VISTA DATA FLOW SYSTEM (VDFS)

---

for VISTA & WFCAM data

.
.
.
.
.
.

WSA Data Flow Document

author
Ian Bond (WFAU Edinburgh)
WSA Programmer
number
VDF-WFA-WSA-005
issue
Issue 1.0
date
2 Apr 2003
co-authors
Nigel Hambly

Contents

1 SCOPE

The Data Flow Document analyses the flows of data from input to the WFCAM Science Archive through to the end user. The contents of the archive are identified and described along with the data products and services offered to the end user. The document starts with a contextual overview of archive functionality and data flows. This is followed by an analysis of data flows from input to storage on the archive, and data flows from the archive to the end user. Finally a description of the software involved is given.

This document is driven by the Science Requirements Analysis Document (VDF-WFA-WSA-002) and is symbiotically related to the Database Design Document (VDF-WFA-WSA-007). The User Interface Document (VDF-WFA-WSA-008) describes entry points that allow the user to access the data products and services described in this document. The hardware on which the data flows will take place and on where the software will be deployed is described in the Hardware Design Document (VDF-WFA-WSA-006).

The intended audience is scientists working on the development of the WFCAM Science Archive, but this document would also be of interest to UKIDSS end users who will be accessing the data products.

2 CONTEXTUAL OVERVIEW OF THE ARCHIVE

2.1 Use case view

In this section, a functional overview of the WFCAM Science archive is given by means of a top-level use case analysis. The purpose of such an analysis is to determine system usage, to identify who or what will use and access the system, and to identify system boundaries involved. This type of analysis is important because identifying how users will use the system, drives how developers will design and build it.

The functionality of the WFCAM Science Archive can be depicted using a UML use case diagram as shown in Fig. 1. This can be thought of as a collection of scenarios. External entities that initiate a given scenario are called ”actors”. These can be either persons or software components. While actors can influence what goes on inside the system, the system has no influence over the actions of the actor. Separating out actors in this way is an important part of the carrying out a use case analysis.

In Fig. 1, the WSA is depicted from V1.0 through to its envisaged functionality with AstroGrid in V3.0. The identified actors are explained as follows:

- CASU. Makes WFCAM pipeline produced data available for download to WFAU.
- The WSA archive scientist operates and maintains the data archive. The role of this actor is to get the data from CASU, ingest this into the archive, carry out post ingestion analysis on the data, and to prepare data products for user access.
- The astronomer client refers to those tools that the end user will use to access the WSA data products and services. These could be standard Web browser tools or other tools such as GAIA that have web client functionality.
- AstroGrid client. This actor refers to those tools that will interrogate the AstroGrid registry and then access the WSA web and grid services deployed at WSA—that is, if the services sought by the Grid client can be found at WSA.

2.2 Data flow view

The use case analysis in the previous sub-section is important from the point of view of data flow as it identifies the scenarios and actors involved in the flow of data from input to the archive to the end user. Here, the WFCAM Science Archive is depicted from the point of view of the data flows themselves. A data flow diagram showing a top level overview from this view point is shown in Fig. 2. From a top level point of view, there are essentially two data flows: that from input to the archive and that from archive to the user. These dataflows are expanded upon in the next two sections.

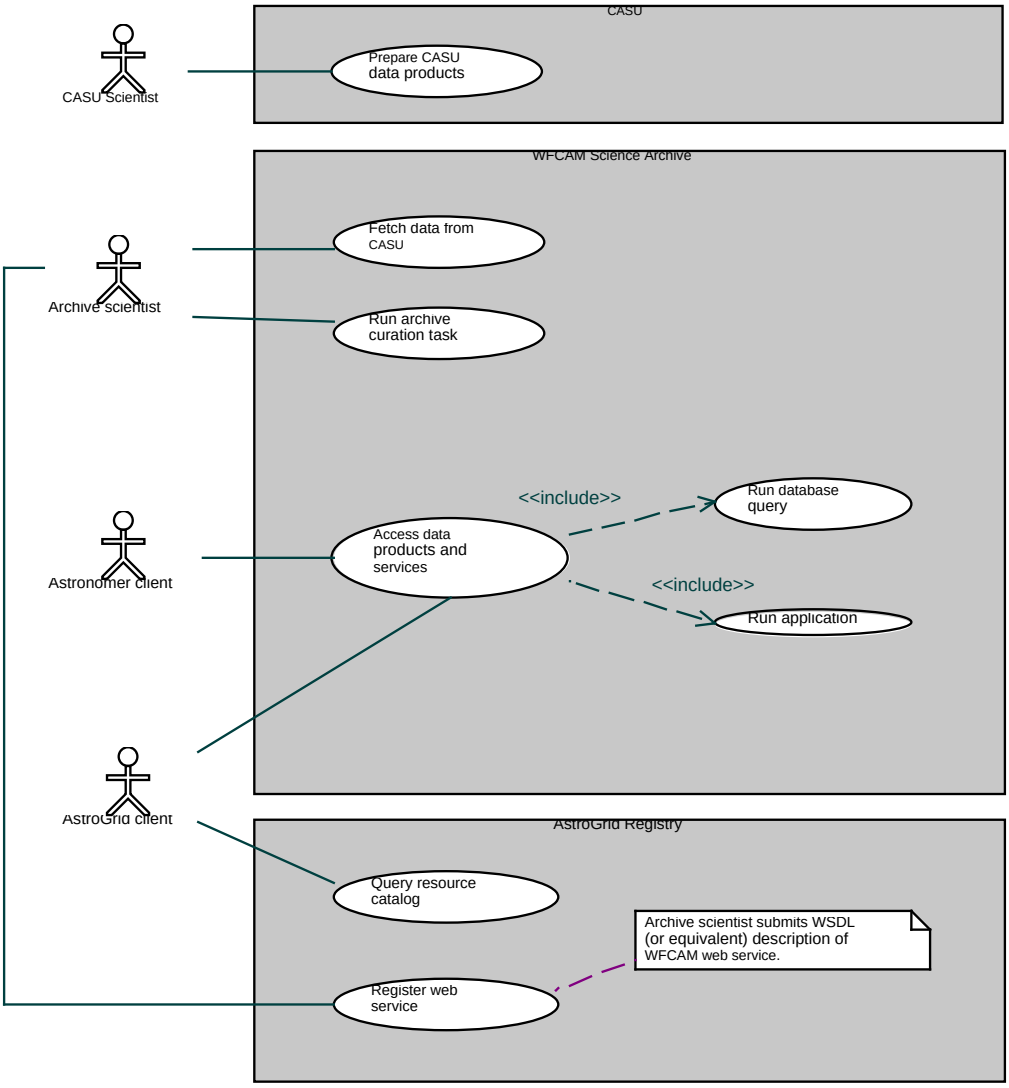

Figure 1: *UML use case diagram depicting the functional relationship of the WFCAM science archive with actors and external systems.*

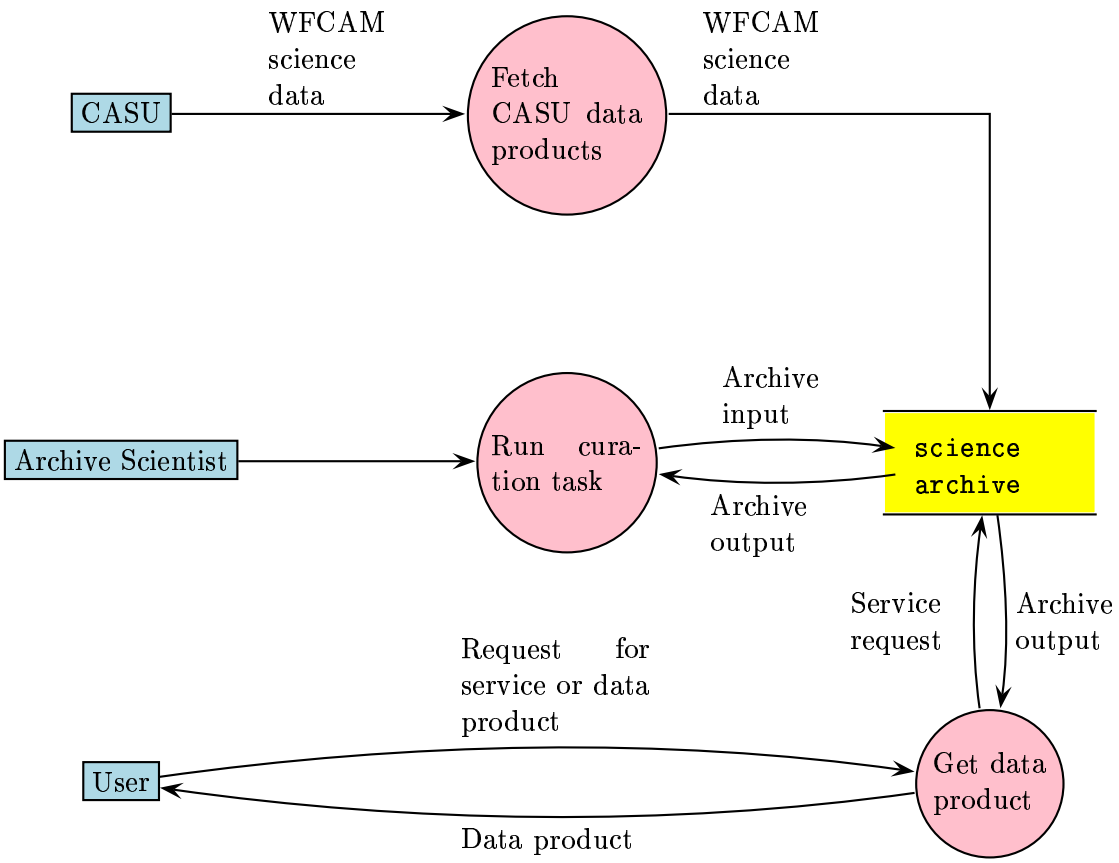

Figure 2: *Top level overview of data flows in and out of the WFCAM Science Archive*

3 CONTENTS OF THE ARCHIVE

3.1 Archived data products for V1.0

Data input from CASU will be archived as a number of ”data products”. Some of these data products involve further analysis and processing at WSA. The Version 1.0 archived data products are listed as follows:

Queryable database

A queryable database containing all source detections, calibration information, image FITS header information, housekeeping data etc. Details of the database design, tables, schema, and management are given in the Database Design Document.

Science Images

A store of WFCAM science images in flat file storage. As specified in the Interface Control Document, these shall be stored as multi-extension FITS files with each HDU corresponding to one interleaved image formed by combining the corresponding microstepped images on one device frame. These files are directly input from CASU. The filenames along with image attributes are tracked in the database.

Compressed Images

A store of compressed science images. Each science image shall have a corresponding JPEG compressed images. The intention here is to allow for rapid internet downloads for visualization in web browsers or users own web clients. They would not be intended for science level data analysis.

Difference Images

A store of ”emission line” images formed by subtracting a K-band image from a corresponding narrow H_2-band image. Observations in the H_2-band will be carried out as a subset of the UKIDSS Galactic Plane Survey. Corresponding difference images will be constructed and archived for all available pairs of K-band and H_2-band images.

Stacked Images

A store of combined images obtained by stacking images corresponding to repeated observations of a given field. There shall be one stacked image corresponding to each device frame field-of-view for each of the UKIDSS Deep Extragalactic Survey (DXS) and Ultra Deep Survey (UDS) in each available passband.

3.2 Planned archived products for V2.0

The V2.0 WFCAM Science Archive will additionally contain:

- Externally provided catalogues and pixel data (UKIDSS complementary imaging and SDSS data releases as available at that time).
- CASU data products based on WFCAM Open Time observations.
- Enhanced database containing information based on post ingestion analysis of UKIDSS source catalogues. New data to be ingested into the database includes proper motions, dereddened colours, flux measurements derived from placing apertures on SDSS pixels at WFCAM detection positions.

3.3 Dataflows

The top level use cases identified in Section 2.1, carried out by the WFAU archive scientist, have been expanded into 20 ”curation” use cases. These are analysed in detail in the Database Design Document but for reference, they are listed here in Table 1. In this document, they are studied in the context of data flows from input to storage in the archive. Each use case is depicted as a process bubble in the data flow diagrams that follow in Figs. 3–7.

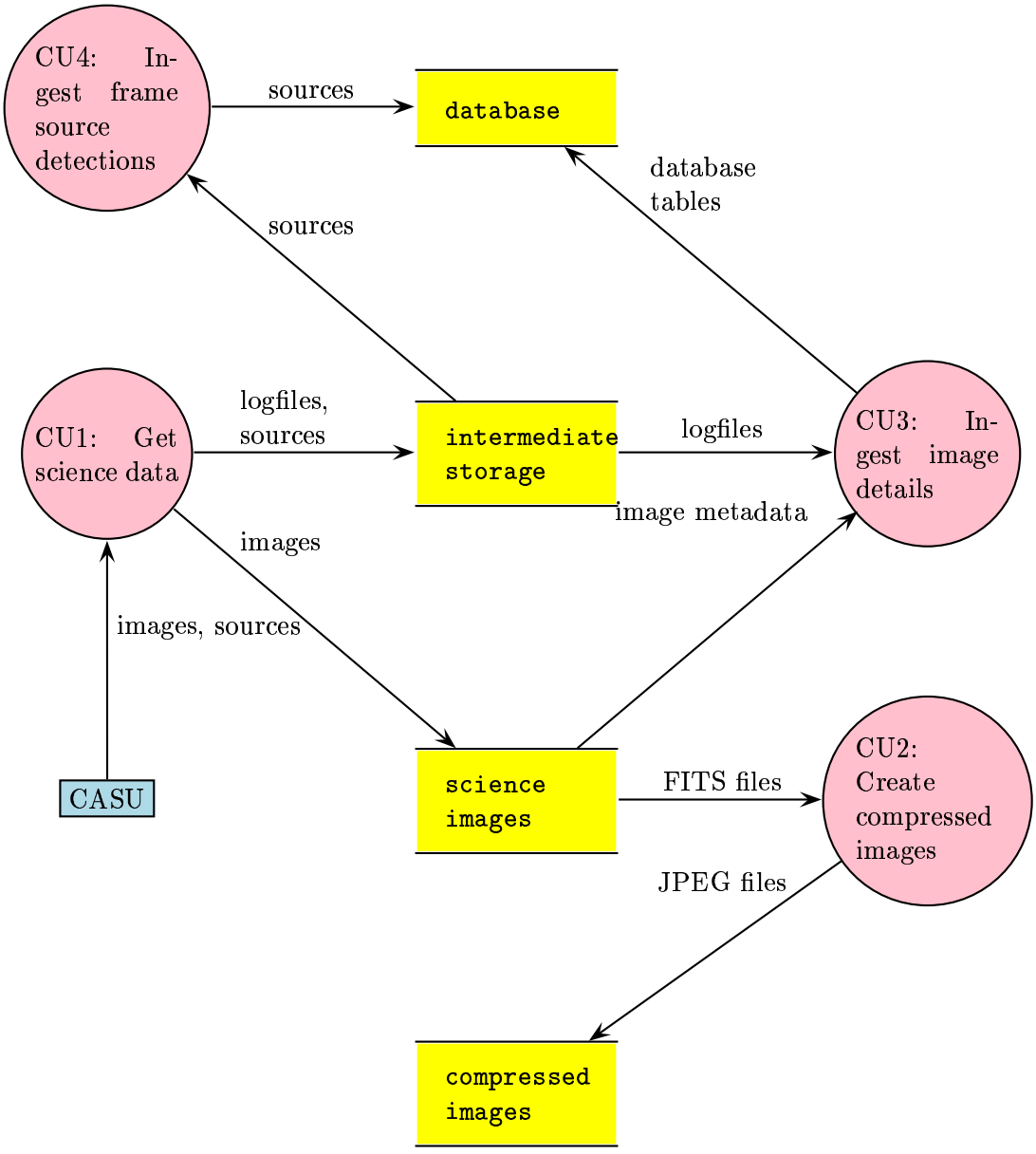

Figure 3: *Data flows involved in the data input and ingestion stages*

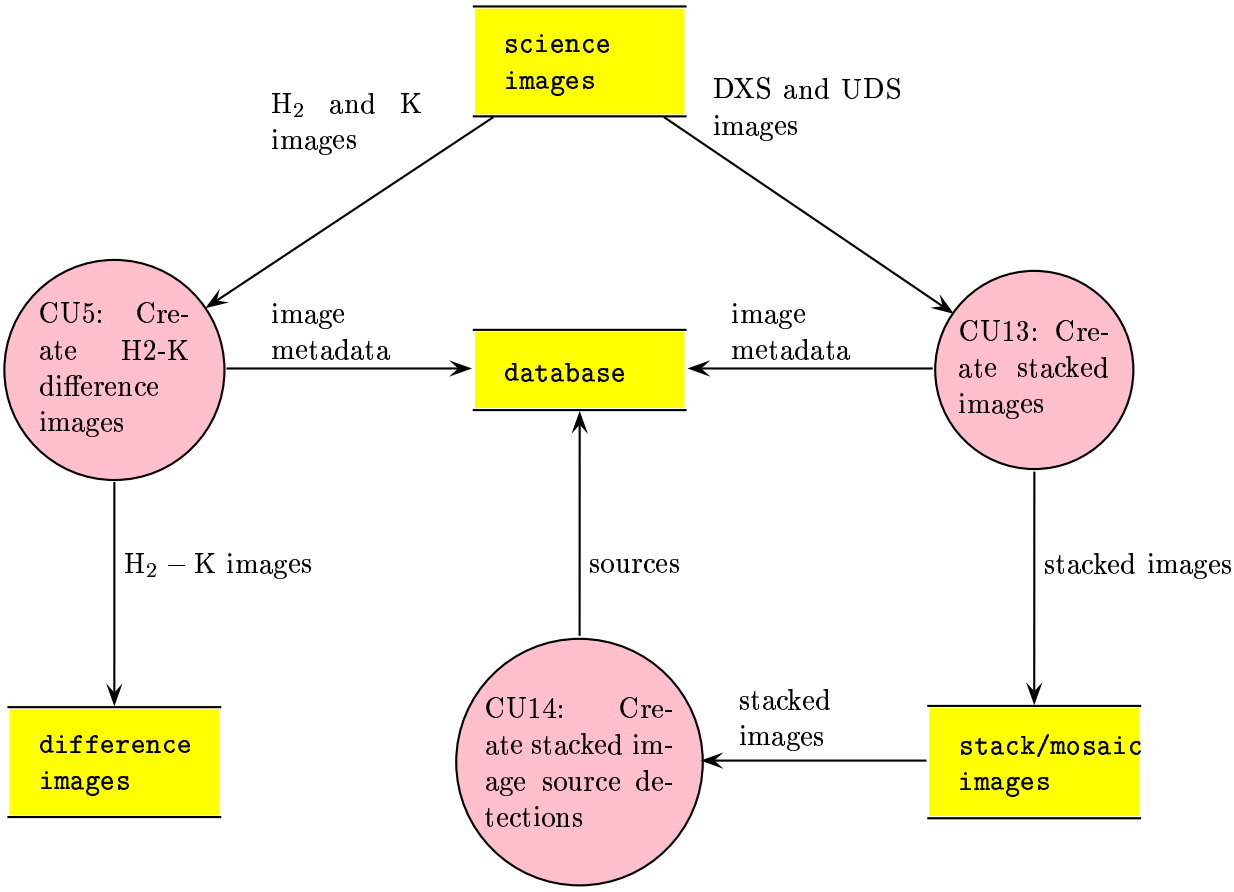

Figure 4: *Data flows involved in generating post-ingestion archived data products*

CU1	Obtain science data from CASU.
CU2	Create library compressed image frame products
CU3	Ingest details of transfered images and library compressed image frame products
CU4	Ingest single frame source detections
CU5	Create library H_2–K difference image frame products
CU6	Create spatial indicies for all new records having celestial coordinates
CU7	Recalibrate photometry
CU8	Create/update merged source catalogues
CU9	Produce list measurements between WFCAM passbands
CU10	Compute/update proper motions
CU11	Recalibrate photometry
CU12	Get publicly released and/or consortium supplied external catalogues
CU13	Create library stacked/mosaic images
CU14	Create standard source detection list from any new stacked/mosaiced image frame product
CU15	Run periodic curation tasks CU6–CU9
CU16	Create default joins with external catalogues
CU17	Produce list driven measurements between WFCAM and non-WFCAM imaging data
CU18	Create/recreate table indicies
CU19	Verify, freeze, and backup
CU20	Release—place online new DB product

Table 1: *Curation uses cases for the WFCAM Science Archive*

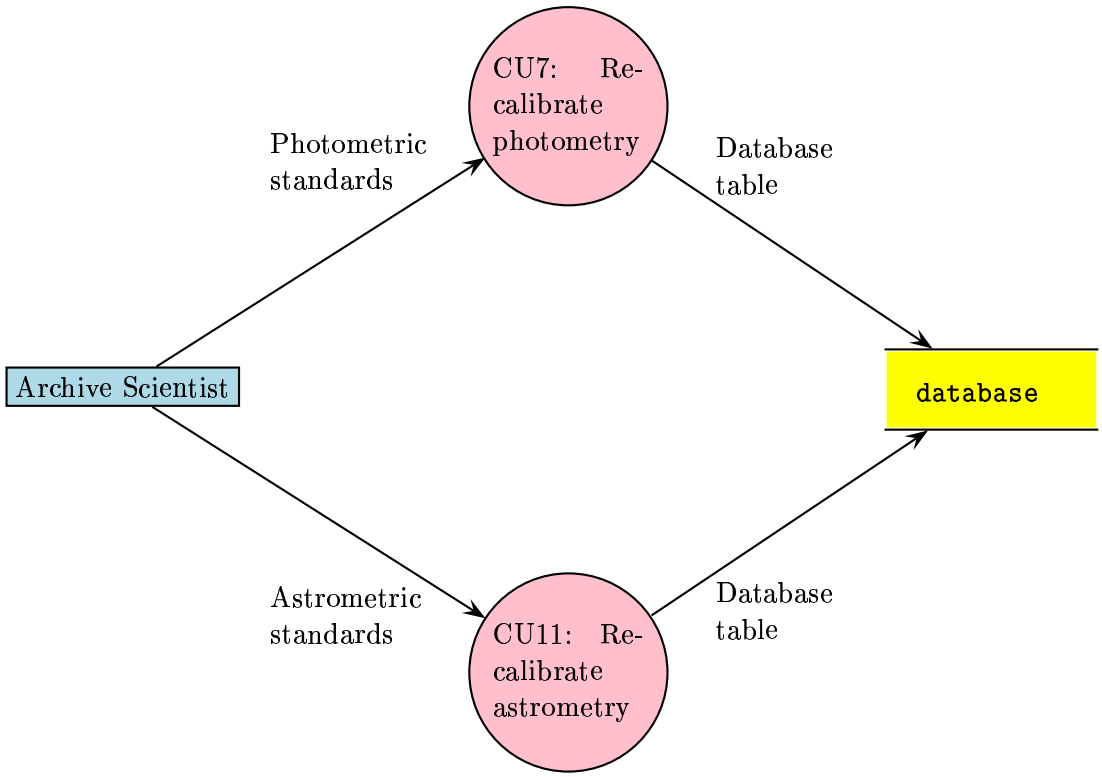

Figure 5: *Data flows involved in photometric and astrometric recalibration*

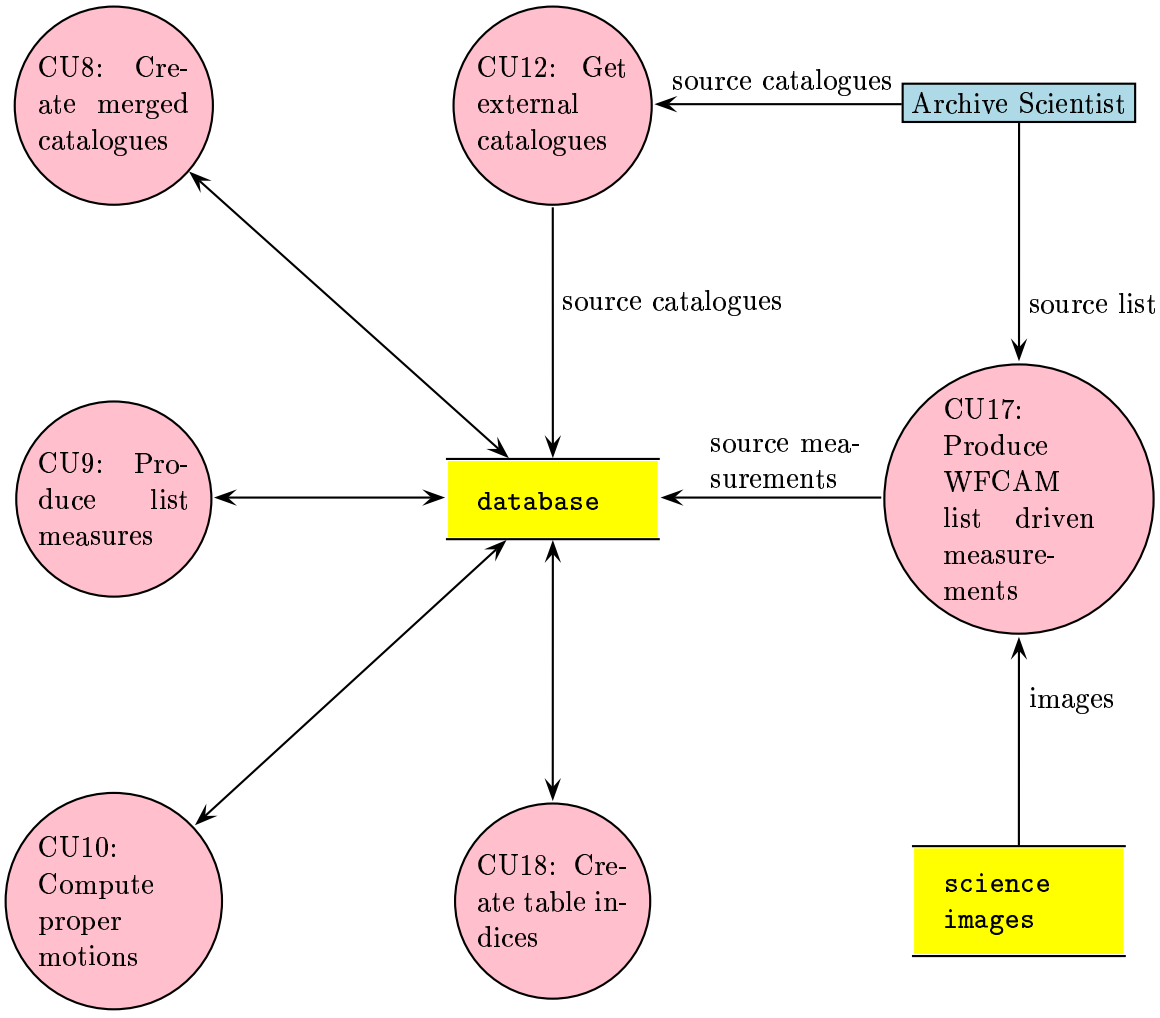

Figure 6: *Dataflows involving access and manipulation of object catalogues*

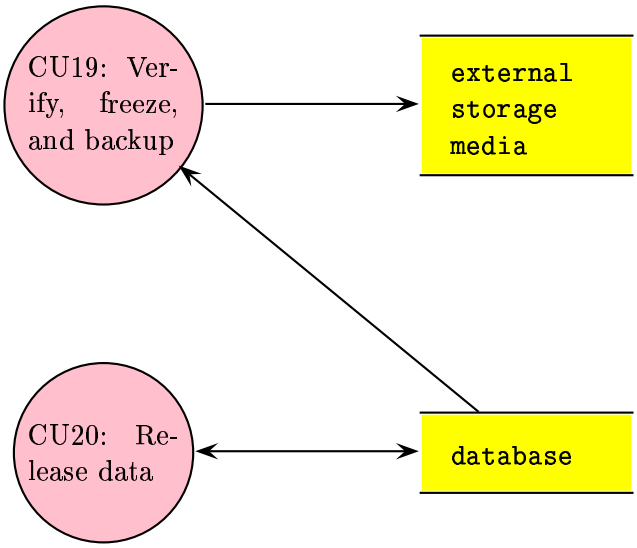

Figure 7: *Database backup and release dataflows*

4 DATA PRODUCTS AND SERVICES

4.1 Definition of data products and services for V1.0

Database Query

The user shall be offered the ability to access information stored in the database by forming and submitting queries based on any stored quantity. It is envisaged that the most common queries will be those that request photometry on detected sources. The reader is referred to the Database Design Document for further details on the contents of the database.

Extracted Image

The science, stacked, and difference images will be generated by the archive scientist and stored on a device frame-by-frame basis. The user will be able to extract sub-frames of these images on a user specified region of the sky limited in extent to that covered by a complete device frame. The pixel data will be served without any pixel resampling and further processing.

This data product will be served as a multi-extension FITS file and shall comprise the following components:

- Image metadata (type of image, provenance information, etc) encoded in the primary FITS header.
- World Coordinate System information encoded in the primary header.
- Image pixel data placed in the primary pixel data array
- *Optional.* Corresponding confidence maps, attached as pixel data in a FITS extension.
- *Optional.* List of sources and their measured fluxes detected in corresponding piece of sky. These data will be attached as a binary extension.

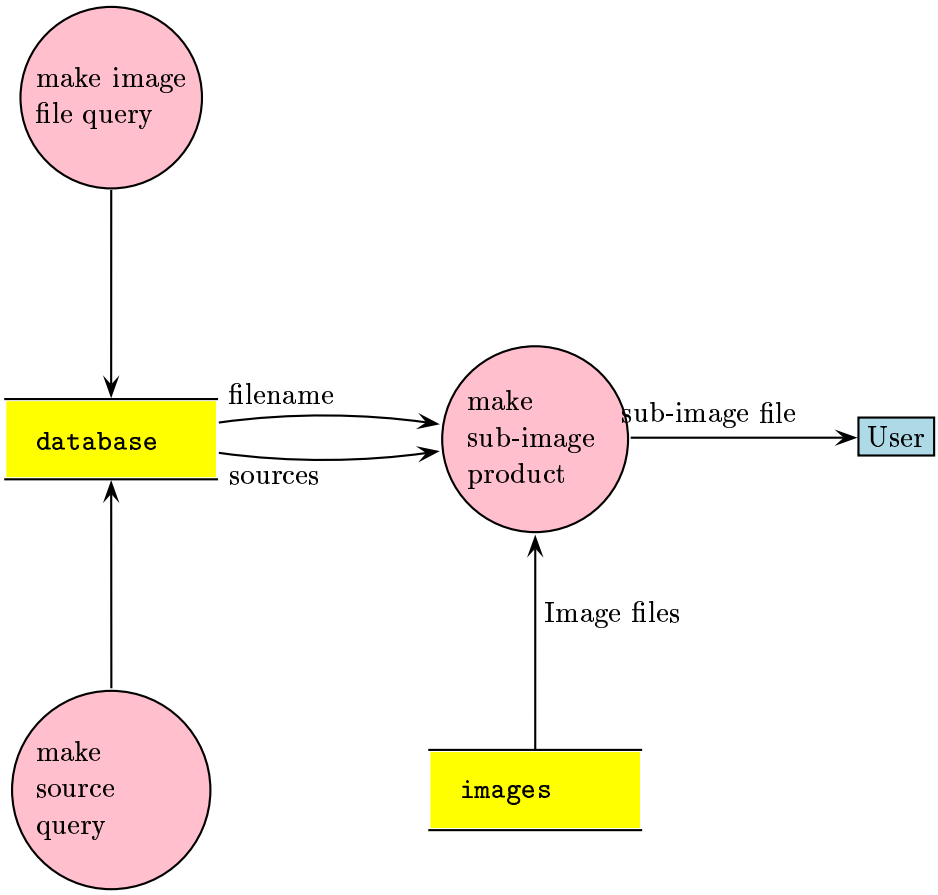

Figure 8: *Image extraction data flow*

A dataflow diagram showing movement of data from the archive to the user in generating an extracted image is shown in Fig. 8. The user specifies a region of the sky and the type of image (science, stack, or difference). The subsequent dataflow involves three processes:

1. Formation and submission of a query to the database in order to find the image filename that contains the user supplied piece of sky.
2. Formation and submission of a query to extract information on detected sources within the specified region of the sky.
3. Running application software that reads a multi-extension FITS file and returns a sub-image corresponding to the specified region of the sky. A corresponding confidence map and/or list of sources will also be attached if required.

Finder Chart

Image derived from a single observation image with object ellipses overlaid.

Mosaic Image

Image derived from combining all available individual science images within a user specified region of the sky. The intention is to build large panoramic images but in V1.0 the size of the mosaic image shall be limited to $0.8^\circ \times 0.8^\circ$. The final image may be pixel re-binned (blocked down) if required. The corresponding dataflows, shown in Fig. 9, are very similar to those involved in generating the extracted image. The main difference is that the appropriate application software for constructing image mosaics will be used. As with the extracted images confidence maps and/or source detection information will be added as FITS extensions.

4.2 Planned products and services for V2.0

Version 2 of the WFCAM Science Archive will allow more user interaction with the data and provide more functionality for data mining operations. The aim here is that the user will be able to carry out a range af analysis tasks on data stored in the archive without having to download huge amounts of data to their local machine. As a step towards AstroGrid integration, the server-side tools will be recast as Web and Grid services.

Data products and services planned for Version 2 are:

Enhanced database query and exploration

The user will be able to carry out queries with the database jointly with an external user-supplied catalogue. Data visualisation and exploration tools will be added to provide the user with the ability to remotely interact with the archive. These tools will allow XY plotting, histogram plotting, and simple model fitting routines.

User controlled image products

Rather than extract image products from the archive image stores, the user will be provided with the ability to generate these products on the fly. Points specific to individual image products are:

- The user will be provided with the ability to select those images to be stacked and will be provided with a range of choices of algorithms for combining the images into the stack.
- The user will have the ability to genrate and download mosaiced images of arbitray (as long as practicable for WSA) size with pixel rebinning. Multi-colour functionality will be provided to allow the user, for example, to generate JHK colour images. As with the stacked images, a range of image combining algorithms will be provided.

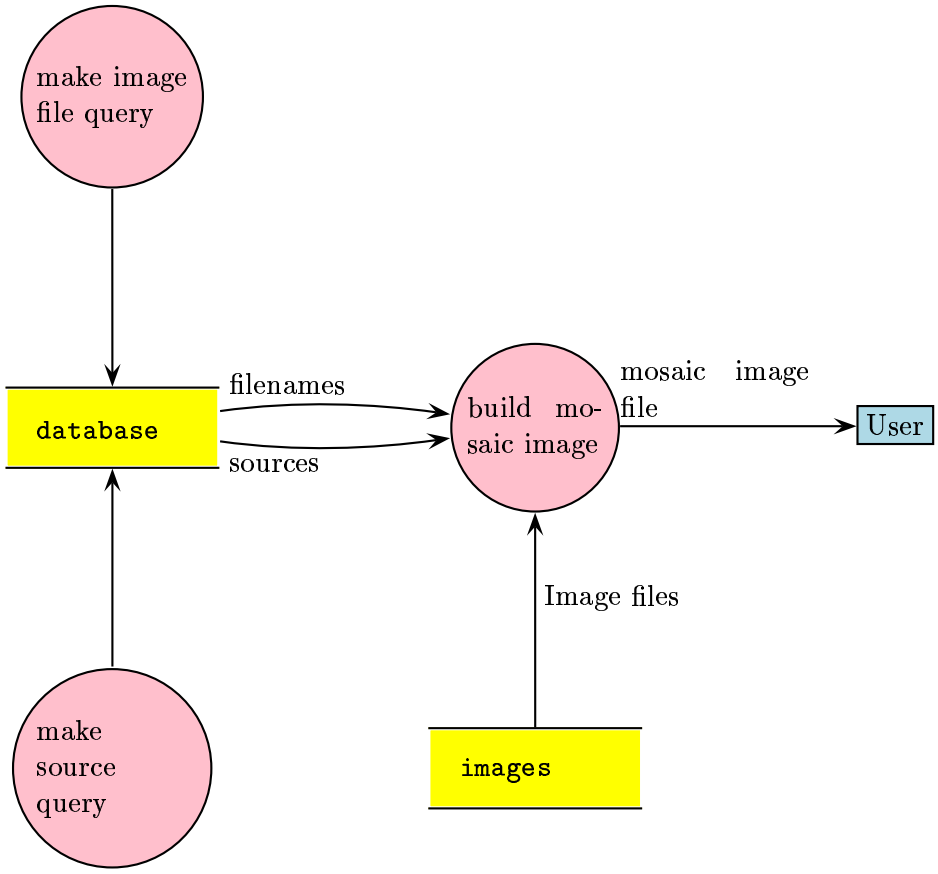

Figure 9: *Mosaic image construction data flow*

- As well as the default H_2–K difference images, the user will be offered a choice of cross passband difference imaging combinations.

User controlled image source extraction and analysis

The user will be provided greater functionality in extracting and/or measuring sources on the various types of images. A choice of tools for source extract/measurement wools will be provided including SExtractor, DAOphot, and CASU standard source extraction. The user will be able to apply these tools to the on-the-fly image products described above, as well as archived images. Two types of source analysis will be possible:

1. Extraction of sources on a given set of images
2. Photometry measurements on a given set of images based on an input list of source positions. The input list can either be an external list of sources supplied by the user, the result of a query on the database, or even the result of operation (1) above.

5 DESCRIPTION OF SOFTWARE

5.1 Decisions taken for V1.0

The generation of the image data products requires a combination of queries to the database and the execution of various software applications. The software will deployed on one or both of the mass storage system server and the web server. These are both PCs that run under Linux as described in the Hardware Design Document. The software executables will be wrapped in mini-pipelines using Perl as a glue language. These mini-pipelines will be tied in to the user interfaces that provide the entry points for users to access the products and services. These interfaces and the underlying worklows are described in the User Interface Document.

FITS file manipulation code

The availability of the very powerful CFITSIO library makes it a straightforward job to write C/C++ code for low level access to data encoded in FITS files and to carry out manipulations on the various components. For operations involving World Coordinate System manipulations, the Starlink AST library (Warren-Smith & Berry, 2002) will be used. For Version 1.0, code will be written to carry out the following.

- Extract a sub-image from a multi-extension FITS file containing WCS information. The application will take the sub-image centre and the dimensions in pixels as input and will find the HDU containing the relevant pixel data. The output will be a FITS file containing the sub-image data. The WCS information will be preserved, but with the appropriate changes to the CRPIX1 and CRPIX2 keywords. Optionally, corresponding pixel data will be extracted from an input confidence map FITS file and attached to the output file as a FITS extension.
- Attach an input list of sources and their properties as a binary table extension to an input FITS file
- Strip all FITS keyword-value-comment triplets from an input multi-extension FITS file.

Stacking and mosaicing tools

These are being developed at CASU and will be implemented in the WFCAM science archive for generating archived stacked images and user demanded mosaic images.

Source extraction and measurement

The CASU-developed standard source extraction/measurement tools will be implemented at WSA for generating merged catalogues.

Difference Imaging

We will implement the difference imaging software that has been used in the MOA microlensing survey (Bond et al 2001). Adaptive kernel matching is used to take into account differences in seeing between the two input images. Also allowance is made for a spatially varying kernel. This software has been implemented in the H_α–short red difference images offered by the WFAU SuperCOSMOS data facility.

5.2 Plans for V2.0

Advanced image stacking software for V2.0 functionality will be developed as a joint WFAU-CASU effort. Prototype software has been developed at WFAU that will generate mosaic images of size limited only be the amount of available disc space. Additional source extraction and measurement software to be added at V2.0 include SExtractor and DAOphot.

6 REFERENCES

Bond, I.A. et al, 2001. *Real-time difference imaging analysis of MOA Galactic Bulge observations during 2000*. MNRAS, 327, 868.

Warren-Smith, R.F. & Berry, D.S., 2002. *Ast: A Library for Handling World Coordinate Systems in Astronomy*, Starlink User Note 211.11

7 ACRONYMS & ABBREVIATIONS

ADnn : Applicable Document No nn
CASU : Cambridge Astronomical Survey Unit
FITS : Flexible Image Transport System
HDU : Header Data Unit in FITS files
UML : Unified Modeling Language
VISTA: Visible and Infrared Survey Telescope for Astronomy
VPO : VISTA Project Office
WCS : World Coordinate System
WFAU : Wide Field Astronomy Unit (Edinburgh)

8 APPLICABLE DOCUMENTS

AD01	WSA Science Requirements Analysis Document	VDF-WFA-WSA-002 Issue: 1.0 2/4/03
AD02	WSA Hardware Design Document	VDF-WFA-WSA-006 Issue: 1.0 2/4/03
AD03	WSA Database Design Document	VDF-WFA-WSA-007 Issue: 1.0 4/4/03

9 CHANGE RECORD

Issue	**Date**	**Section(s) Affected**	**Description of Change/Change Request Reference/Remarks**
Issue 1.0	2/4/03	All	New document

10 NOTIFICATION LIST

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

WFAU: P Williams, N Hambly
CASU: M Irwin, J Lewis
QMUL: J Emerson
ATC: M. Stewart
JAC: A. Adamson

__oOo__