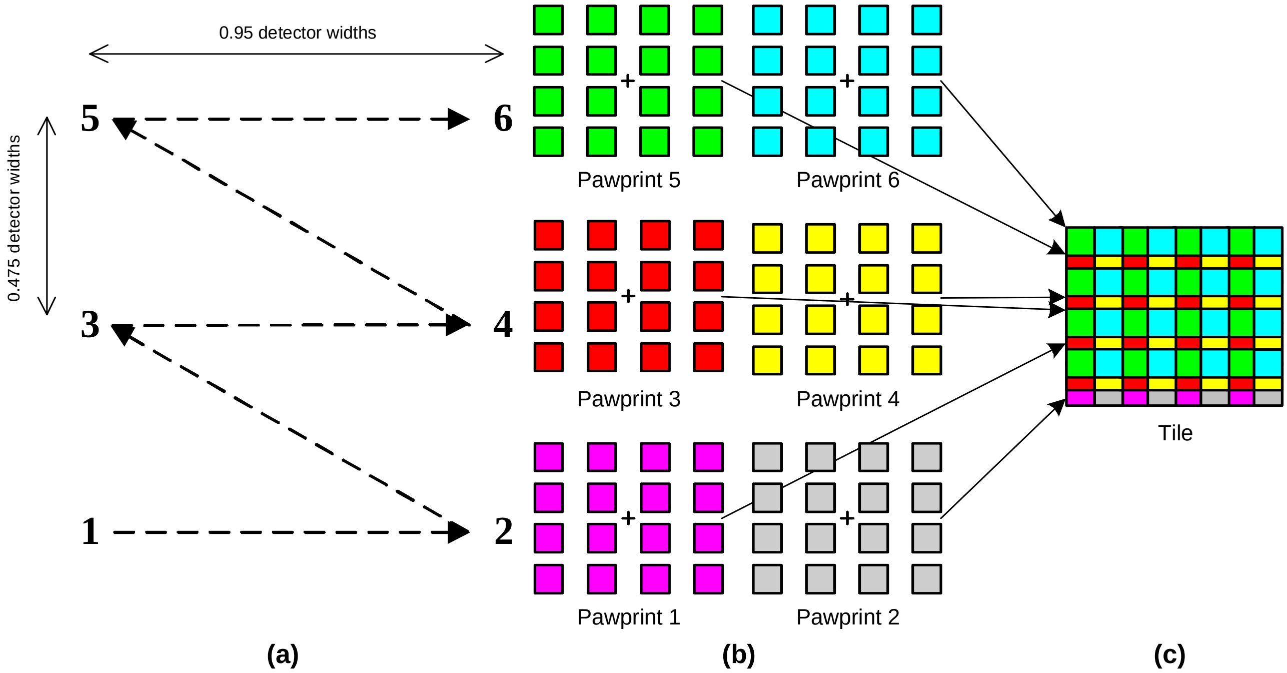

Figure 1. A contiguous image, or "tile", of the sky can be made with VISTA by stepping the telescope (a), to make six "pawprint" exposures (b), which are then overlapped and combined together to make a single contiguous image (c). Each piece of sky is covered at least twice except at the edges. The shadings (colours in the CD version of these proceedings) in (c) indicate the last pawprint to be added in any position, for the order shown. The 6 pawprint positions can however be made in any order.

Small-scale artefacts, caused for example by cosmic ray events or bad pixels, can be eliminated by stacking several "pawprints" together using small telescope offsets; a strategy known variously (for historical reasons) as "jittering" in the case of VISTA but as "dithering" in the case of WFCAM. A microstepping strategy can also be employed, stepping the telescope by exact fractions (e.g. n+0.5) of a detector pixel and interleaving the data to increase the sampling frequency.

1.3 Data rates and volumes

VISTA is required to be capable of sustaining an exposure every 10 seconds for a 14 hour observation period, leading to a maximum nightly data rate of 1.35 TB/night as indicated in Table 1 which compares the requirements of VISTA and WFCAM. On most nights such a rate may not be required and a realistic average nightly rate might be ~30% of this, though we continue to use the maximum rate for purposes of this discussion. Applying the same number to WFCAM produces a nightly rate ¼ of that of VISTA because WFCAM has ¼ the number of detectors. Each set of 16 (4 for

WFCAM) detector files forming a pawprint is in due course merged into a single multi extension FITS file, but this does not change the data rate or volume which the quality control and calibration software needs to process.

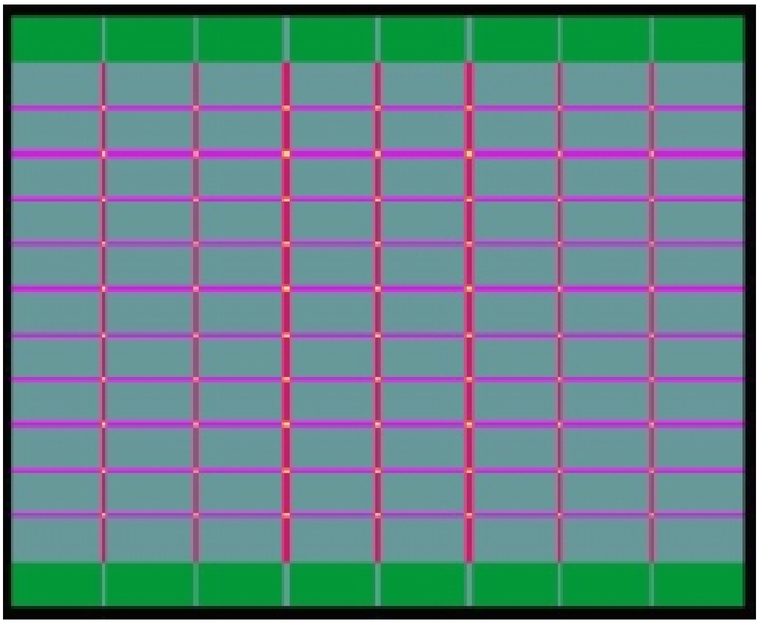

Figure 2. Map of exposure time for a filled (unjittered) VISTA tile of 6 pawprints. Green = 1 (at top and bottom), grey green = 2 (most of image), magenta = 3 (in horizontal stripes), red = 4 (in vertical stripes), yellow = 6 (at interstices), in units of the single-pawprint exposure time.

	VISTA	**WFCAM**
Pawprints		
Detector format:	2048 x2048 pixels of 4 bytes	
Size of one detectors output:	16.78 Mbytes	
Number of detectors:	4 x 4 = 16	2 x 2 = 4
Total number of pixels in a single pawprint:	67.1 Mpixels	16. 8 Mpixels
Size of one pawprint's data:	268.4 Mbytes	67.1 Mbytes
Data Rates		
Maximum sustained exposure rate:	1 pawprint every 10 sec	1 pawprint every 10 sec
Sustained storage rate required:	26.8 Mbytes per second	6.7 Mbytes per second
Volumes / night		
Maximum observing duration:	14 hours	14 hours
Maximum raw data storage per night:	1.35 Tbytes per night	0.34 Tbytes per night
Typical raw data storage per night (~30% maximum):	0.45 Tbytes per night	0.11 Tbytes per night
Volumes / year		
Nights scheduled per year:	~350	180
Nights operated per year: (assuming ~85% usable)	~300 nights	~150 nights
Raw Data storage per year: (at 30% maximum rate)	135 Tbytes per year	17 Tbytes per year
Lifetime Raw Data Volumes		
Years operated:	15 years +	7 years +
Lifetime raw data volume: (at 30% maximum rate)	>2.03 Pbytes	>0.12 Pbytes

Table 1. Raw data rates and volumes required for VISTA compared with WFCAM (no data compression is included)

Table 1 shows and compares the raw data rates, and the accumulated volumes, for VISTA and WFCAM. Although the volume of data to be processed per observing night only differs by a factor of 4, VISTA will be used whenever weather and maintenance permit. WFCAM however will compete with other instruments, and weather conditions can be more severe at UKIRT than at the Cerro Paranal Observatory, so WFCAM will operate for perhaps half the nights per year of VISTA leading to a difference of annual raw data volumes of a factor of ~8 (the actual figure will only be known in light of real experience with both VISTA and WFCAM). WFCAM is expected to operate for at least 7 years (during which 1000 nights are allocated to the UKIDSS surveys[7]) whereas VISTA is expected to operate for at least 15 years so the total raw data volume after 7 years differ by a factor of 8 and after 15 years by a factor of 17.

To reduce the data storage, I/O overheads and transport requirements, we intend to use, as much as possible, the lossless Rice tile compression scheme as used transparently, for example, in CFITSIO.[8] For this type of data (32 bit integer) the Rice compression algorithm typically gives an overall factor of 3–4 reduction in file size.

1.4 Requirements of organisation operating the telescopes

The Joint Astronomy Centre (JAC) and European Southern Observatory (ESO), the organisations that operate WFCAM and VISTA respectively, both require processing of the raw data at the telescope site to derive quality control measures which can be used to monitor instrument performance. They both also require software to remove instrumental artefacts and derive calibrated science data from observations made during each night. JAC further require quick look science data at the telescope. For both WFCAM and VISTA the raw data is returned to Europe roughly weekly on LTO2 tape (WFCAM) or disk (VISTA), and is then copied using the internet giving duplicate copies of all the raw data in the UK and at ESO.

1.5 Science requirements & the Virtual Observatory

Whilst some science users will want to work directly with, or ‘data mine’, the calibrated images from the surveys, many more will want to mine the catalogues of objects derived from the images. Based on previous experience we estimate that the volume of catalogue data will be ~10% of the raw data volume (1.7 & 13.5 Terabytes per year for WFCAM and VISTA respectively), and at such annual volumes the handling and exploration of catalogues becomes a task requiring very careful planning. The data volumes and the need for data mining strongly indicated that designing the science archives in the context of the emerging Virtual Observatory (VO) would be the best approach for handling the vast VISTA & WFCAM data volumes, and the many uses to which users might want to put the data. Originally making the VDFS a component of the UK’s AstroGrid[9] VO work was considered, but it was later decided that the priorities and schedules of both projects were best served by VDFS concentrating on the WFCAM and VISTA specific aspects of the problem, and how these should be integrated into the wider VO, using AstroGrid tools.

2 APPROACH

2.1 Similarity of WFCAM and VISTA tasks

The similarities between WFCAM and VISTA data, their parallel scientific purposes and data exploitation requirements, immediately suggested that synergies in the area of survey processing and production should be exploited, for reasons of both efficiency and cost. Furthermore the experience of handling WFCAM data would provide invaluable experience for dealing with the later and larger data rates and volumes from VISTA, which could otherwise easily overwhelm a conventional processing system. Essentially the approach adopted was to engineer a scalable system that could handle both WFCAM and VISTA data, whilst retaining as much flexibility as possible to modify the plans for VISTA in the light of experience with actual WFCAM data. The system name adopted was the VISTA Data Flow System (VDFS).

Processing of data at the telescopes is covered by separate modules for WFCAM and VISTA (recognising the differences in observatory practise). Extrapolating the modular approach that led us to keep VDFS as a distinct entity from AstroGrid, it was decided to deal with the processing of the image data to calibrated form by developing a common set of software modules at the Cambridge Astronomical Survey Unit (CASU). The calibrated science products are then placed in a curated science archive, designed and operated by the Wide Field Astronomy Unit (WFAU) of Edinburgh University. This approach allows CASU to focus on the immediate problems of calibrating the latest data to arrive, whilst WFAU focuses on the problems of combining or comparing many calibrated pawprints (to go deeper, or wider, or to look for time evolution), and of providing the means for the many users to access and mine the vast amounts of data products.

An indication of the complexity of the system is provided by Fig 3 which shows the flow of data from UKIRT (on the left) to CASU (in the middle) where the calibration work is done, and then on to WFAU (on the right) where the calibrated frames are ingested into databases. Fig 3 also shows the interaction with users and the VO and various feedbacks. Careful control of the interfaces between the parts of the system is ensured by producing an Interface Control Document (ICD) describing the format of the raw image files which form the interface between the data from the telescope data storage systems and the CASU pipeline (example for WFCAM[10]). A second ICD describes the format of the calibrated images and extracted catalogues, defining the interface between the CASU pipeline and the science archive at WFAU (example for WFCAM[11]). The interface between the science archive and the users will continue to evolve with user feedback and as the capabilities of the VO evolve, and is described with the data products. The science archive's goal is to be compatible with VO standards wherever possible.

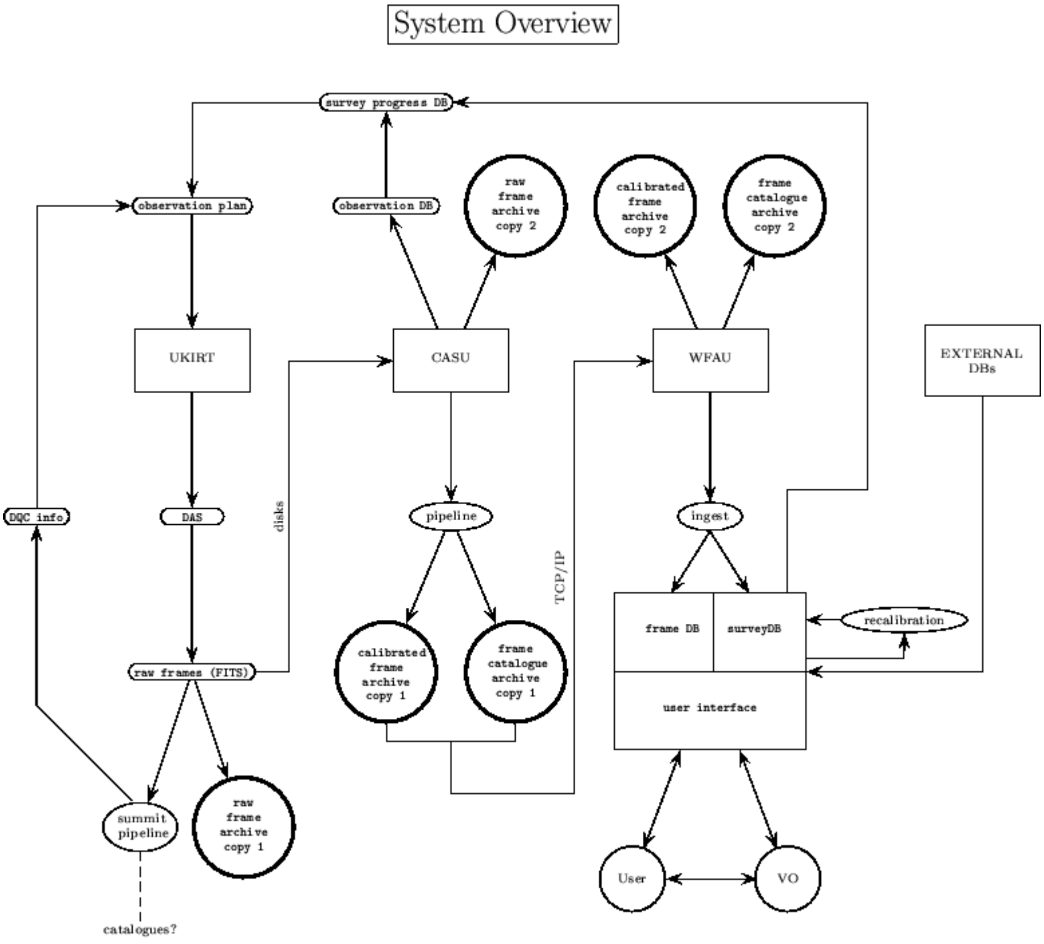

Figure 3. Data flow (for WFCAM). CASU is the pipeline centre, and WFAU the science archive centre.

2.2 Community liaison

The requirements of the organisations operating WFCAM and VISTA (JAC and ESO respectively) are unambiguously known, but the detailed requirements of the science users in the UK were not initially well defined in detail. In the case of VISTA basic processing requirements had been laid down in the project's Science Requirements Document (SRD), though formally they are on the end to end VISTA system, and the project has a full time project scientist and a Science Committee with members from all 18 Universities of the VISTA Consortium. With the addition of a VDFS Users

Committee to ensure the representation of other users the mechanism for inputting VISTA requirements were clear, once the expected performance, including overheads, of VISTA could be fixed after the design phase was completed.

However with WFCAM originally due ~3 years before VISTA, and the user communities in the UK being largely the same people, the procedure adopted was first to focus on the processing requirements for WFCAM as represented by the members of the UK Infrared Deep Sky Survey consortium (UKIDSS[12]). Through a series of proposals, iterations and meetings an agreed set of written requirements and goals for the pipeline and the science archive were drawn up and prioritised, in conjunction with the VDFS team, allowing a schedule to be drawn up indicating at what stage each of the requested products or capabilities would be made available for WFCAM. The schedule and progress are regularly presented to the UKIDSS consortium, and feedback noted, helping ensure that the system will meet the needs of the community. To further verify their needs are adequately covered the UKIDSS consortium have obtained 6 months funding for a 'Consortium Science Verifier' to provide independent checks that the pipeline and archive are indeed behaving as expected. This process should be very important in rooting out the odd bugs and omissions that may otherwise slip into (or fall out of) the system. As the final design reviews for the construction of VISTA itself have recently all been completed, a similar VDFS exercise will soon begin focussing on VISTA data.

2.3 Security

Although much of WFCAM and VISTA will be devoted to 'public' surveys some 25% will be for more traditional Principal Investigator (PI) type proposals. Whilst the proprietary periods for big 'public' surveys will presumably differ from those for normal PI type proposals some form of control, probably in the form of periodic data releases, will certainly be needed even for 'public' surveys, and so the VDFS must have systems to ensure respect of data access and security rights on all data, in accordance with any data rights or restrictions associated with the awards of time by the awarding bodies, either in the UK or ESO. These security systems will ensure that only those with rights to access data are able to do so.

2.4 The development process

The design plans for both the pipeline and science archive went through a peer review culminating in a final design review of both software and hardware, designs being scalable to VISTA data rates and volumes. Purchase of hardware has been delayed as long as possible to take advantage of Moore's law[13].

We have found it essential to follow a spiral model of software development — developing a series of increasingly complex prototypes rather than attempting to follow a linear path (waterfall model). The schedule for the deliverables to the observatories are defined by JAC and ESO respectively. VDFS pipeline and science archive deliverables for the science users in the UK are designed around a series of incremental releases. v1 (released end 2003), v2 & v3 will each operate on WFCAM data, whilst v4 onwards will operate on both VISTA and WFCAM data. v3, to be released before VDFS commissioning for VISTA data, will also contain the functionality needed for VISTA and v4 will effectively be a thoroughly debugged version of v3. Scalability tests for v3 and v4 will be carried out by re-processing the complete WFCAM raw data archive (which should be ~10 TB by mid 2005, i.e. equivalent to 3 weeks of VISTA data) because it is expected that the successive versions of the VDFS pipeline will increase the quality of the reduced frames and the parameters of the sources detected from them. This re-processing will be, in parallel with regular processing of new data as it arrives. This increase in the volume of data needing processing will itself also be a partial test of the scalability requirement for VISTA data. Similarly, the transfer to the Science archive at WFAU and archiving of these re-processed data will test the scalability of data transfer and ingestion to volumes expected from VISTA.

Although it might seem simpler and "cleaner" to overwrite the data from earlier versions of the pipeline, we recognise that this would be unacceptable to many users (e.g. those who have defined samples for spectroscopic follow-up through selections made on data from earlier versions) and so the archive will retain multiple versions of some data products.

3 ON-SITE QUALITY CONTROL MEASURES

3.1 Observation software

Although observation software is part of the control software provided by the organisations operating the telescopes, the need to process the data requires the VDFS pipeline to have a role in the definition of the observing procedures, to ensure that data taken can be handled and calibrated. Observations are specified in Minimum Schedulable Blocks (MSBs) for WFCAM, and Observation Blocks (OBs) for ESO. These MSBs or OBs are prepared in advance by observers (or survey consortia), and used to sequence the camera and telescope through the steps needed to make each observation. The OBs are (in the ESO case - WFCAM is analogous) made up of templates for: calibration of science observations (reset frame, dark, dome flat, twilight flat); calibration of science detector properties (linearity, cross-talk, persistence, etc…); single "pawprint" observation; sequences of observations to build a "tile", with user-selectable nesting sequence for filter exchange, tile building, jittering (known as dithering in the case of WFCAM) and microstepping.

3.2 Quality control measures and trending

Both JAC and ESO require a pipeline at the telescope to derive quality control measures to allow assessment of instrument health and performance. In the case of ESO the derived measures are (as with all ESO telescopes) sent to ESO HQ in Garching where they are examined and the results of the analysis returned during the (Paranal) day so that any necessary actions can be initiated before the next observing night. Only the means to derive the data quality measures need to be provided to ESO, as they deal with the analysis of trends. In the case of WFCAM the JAC require the summit pipeline to do quality monitoring on-site, and so we have provided tools to do this. The quality control measures derived at the telescope are specifically designed to address the performance of the instruments so their products are not themselves included with the raw data returned to the UK & ESO. Some of the measures are re-derived in the pipelines in Europe for purposes of assessing science quality of data, which is particularly important when deriving stacking or tiling, and for large surveys whose reliability and completeness would ideally be uniform across the areas surveyed.

4 REMOVAL OF INSTRUMENTAL ARTEFACTS & CALIBRATION

4.1 Instrumental artefact removal

The aim of instrumental artefact removal is to produce, as near as possible, an image as though it were taken with a perfect, linear, blemish-free camera. Instrumental artefact removal at the individual detector level is conceptually similar to that for a single 2Kx2K IR detector, and much of the processing will indeed work on single detectors, which can be processed in parallel with suitable CPUs. The steps involve corrections for dark frame, linearisation, gain, background, electrical crosstalk, persistence, non-uniform illumination, and bad pixels.

The large number of detectors, and the presence of autoguiding and wavefront sensing CCDs in the IR focal planes give greater potential for cross-talk between channels and detectors than usual. For example each of the 16 VIRGO detectors for VISTA will be read out using 16 adjacent channels. The signal from one channel can interfere with the signal on other channels or detectors that are clocked out at the same time. The crosstalk may be negligible, but needs to be specifically characterized prior to, or during, commissioning and monitored thereafter. The effect will be characterized and removed via a crosstalk matrix of 256x256 entries.

The pipelines are required to be capable of operating normally if a subset of the detectors is missing or non-functional.

4.2 Calibrating the images photometrically and astrometrically

Both photometric and astrometric calibration are required. In the case of a VISTA survey region the absolute photometric accuracies required (in J, H and K_s) are 0.02^m (with a goal of 0.01^m) on sources bright enough so there is negligible error due to photon noise. The VISTA astrometric requirements, to an airmass of 2, are:

differential astrometric accuracy of 0.1″ rms over the whole of the field covered by the IR mosaic.
differential astrometric accuracy of ≤ 0.03″ rms within the field covered by each individual IR detector.

absolute astrometric accuracy of $\leq 0.3''$ rms over the whole of the field covered by the IR mosaic.
for sources bright enough so that there is negligible centroiding error due to photon noise.

Source extraction software is run on the uncalibrated images to produce a catalogue for each detector, with positions based on the pointing position reported by the telescope to the FITS header. There will be enough sources in each detector field with positions known to sufficient accuracy (e.g. from 2MASS, USNO-B or SuperCOSMOS catalogues) to allow derivation of a FITS-standard World Coordinate System (WCS) using the Zenithal polynomial (ZPN) projection[14]. The fixed offsets between the various detectors (assumed stable, and if not then checkable) aid in the accuracy of this calibration. The field distortions due to the wide fields of view of the two instruments are represented to high precision by the ZPN projection used.

The same catalogues derived from the images are also compared with 2MASS catalogues to produce a preliminary calibration on the 2MASS photometric scale (with associated colour terms). Secondary standard fields for VISTA & WFCAM have been defined and will be measured as part of their observing programmes. These will enable a photometric calibration to be made in the VISTA photometric system by observing sufficient of the photometric standard fields each night to enable the pipeline to produce photometric zero points and extinction measures for each detector in the normal way. Again the multiple detectors can be used to refine the calibration as they will (usually) all experience the same extinction. Non-photometric nights will be flagged via error estimates in this calibration. The coefficients of this internal calibration will be written into the FITS headers of the images and those of the catalogues extracted from them. Because the calibration is not applied directly to the images or catalogues themselves, it will be relatively simple for the science archive to update the calibration in the light of longer term experience with the two instruments. The raw pixel data is archived at CASU, and the instrument artefact-free frames, and image catalogues from them, with associated astrometric and photometric calibration information are transferred up to the science archive at WFAU in Edinburgh.

Material in Sections 3 & 4 above is described in greater depth and accuracy in the accompanying paper by Irwin et al[1].

5 SCIENCE ARCHIVE

5.1 Access and curation

When each batch of the instrument artefact-free frames, and extracted catalogues, with associated astrometric and photometric calibration arrive (over the network) at WFAU they are immediately ingested into a database system as described in Hambly et al[2]. The actual pixel values in the images are not held in the database, but the location of each image is held, so that it can be accessed quickly when required. The catalogues are put into the database, and into faster access locations than is the pixel data, as it is anticipated that most users will want to mine catalogue data. At this stage the recent output from the pipeline can be made accessible to authorised users. Note that no stacking or mosaicing of pawprints has to be done in the pipeline. All such operations on the instrument artefact-free calibrated data is referred to as data curation.

At the level of each detected object it is clear that later observations in the same or different filters may cover the same piece of sky, so the catalogue database has to provide some 'best' multiband values for the parameters of each detected object, or some user specified means for deriving these. Combination of catalogues from single pawprints can however never throw up new objects not bright enough to be seen during the exposures responsible for the individual images. Therefore, as part of the curation process, images may be stacked and new catalogues derived from the stacked images. Time series of images may also be differenced to seek moving or photometrically variable objects, and such objects catalogued. Finally objects of interest are often selected from large surveys by their colours, so each catalogued object will have information available about its detection history in each wavelength band. For objects undetected in some filters in the original catalogues source detection software can be run at the position of the object to provide an upper limit (or detection) in the other filters.

These curation operations will be systematically performed on the available data to maintain its utility, or produced when required for a data release, (or, if resources permitted, for an authorised user). For example the products required for the UKIDSS surveys[8] will be available. Most of this routine work will be database driven. Many of the tasks involved, including deep stacking and mosaicing, merging catalogues, list driven photometry and overlap calibration, will use the combined skills of the pipeline team and the archive team.

Once these products are available they must of course be made available to the authorised users, who will be very numerous when the surveys become fully public, and no doubt very demanding on resources given the volume of data available. For this reason, and to decouple user support problems from the actual curation process, we have chosen to maintain two versions of the database on different servers, one that is continuously available to users and static at any time, and the other which is not available to users, because within it curation tasks are being continually carried out. The user interface to the more static version will follow the emerging VO standards as far as practicable, as described in detail by Hambly et al[2]. Data releases will be made periodically, through publishing static data sets from the curation database to the online server

5.2 Relationship to AstroGrid & the Virtual Observatory

The VDFS is in an exciting position within observational astronomy in that it is being designed during the genesis of the Virtual Observatory. Our goal from the outset has been that both the science archive (initially) and pipeline (ultimately) should be VO compliant and fully exploit the UK’s investment in AstroGrid, with the goal of running within the AstroGrid environment. VDFS has been designed to be able to exploit the AstroGrid framework and components. In the short term, the 'users' of the pipeline and curators of the science archive will predominantly be the VDFS operation teams, but as time progresses we intend that 'external users' will incresingly be able to invoke modules of the archive, and even the pipeline, as grid services that run via an interface within the VDFS science archive.

We also intend to enable users to mine the catalogues and images with their own code in due course, through VO interfaces. For example data treated to optimise detection of point and small sources might need processing with different algorithms (or the same algorithms with different paramaters set) to make it optimal for finding low surface brightness objects. External users will want to use their own algorithms (or specify parameters of those provided) for some processing modules to deal with such cases. The time at which all this capability begins to become available is not yet determined, in principle it should be as soon as possible, but in practise we first need to build experience operating on WFCAM and VISTA data as it arrives, and monitor developments elsewhere in the VO. Undoubtedly, with the data volumes of WFCAM and VISTA there are many classes of operation that will be hard to do elsewhere, and so such facilities will become increasingly essential as the data volumes grow. The ultimate goal would be to allow users to derive a complete survey in a specified manner reprocessing from raw data on the fly. Whilst this will not be a reality for some years it can be thought of as the ultimate goal towards which the VDFS intends to develop.

The science archive for WFCAM is described in greater depth and accuracy in the accompanying paper by Hambly et al[2].

6 CONCLUSIONS

The VDFS software is into its implementation phase and will be used on WFCAM data towards the end of 2004, and on VISTA towards the end of 2006. The required data rate is challenging for the pipeline, but the total accumulating volume of data is even more challenging for the science archive. Both can be handled, with the hardware help provided by Moore’s law, and with the data mining capabilities being developed in the context of the Virtual Observatory. We have adopted a spiral development model in which a series of increasingly complete prototypes are released. The companion papers on pipeline processing and science archive, presented at this conference by Irwin et al.[1], and by Hambly et al.[2], should be consulted for further details.

ACKNOWLEDGEMENTS

The VISTA Data Flow System is funded by grants from the UK Particle Physics and Astronomy Research Council.

REFERENCES

1. M. J. Irwin, P. Bunclark, D. Evans, S. Hodgkin, J. Lewis, R. McMahon, J, Emerson, S. Beard, M. Stewart, “VISTA Data Flow System survey access and curation: The WFCAM Science Archive”, in *Optimizing Scientific Return from Astronomy through Information Technologies,* P.J. Quinn & A. Bridger eds., *Proc SPIE* **5493**, paper 32, 2004.

2. N. C. Hambly, B. Mann, I. Bond, E. Sutorius, M. Read, P. Williams, A. Lawrence, J. Emerson., “VISTA data flow system survey access and curation: The WFCAM Science Archive”, in *Optimizing Scientific Return from Astronomy through Information Technologies*, P.J. Quinn & A. Bridger eds., *Proc SPIE* **5493**, paper 31, 2004.
3. D. M. Henry et al., “Design status of WFCAM: a wide Field camera for the UK infrared telescope”, in *Instrument Design and Performance for Optical/Infrared Ground-based Telescopes*, M. Iye and A. F. M. Moorwood, eds., *Proc. SPIE* **4841**, pp. 63-71, 2003.
4. J. Emerson and W. Sutherland, “Visible and Infrared Survey Telescope for Astronomy: overview”, in *Survey and Other Telescope Technologies and Discoveries*, J.A. Tyson & S.Wolff eds., *Proc SPIE*, **4836**, pp. 35-42, 2002
5. A. M. McPherson, A. Born, W. Sutherland, J. Emerson, “The VISTA Project: a review of its progress and lessons learned developing the current program”, in *Ground-based Telescopes*, J.M. Oschmann & M. Tarenghi eds., *Proc SPIE* **5489**, paper 46, 2004.
6. G. Dalton, M. Caldwell, K. Ward, M. Strachan, P. Clark, W. Sutherland, J. Emerson, “The VISTA IR Camera”, in *Ground-based Instrumentation for Astronomy*, A.F.M. Moorwood & M. Iye eds., *Proc SPIE* **5492**, paper 34, 2004.
7. S. Warren, “Scientific goals of the UKIRT Infrared Deep Sky Survey”, in *Survey and Other Telescope Technologies and Discoveries*, J.A. Tyson & S. Wolff eds., *Proc SPIE*, **4836**, pp. 313-320, 2002
8. W. D. Pence, “New image compression capabilities in CFITSIO”, *Proc. SPIE* **4847**, pp. 444–447, 2002.
9. N.A. Walton, A. Lawrence, A.E. Linde, “Scoping the UK's Virtual Observatory: AstroGrid's Key Science”, in *ADASS XII, ASP Conference Series*, eds. H. E. Payne, R. I. Jedrzejewski, and R. N. Hook, **295**, p.25, 2003
10. JAC to CASU Interface Control document (for WFCAM: Telescope Data Storage to Pipeline), www.jach.hawaii.edu/JACpublic/UKIRT/instruments/wfcam/ICD, v1.2, Oct 2003
11. N. Hambly, M. Irwin. J. Lewis, WFCAM Science Archive Interface Control Document: (Pipeline to Science Archive) http://harris.roe.ac.uk/~nch/wfcam/VDF-WFA-WSA-004-I3/VDF-WFA-WSA-004-I3.html , Oct 2003
12. www.ukidss.org
13. I. Tuomi, “The lives and death of Moore's Law." First Monday (peer-reviewed journal on the internet), Volume 7 No. 11: http://www.firstmonday.dk/issues/issue7_11/tuomi
14. M.R. Calabretta and E.W. Greisen, “Representations of celestial coordinates in FITS”, *A&A* **395**, pp. 1077-1122, 2002