# VISTA DATA FLOW SYSTEM (VDFS)

for VISTA & WFCAM data

## WFCAM Science Archive : Overview

**author**
A.Lawrence (WFAU Edinburgh)
VDFS co-investigator

**number**
VDF-WFA-WSA-001
**issue**
issue 1.0
**date**
01 Apr 2003

**co-authors**
N.Hambly, P.Williams, I.Bond, E.Sutorius, M.Read, R.Mann, H.MacGillivray

# Contents

# 1 SCOPE

This document is an overview of the WFCAM Science Archive (WSA). It is part of the document collection for the VISTA Data Flow System, which includes WSA as a testbed project. It is a general purpose summary which acts both as an internal project document and as a readable source of general information for interested external parties.

# 2 CONTEXT

## 2.1 The VISTA Data Flow System project

VISTA is a new dedicated 4m survey telescope designed to have both IR and optical cameras operating from 2006, although the optical camera is currently unfunded. It began as a UK initiative, but became an ESO telescope as part of the UK's accession to ESO. VISTA presents considerable challenges both in sheer data volume terms and in terms of presenting transparent data mining opportunities to astronomers across Europe in a developing Virtual Observatory context. The VISTA Data Flow System (VDFS) is a UK project aimed at being ready to cope with these aspects of VISTA. The UKIRT Wide Field Camera (WFCAM) is an important intermediate step both in data rate (20TB/year versus VISTA's expected 100TB/year) and in time (survey data arriving from 2004 and 2006 respectively). A decision was therefore taken to take an integrated approach to WFCAM and VISTA. The WFCAM project is scientifically important in its own right, but also acts as a *testbed* for VDFS. The testbed approach applies not just to general experience, but to actual design and construction. The pipeline and science archive systems for WFCAM should lead directly on to the corresponding VISTA systems.

## 2.2 The UKIRT Wide Field Camera

WFCAM is a wide field IR array camera being built for UKIRT. It is under construction at the UKATC at ROE (see the WFCAM project web page at http://www.roe.ac.uk/atc/projects/wfcam). The current schedule has WFCAM being delivered to Hawaii in November 2003, with science observations starting early in 2004. WFCAM is based on a novel optical design using a quasi-Schmidt camera placed forward of Cassegrain focus that gives a 1 degree field of view. It will be used in conjunction with UKIRT's tip-tilt secondary which primarily corrects for wind-shake and dome seeing and is expected to deliver median seeing of 0.4" across the whole of this field of view. The camera employs four Rockwell PACE $2048^2$ HgCdTe arrays operating across the JHK spectral range, with 0.4" pixels, and placed at 90% spacing. The edge to edge diameter is therefore 0.66 degree, and the instantaneous field of view is 0.21 sq.deg, in four separated areas. The pixel-size is somewhat under-sampled with respect to the median seeing, so micro-stepping will normally be employed, producing smaller final data pixel-steps. WFCAM will be capable of a variety of micro-stepping patterns, but it is expected that it will normally be used in a $2 \times 2$ pattern. A filled "tile" on the sky can be made by making a $2 \times 2$ macro-step pattern, which together with the micro-stepping would produce an image with 0.2" pixel-pitch and diameter 0.87 degrees. However, a variety of mosaic patterns will be used.

Key facts :

- pixel size : 0.4 arcsec
- array size : 2048 pixels = 13.65 arcmin
- spacing = 90% width = 12.29 arcmin
- total diameter = 2.9 array diameters = 39.59 arcmin
- instantaneous field of view = 4 x $13.65^2$ = 0.21 sq.deg.

### 2.3 Science with WFCAM

WFCAM will be used for both pre-planned public survey operations (through the UKIDSS project), and for competitive smaller allocations of time (through the PATT system). The pipeline and archive supports all WFCAM users, but its requirements are driven by UKIDSS. The five UKIDSS surveys are as follows :

- Large Area Survey (LAS) : 4000 sq.deg. K=18.4
  short exposures; two passes
- Galactic Plane Survey (GPS) 1600 sq.deg. K=19
  short exposures ; two passes; some deeper components
- Galactic Cluster Survey (GCS) 1400 sq. deg. K=18.4
  short exposures; two passes
- Deep Extragalactic Survey (DXS) 35 sq.deg. K=21
  long exposures, stacked over several nights
- Ultra Deep Survey (UDS) : 0.75 sq.deg. K=23
  very long exposure, stacked over many many nights

### 2.4 Responsibilities for WFCAM

End-to-end responsibility for WFCAM lies with the Joint Astronomy Centre (JAC), who operate UKIRT. The UKATC has responsibility for camera construction. JAC are responsible for WFCAM operations. The pipeline and archive is a collaboration between JAC, Cambridge Astronomical Survey Unit (CASU) and Edinburgh's Wide Field Astronomy Unit (WFAU). CASU is primarily responsible for the pipeline, and WFAU for the science archive. The science programme divides into two parts. For PATT-allocated time, PIs are of course responsible, but can expect data to be more or less *pret a porter*, provided their mode of observing adheres to one of a small number of standard protocols. For UKIDSS surveys, the consortium are responsible for survey design and staffing of operations, but are not responsible for data reduction and archiving, which are done by WFAU and CASU. Actual scientific analysis, expected to be done primarily through use of the WSA, is open to all astronomers across Europe, and after a short proprietary period to all astronomers worldwide.

## 3 WFCAM OPERATIONS AND DATA FLOW

All WFCAM operations will be in a queued mode, using UKIRT's Observing Management Protocol (OMP) system. There will be very few observing modes. For both UKIDSS and open-time observations fixed standard calibration procedures will be enforced. All these simplifying factors make a standard pipeline possible for all WFCAM data. Of course no standard pipeline will ever squeeze all the possible information out of the data, and there will always be users, or types of question, that require different assumptions or algorithms in the processing. We are not attempting to construct an all-purpose completely flexible pipeline toolkit, but rather a processing pipeline to produce pre-agreed standard data products. These should be good enough for most purposes, but the design is optimised to produce the survey products demanded by the UKIDSS project, with optimal stacking, mosaicing and source extraction, and uniform astrometric and photometric calibration across survey fields. These demands are embodied in the **Top Level Requirements** and **Science Requirements Analysis Document (AD01)** agreed with both JAC and the UKIDSS consortium.

The overall system flow is illustrated in Fig. 1. There are several stages - data acquisition, the summit pipeline, the standard pipeline, ingestion to the archive, further survey-wide processing, refinement of calibration, and serving the data to users. At the summit, data from all exposures within a single night at the same telescope pointing position (including micro-steps within that position) and using the same filter, are co-added on the spot by the *Data Acquisition System (DAS)*. Regardless of the macroscopic dither pattern through the night, data from the four arrays are kept distinct, as are data from different pointings, so that the data written to media consist of a collection of $4096^2$ co-added *frames* within one night. The main purpose of the *summit pipeline* is to generate near real-time Data Quality Control (DQC) information from the co-added frames. It will use fixed library frames for instrument signature removal (e.g. flat fielding) and will do a first cut source extraction. The reduced frames, DQC, and a statistical analysis of the source lists, will be examined in Hawaii by UKIRT and or UKIDSS staff and used to update a *survey progress database*, which then feeds back to the observing queue.

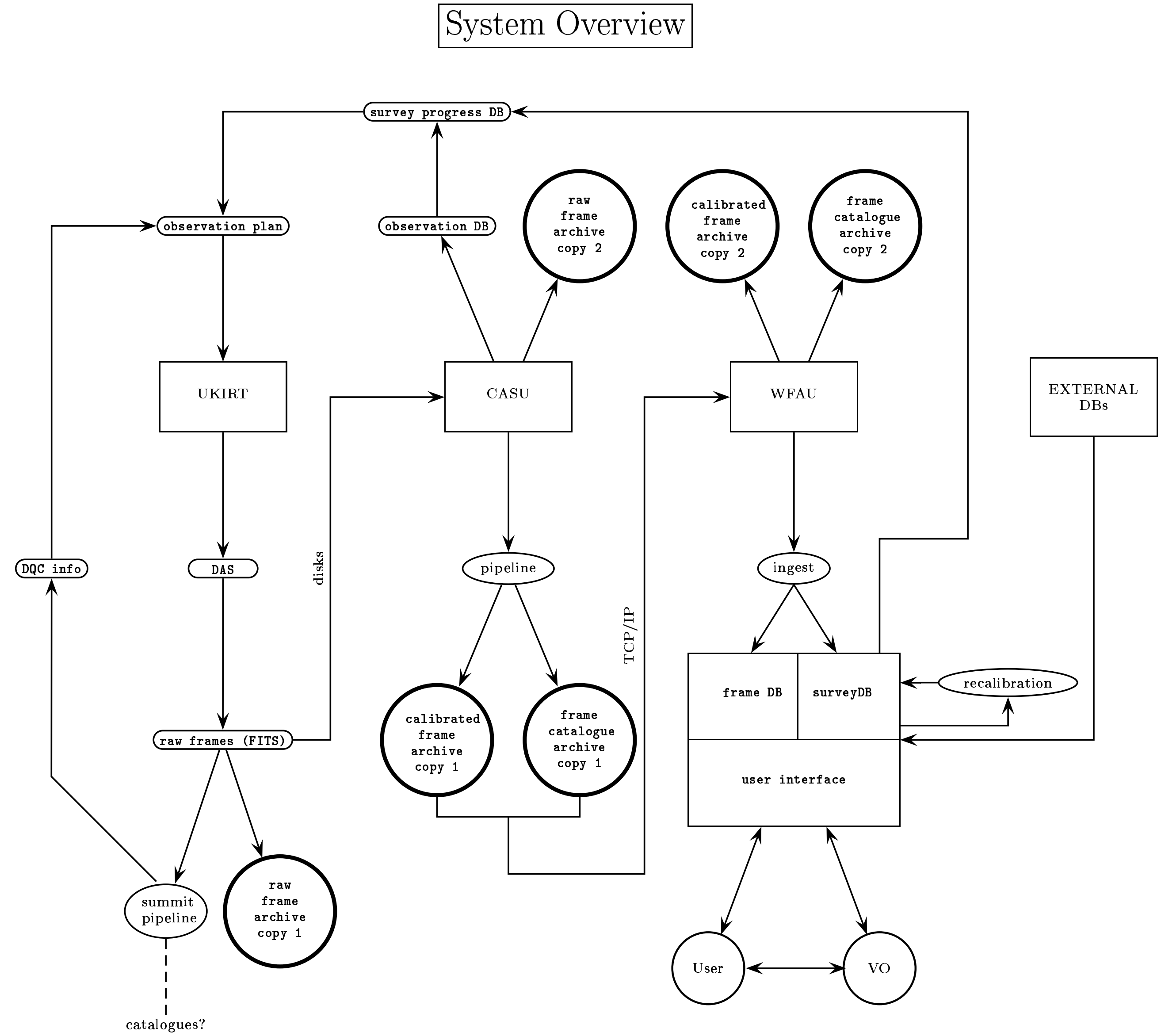

Figure 1: *Illustration of overall data flow for WFCAM*

The raw data (i.e. the collection of co-added 4K frames from each night) will be archived in Hawaii but also sent to the UK on a daily basis for placing in a duplicate archive in Cambridge, where further processing with the *standard pipeline* take place. The data will be sent either on a nightly tape, or on a hot-swappable hard disk drive, as used by the ESO NGAST system. (Final review and decision is due shortly). The standard pipeline will use the same software as the summit pipeline, but it will process the real calibration data, and will estimate a separate PSF from each frame. The pipeline removes instrument signature, does a photometric and astrometric calibration, and a source extraction. The result is a calibrated version of the collection of 4K frames, and a separate source list for each frame.

The calibrated frames and associated source lists will be archived in Cambridge, but also sent to Edinburgh for placing in a duplicate archive. Transport to Edinburgh will be on a nightly basis over the Internet. A series of *further processing steps* is needed to make final survey products, but these steps can only be carried out as the survey data accumulates in the science archive, so they need to be especially carefully planned between the CASU and WFAU teams. The most obvious thing is optimal stacking of matching frames from different nights, and mosaicing to produce a final large pixel-map for each survey. (Once again the first aim is a standard single pixel map, but we will also develop the facility to build images of any given sky-area on the fly from the constituent frames using different sampling and stacking choices). The next step is improved PSF generation including variations over the field and within a frame stack. Next there is improved source extraction from the stacked data, including detection and parameterisation of Low Surface Brightness Objects and transient events. Then we have *pairing* of sources across catalogues in different filters to make YJHK colours, and pairing with objects in external catalogues, such as the SDSS. Finally, as large area surveys are accumulated, we will revisit astrometric and photometric calibration looking for systematic gradients and step functions, and making external checks, and eventually, over several years, deriving proper motions and variability parameters.

Finally the data is ingested into a *public science archive* housed in Edinburgh, with a user friendly interface and fast search and analysis capabilities. This will have both interactive and batch modes, and all data in it will be well calibrated and documented. It will contain all the final survey pixel maps and paired source-lists, but will also contain all the constituent nightly 4K frames and their standard one-filter source lists, as well as a browsable database of the available frames. The functionality delivered by this science archive will come in three stages, as described in the next section.

# 4 SCIENCE ARCHIVE FUNCTIONALITY

## 4.1 General Requirements

The archive is a *Science Archive*, not a repository. Our expectation is that a large fraction of the science done with WFCAM will be done through the science archive. It is also the first major astronomical archive required to be “VO-ready” from the outset, as opposed to being retro-fitted. These considerations puts four key requirements on the archive :

- The data must be reliably **calibrated**, and include data quality information.
- **Tools** for querying and analysis must be provided, as well as data download.
- The whole system must be **documented**.
- Methods and formats must be **VO compatible**.

User expectations are higher than can be met initially. This, together with pragmatism concerning project development, leads us into delivering WSA functionality in three distinct stages, as detailed in sections 4.3 to 4.5.

## 4.2 User expectations in the VO era

Straightforward *data access* is the standard today. One can offer either distinct data subsets such as plates or frames, chosen from a browsable list, and downloadable via a hypertext link, or pixel maps and catalogues over user-defined areas, created from some survey database.

With the second level of service, *complex queries*, one can construct questions in SQL or a similar language, along the lines of "give me a list of objects redder than X in this area of sky, with measurable proper motion, that have such-and-such quality flag better than Y (unless R-mag is brighter than Z, in which case accept anything), and that were found on a Tuesday". This kind of service is becoming slowly more common now. An example is the bf SDDS science archive, or the **6dF Galaxy Redshift Survey**. The leading edge now, which can be seen as Level-2a, is in *federated queries*, i.e. the ability to make joint queries of arbitrary databases distributed round the world - e.g. "give me all the objects in the UKIDSS LAS survey which were not seen in the SDSS but do have an X-ray ID in either a Chandra or an XMM observation, and check the list against the ESO VLT observing log". This is the key problem being tackled by the various VO projects worldwide. Although truly distributed queries will soon become common, in the short term we can facilitate this kind of science by *warehousing* selected external databases along with the WSA.

The third level of service involves *database manipulations*. What we have in mind is things like calculation of correlation functions, cluster analysis in N-D parameter space, making statistical digests so that one can find objects $5\sigma$ outside the main clump, visualisation and exploration of multi-faceted datasets, and so on. Today, such data-intensive calculations are the province of specialised "power users" on their own machines, but we expect that such calculations will increasingly be provided as a standard, fast, service at the data centre, and that it will become common to do exploratory analysis this way as well as rigorous calculations. Furthermore, because of the increasingly large archive volume and network limitations, it will be more practical to use a service provided by a data centre than to download huge amounts of data and hack your own code.

## 4.3 Version-1 Science Archive

This will be available immediately following first light (January 2004). It will allow access to the basic data products, will have a web-based interface, and documentation. It will include access to 2MASS, SSS, SDSS-DR1, and USNO-B catalogues as well as WFCAM data. It will allow the user to :

- Browse the database of calibrated observation frames and associated catalogues, and download any one of these
- Produce small images and associated multi-colour merged catalogues, from either the observation frame collection, or the accumulating UKIDSS surveys
- Produce mosaic images on demand up a width of 0.8 degrees
- Search UKIDSS surveys on a position rectangle or circle in Celestial or Galactic co-ordinates, and within a radius of a resolveable source name
- Allow joint position searches as above, and produce merged catalogues, with the combination of UKIDSS plus any of 2MASS, SSS, SDSS-DR1
- Make SQL-like queries in any sensible arithmetic combination of the catalogue parameters, for example colours
- Allow general queries on the combination of UKIDSS, 2MASS, SSS, SDSS-DR1

### 4.4 Version-2 Science Archive

This is intended to be available one year after survey operations begin (i.e. Jan 2005). It will be compliant with any relevant protocols and formats produced as part of the various Virtual Observatory initiatives, and will work especially closely with AstroGrid. It will allow the user to :

- Search open-time data as well as UKIDSS data
- Plot returned parameters in histograms
- Plot pairs of returned parameters as (x,y) graphs
- Generate on request images and catalogues across observation frame boundaries
- Generate mosaic and stacked images and source catalogues from any area, from user specified observation frames
- Choose a range of stacking and source extraction algorithms on request, with a menu of tuneable parameters

Probably other simple user tools will be available, such as the *SED tool* developed for the AVO first-light demo with the GOODS dataset. Version-2 of WSA will also :

- Return catalogues and similar data in VOTable format, and images using binary data access protocols still under development
- Recast all the V1 services as *web services* for publication in the AstroGrid Registry

### 4.5 Version-3 Science Archive Archive goals

Version-1 and Version-2 represent two stages of functionality to which we are committed. However it is our *goal* to go beyond this and provide server-based *data analysis tools.* We hope, given further resources, to :

- Develop (or install) a suite of advanced visualisation tools, e.g. pannable large area imaging, multi-dimensional catalogue plotting and rotation
- Develop (or install) a suite of data analysis tools, e.g. cluster analysis, principal component analysis, etc
- Implement both the above as server-based tools
- Develop a system to allow uploadable user-specified analysis algorithms
- Recast our web services as *grid services*, in order to allow queries and analysis using *shared managed resources* with other data centres

# 5 WFCAM DATA PRODUCTS

## 5.1 Data produced at telescope

Integrations are made up of many short exposures, most of which are accumulated on the spot by the *Data Acquisition System.*

- **exposure** = smallest unit
  normally $t_{exp} = 10$ sec
  can have multiple non-destructive readouts (NDRs)
- **stare** = sum of repeated exposures before any movement
  $t_{stare} = R \times t_{exp}$ ; often $R = 1$
- **pointing** = interleaved set of $N \times N$ microstepped stares
  $t_{point} = N^2 \times t_{exp}$ : usually $N = 2$
- **observation** = accumulated pointings at a single telescope position at a single epoch
  $t_{obs} = M \times t_{point}$
  for shallow surveys $M = 1$

Only the data accumulated during an *observation* are recorded. The data from a single array form a *frame* which is the fundamental data product. Data from all four frames at a single observation is a *multiframe.* Note that there will be *calibration frames* as well as *science frames.* The *summit pipeline* will produce many intermediate data frames which are not kept, but will also produce *summit catalogues*, and *Data Quality Control* (DQC) information which is recorded with the saved frames. In addition, a *survey progress database* will be accumulated, from which the *WFCAM schedule* will be derived. As well as being updated daily in Hawaii, information from the far end of the processing chain, i.e. the accumulating completed surveys, will be fed back into the survey progress database. In summary, the data products at this stage are :

- Raw frame collection (including DQC information)
- Summit catalogue collection
- Survey progress database

## 5.2 Archived pipeline data products

The standard pipeline instrumentally corrects each *frame* as best it can, produces the best possible astrometric and photometric calibration for that single frame, determines the PSF, and extracts a *source catalogue.* The collection of *calibrated frames* and their associated catalogues and housekeeping data is the basic archived product. We also produce a database of these frames, which can be browsed and queried, so that individual frames can be downloaded. (The **frame collection**, which is a collection of flat FITS files, should be carefully distinguished from the **frame database** which is a structured database within a DBMS system, but containing only the frame metadata, with links to the actual data files).

Further processing stages carried out at the science archive build the actual UKIDSS survey products, stacking and mosaicing the frames, adjusting the calibration across the whole survey, and producing final merged source catalogues, including colours, proper motions, and variability measures. A similar recipe could in principle be applied to user-defined surveys with open-time data.

The following products will be housed at CASU :

- raw frame collection - Copy 2
- calibrated frame collection - Copy 1
- frame catalogue collection - Copy 1
- library calibration frame collection - Copy 1

The following products will be housed at WFAU :

- calibrated frame collection - Copy 2
- frame catalogue collection - Copy 2
- library calibration frame collection - Copy 2
- frame metadata database
- frame catalogue database
- UKIDSS survey final (accumulating) pixel maps
- UKIDSS survey final (accumulating) merged catalogue database
- photometric calibration database
- astrometric calibration database

Note that the flat-file collections are all duplicated at CASU, but the databases are not, so there will be a tape-backup system for these elements.

### 5.3 On demand data product types

Users will want subsets and combinations of data from the archive. Furthermore the standard UKIDSS survey products will not always be the last word in how one can combine frames or extract sources. We need therefore the ability to construct various standard data product types on demand from the frame collection and/or UKIDSS survey data. Those envisaged are as follows :

- Small image - subset of one frame, or area from UKIDSS pre-merged survey, plus associated catalogue
- Finder chart (small images with overlaid object ellipse plots)
- Colour image - from user specified combination of frames
- Mosaic image - i.e. crossing frame boundaries
- Big mosaic image - offering blocked down or smoothed pixels
- Stacked image - with selectable algorithm
- Difference image
- Merged source catalogue - from user specified frame combination

# 6 WSA DESIGN ISSUES

The WSA project has several overlapping development phases - initial conception, requirements capture, design and analysis, build, and integration and test. (See the **Management Plan** (AD02) for details). At the current version of this documeent, we are just completing the analysis and design phase. The result is a series of detailed design documents (AD03-07), and a Critical Design Review scheduled for April 15-16 2003. In this section we summarise some of the key issues and decisions that arose during the analysis and design phase. The overall logical design of WSA operations is illustrated in Fig. 2.

## 6.1 DBMS design

After an initial review and substantial experimentation (described in section 7) we decided that the WSA would be built around a commercial relational DBMS. (This only applies to the structured database components of the archive - the pixels will be stored in a flat file collection with links to the necessary databases). Key factors in this decision included the following. (i) Minimising work by collaborating closely with our colleagues in the SDSS science archive at JHU and Fermi-Lab, using their experiences and occasionally their actual software. (ii) Maximising ease of working with other astronomical databases and projects, again especially with SDSS. (iii) Ease of use, both in system development and curation. (iv) License fee, and the academic discounts available. (v) Scaleability to VISTA volumes. (vi) Collaborative possibilities with database vendors. We have formed good relationships with individuals in the research arms of all of Microsoft, IBM, and Oracle.

Based largely on factors (i), (ii), and (vi), i.e. the SDSS-DR1 system, and a working collaboration with Jim Gray of Microsoft Research, we have decided to implement Version-1 of the WSA using MS SQL Server on a Windows cluster. However, there are some questions over parallelisation of SQL Server, and hence its scaleability, so the Version-2 implementation is still considered open, with IBM DB2 the other strong contender.

We are now proceeding with detailed database design, described in more detail in the **Database Design Document** (AD02). The **Interface Control Document** (AD04) agreed with CASU specifies the detailed output from the pipeline process, the general principles of which have been agreed with the UKIDSS consortium. Based on these and on the functionality requirements, we have developed a series of Relational Models, which will then lead onto to DBMS schema for various tables, and thus to scripts for ingestion.

## 6.2 Hardware isssues

**Data Volume Requirements**. The table below shows the estimated volume of the main data products, after one year and after seven years (the estimated duration of the UKIDSS surveys). Core-parameter catalogues contain only the most important object parameters, which will satisfy many queries.

**Access, search and analysis requirements**. For pixel subsets and small catalogue collections, **access** should be fast as long the data are sensibly organised and linked, so this is not a hardware driver. Likewise most common SQL-like **queries** should be fast as long as the database tables are indexed. This is non-trivial for proximity searches where two parameters are related by some kind of metric (for example position on the sky), but such problems have been the subject of much research over the last few years leading to solutions such as HEALPix and HTM. However, there will always be circumstances where there is no alternative but to search a large fraction of the database, so we need to maximise the em trawl rate, which will be I/O limited. Typical disk-CPU bandwdiths are around 10MB/s, so that to trawl catalogues of around 100GB would take several hours. From the

Table 1: *Estimated WFCAM data rates/sizes*

| | |
|---|---|
| Peak data flow at instrument | 12 MB/s |
| Peak recorded data rate | 230 GB/nt |
| Average recorded data rate | 100 GB/nt |
| Frame data accumulation | 20 TB/yr |
| Catalogues and ancillary data accumulation | 2 TB/ yr |
| 2010 archive : frame+cat data | 154 TB |
| 2010 archive : stacked survey maps | 50 TB |
| 2010 archive : full survey object catalogues | 5 TB |
| 2010 archive : core parameter catalogues | 500 GB |

consultation with UKIDSS, we have set a goal of 100 seconds for archive-response for such trawls. This is a significant hardware constraint, requiring some kind of I/O parallelism. **Analysis** will often mean relatively simple pixel transforms, which will not set significant hardware constraints. However, more ambitious types of datamining, such as computation of correlation functions, will be strongly CPU bound, and are likely to benefit from CPU parallelism - i.e. either a PC cluster, or a proper SMP machine. This is a Version-3 issue, so we have not yet set a formal requirement.

**Curation Requirements**. The plan is that calibrated frames and catalogues will be transported from CASU to WFAU on a nightly basis over the Internet. The expected average rate is of the order 100GB/day. For night-time transfer we then need therefore to achieve *transmission bandwidth* of at least 5MB/s. As the data arrive, we need to make sure that ingestion is not held up by external use of the databases. We therefore need a *separate load server*. Finally we need to consider data security. This is mostly achieved by having two spinning-disc copies of each major collection/database (e.g. raw frames at JAC and CASU; calibrated frames at CASU and WFAU). However at the far end of the chain, final survey data will be backed up on a tape library.

**Hardware solutions**. Our main conclusion is that we should have distinct hardware solutions for pixel storage, queries and analysis. The proposed solution for **pixel storage** closely follows the ESO NGAST method, with multiple units of PC plus IDE RAID, costing something like 2K/TB. For our **query engine**, I/O parallelism is achieved by hanging many RAID-striped discs off one high-end processor, adding up to around 2TB. We should be able to get close to the limit set by the PCI bus, of around 1GB/sec. We envisage three such systems - two as catalogue servers, and one as load server. The solution for **analysis engine** has been deferred. For **bulk data transmission**, the Janet backbone and our local network are easily sufficient, with the bottleneck on the *end stations*. The solutions are mostly in software - e.g. tuning TCP buffers, and use of GridFTP. We expect to reach the necessary bandwidth.

## 6.3 User interface

The overall configuration of the user interfacee is illustrated in Fig. 3. The user interface will be web-based throughout. However both user expectations and the technology available in this area are developing fast, so the three-stage approach to functionality is especially important here. For **Version-1** we need to provide three types of functionality. (i) Access to frames and small catalogues, and the ability to select from the collection. This will be achieved via a DBMS holding the metadata, but it requires both a *query form* and an *interactively browsable interface*, which will in the first instance be just a table of links. The return will be the request files. (ii) The ability to make SQL-like queries, and return tables, in a variety of formats. We will provide a *button driven query interface* which will construct SQL statements on the user's behalf, as well as allowing direct SQL entry. (iii) Request of small pixel images on demand, returning either FITS or JPEG images. Requests will be through a simple *parameter range form.*

For **Version-2**, the first priority will be to update the technology of the interface, for example to pass SOAP messages rather than just a CGI stream, to return XML (e.g. VOTable rather than just FITS tables), and to implement the web-interface dynamically and configurably, through a portal technology such as Cocoon or Zope. There will also be increased functionality. For the *pixel processing engine*, we need to deliver a larger range of on-demand products, but this will not change the UI very much. However we also wish to provide some simple *user tools* such as XY plotting of returned tables, or SED plots. We expect that these tools will be written elsewhere, and our job is to implement them either as applets (like Aladin) or as server applications delivering html output to the user's browser. For **Version-3** we hope to implement more advanced tools, server based streaming visualisation, and so on. The impact of these on the UI is not yet designed.

### 6.4 Curation

Operation of the archive implies both development of software components and continuing operational staff effort. As well as ingestion and controlled release of useable products, the science archive has the responsibility to merge the survey data and produce final survey-wide calibration. Curation tasks divide into four relevant timescales.

On a **daily basis** :

- transfer data from CASU, verify it, and log it
- ingest data into the frame collections and DBMS
- create compressed image products

On a **periodic (weekly/monthly) basis** :

- create "library" frame products - e.g. difference images etc
- recalibrate astrometry and photometry
- create spatial indices
- compute proper motions and other multi-passband derived quantities
- produce list driven quantities across WFCAM passbands

On an **occasional basis** :

- ingest external catalogues
- create default joins with external catalogues
- create variety of list driven quantities with external pixel data

For **releases** :

- create “library” survey products - stacked pixels, revised catalogues
- create “world access” subsets depending on date
- verify, freeze, release and backup
- place frozen products online

Note that we do not archive all past versions of the data products - only the latest versions.

### 6.5 Software components

Delivering the intended functionality requires software components in several separate areas. (i) The DBMS, and associated scripts. (ii) The user interface requires a mixture of HTML pages, CGI scripts, and Java classes. It also needs an interface to the DBMS. This will use Apache Tomcat to strip out the SQL from the HTTP stream and pass to the DBMS. (iii) Software components are associated with all the curation tasks. These will vary from simple shell scripts to C or C++ programmes. (iv) The data services we intend to deliver implies a series of *processing engines* behind the scenes, which will be written in C or C++. this includes a basic *cut-out engine* which creates and delivers image subsets, and does on-the-fly pairing and so on as necessary, and a *pixel processing engine* which creates stacked images, large images across frame boundaries, and so on. Our goals include an *advanced pixel engine* which can do streaming visualisation; a series of *user tools* for datamining (most of these will be written outside the project and wrapped within the system); and a system to allow users to upload their own analysis algorithms, which implies some kind of *script engine* converting scripts to WSA-native code. These goals are only loosely defined so far.

## 7 EXPERIMENTAL AND PROTOTYPE DEVELOPMENT PROGRAMME

Our design approach has been close to a traditional waterfall, with requirements followed by design and analysis and then build. However hand-in-hand with this top down approach we have built a series of bottom up prototypes which have tested ideas, trained staff, and influenced the top-down design. Many of the other projects the WFAU has undertaken in recent times have been at least in part, and sometimes mostly, undertaken as WSA prototypes.

One key issue has been **data volume and scaleability**. The prototype here is the SuperCOSMOS Science Archive (SSA), for which we have dealt with approximately 10GB/day for several years, and a multi-TB archive. WFCAM will involve 100GB/day from 2003, and tens to hundreds of TBs of on-line data. The WSA itself is a testbed for the VISTA Science Archive (VSA) which will deal with 600GB/day from 2006. From the *curation* point of view, including survey-wide processing and calibration, we expect the SuperCOSMOS experience to apply directly, and on to VISTA. From the hardware point of view, we have come to the conclusion that the SuperCOSMOS solution is not scaleable and we must take a new approach, especially looking towards VISTA. The WSA is more than just a testbed for VISTA. It is being developed as part of an integrated project with the VSA in aim. Our intention therefore is not just that the ideas and concepts, but actual software components, are re-used for VISTA.

The second key issue has been **deployment of commercial DBMS systems**. We first explored *object oriented databases* by installing a mirror of the SDSS EDR system using Objectivity. For a variety of reasons we abandoned this route and explored relational systems. It was vital to get hands-on experience with real data, and delivering a real service, as soon as possible. We therefore made the decision to use MS SQL Server for Version-1, and ingested the 6dF Galaxy Redshift Survey (6dFGRS) into this system and released it as a public service. We have also installed a subset of the SuperCOSMOS data into SQL server and we are using it for trawl rate experiments. We intend shortly to ingest the whole database, and release a new public version of the SSA, with the user interface following the WSA and VSA intentions as closely as possible. Meanwhile we are also ingesting the SuperCOSMOS data into DB2, but only for experimental purposes, not as a public service.

The third key issue has been **web technology and the user interface**. Our existing SuperCOSMOS service is web based and uses simple parameter range forms, passed through a CGI script. To prototype an SQL-query interface to an RDBMS, we followed the example set by the SDSS SkyServer system, deploying a similar system to the 6dFGRS public archive. This included experience with new pieces of technology, such as Tomcat, as well as new styles of ergonomic design. The Version-1 user interface will be different only in detail.

Next, we have undertaken a significant **hardware test programme** in order to arrive at purchase decisions. In this we have been lucky to have the help of Eclipse Computing, who have loaned a variety of different systems, for example to compare SCSI, IDE, and fibre linked discs, etc. Most of this work has been aimed at maximising aggregate I/O and checking scaling versus number of discs, RAID level, etc. We have also been undertaking bulk data transfer tests between WFAU and CASU, tracing routes and bottlenecks and so on. Substantial improvements can be made by firewall upgrading and by TCP tuning, but we are also likely to install *Globus* at both sites, so as to use GridFTP, which opens up several FTP sessions in parallel.

Finally we have prototyped **processing components** such as stacking and mosaicing software. Many of these are being supplied from outwith the WFAU, so prototyping includes getting working versions of these components for testing integration and functionality. Experimentation with external components is part of the process of deciding when to use them and when to cook our own.

**__oOo__**

## 8 ACRONYMS & ABBREVIATIONS

ADnn : Applicable Document No nn
ATC : Astronomy Technology Centre
CASU : Cambridge Astronomical Survey Unit
DQC : Data Quality Control
ESO : European Southern Observatory
PATT : Panel for the Allocation of Telescope Time
PSF : Point Spread Function
SDSS : Sloan Digital Sky Survey
VISTA: Visible and Infrared Survey Telescope for Astronomy
VPO : VISTA Project Office
UKIDSS : UKIRT Infrared Deep Sky Survey
UKIRT : UK InfraRed Telescope
WFAU : Wide Field Astronomy Unit (Edinburgh)

## 9 APPLICABLE DOCUMENTS

| | | |
|---|---|---|
| AD01 | Science Requirements Analysis Document | VDF-WFA-WFCAM-002<br>Issue: 2.0 15/04/003 |
| AD02 | Management and Planning | VDF-WFA-WFCAM-003<br>Issue: 1.0 15/04/003 |
| AD03 | WSA Interface Control Document | VDF-WFA-WFCAM-004<br>Issue: 1.0 15/04/003 |
| AD04 | WSA Data Flow Document | VDF-WFA-WFCAM-005<br>Issue: 1.0 15/03/003 |
| AD05 | WSA Hardware/OS/DBMS design documeent | VDF-WFA-WFCAM-006<br>Issue: 1.0 15/04/003 |
| AD06 | WSA Database Design Document | VDF-WFA-WFCAM-007<br>Issue: 1.0 15/04/003 |
| AD07 | WSA User Interface Document | VDF-WFA-WFCAM-008<br>Issue: 1.0 15/04/003 |

## 10 CHANGE RECORD

| **Issue** | **Date** | **Section(s) Affected** | **Description of Change/Change Request Reference/Remarks** |
|---|---|---|---|
| Issue 1.0 | 1/04/03 | All | New document |

## 11 NOTIFICATION LIST

The following people should be notified by email whenever a new version of this document has been issued:

**__oOo__**

| | |
|---|---|
| **WFAU:** | P Williams, N Hambly |
| **CASU:** | M Irwin, J Lewis |
| **QMUL:** | J Emerson |
| **ATC:** | M. Stewart |
| **UKIDSS:** | S Warren, A Lawrence |
| **JAC:** | A. Adamson |

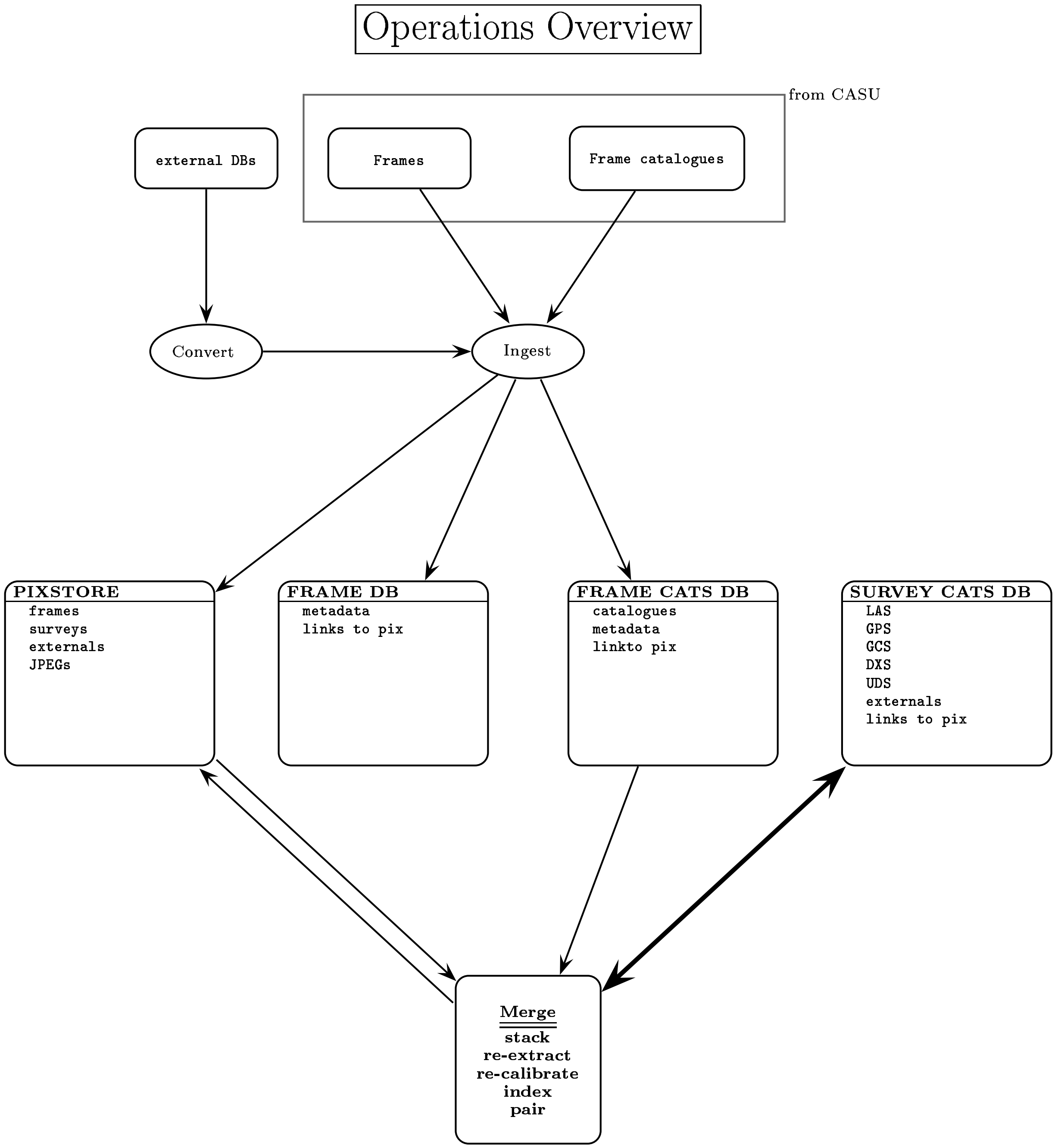

Figure 2: *Illustration of overall WSA operations*

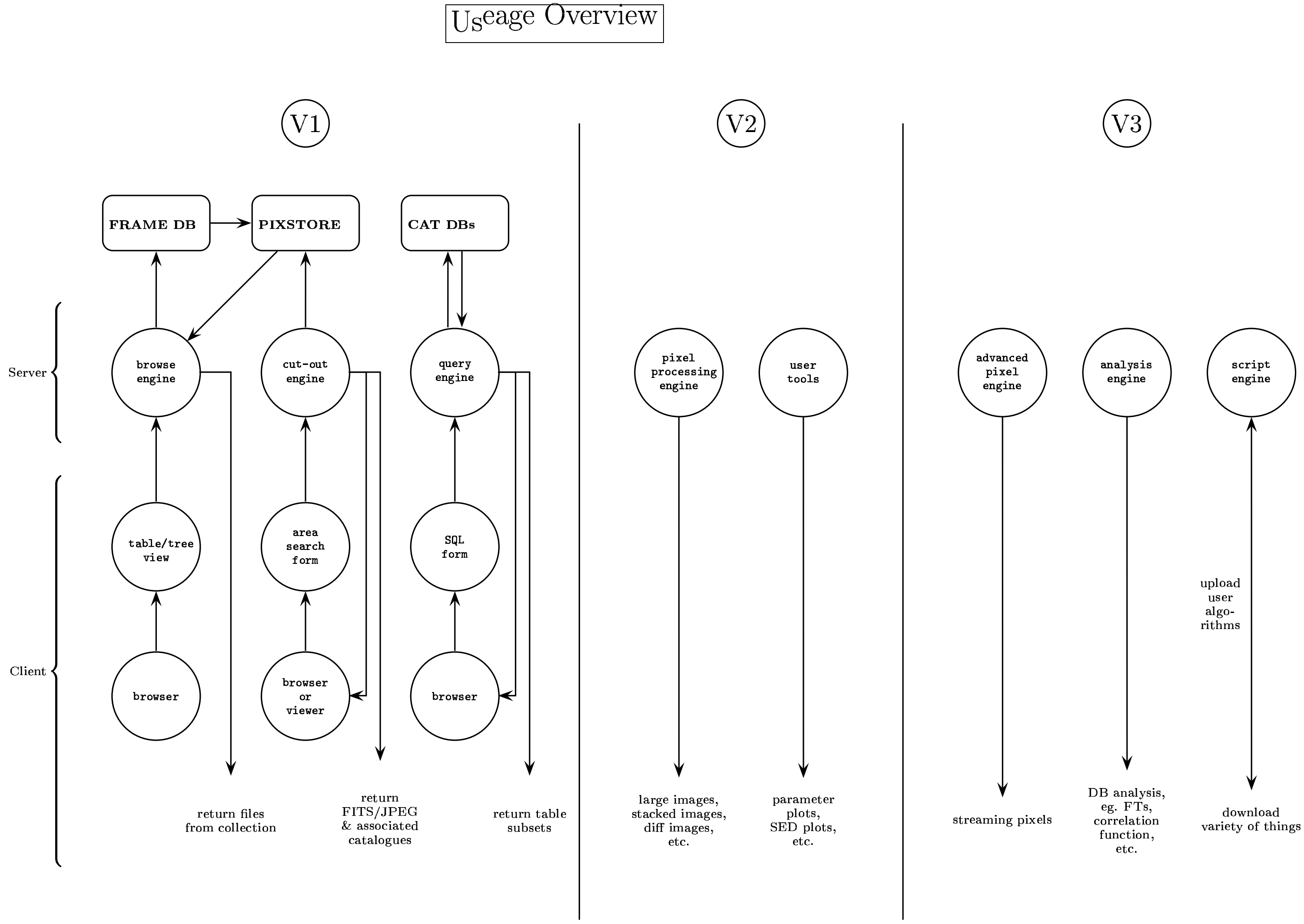

Figure 3: *Illustration of logical design of the WSA user interface*