WFCAM SCIENCE ARCHIVE

SCIENCE REQUIREMENTS ANALYSIS DOCUMENT

Nigel Hambly, Ian Bond & Bob Mann

(with contributions from the UKIDSS Consortium, Andy Adamson & Jim Emerson)

Wide Field Astronomy Unit (WFAU), Institute for Astronomy, University of Edinburgh

Modification history:

Version	Date	Comments
1.00	28/10/02	Original version (NCH & IAB)
1.10	04/12/02	Updated following input from UKIDSS, AA & JPE (NCH)
1.20	08/01/03	Updated following CASU & WFAU review (NCH)
1.30	20/03/03	Inserted 'analysis' in title

Contents

1 INTRODUCTION

Standard, top–level analysis for complex digital systems consists of:

1. definition of requirements and specifications,
2. undertaking analysis and design,
3. code development and debugging,
4. unit and integration testing,
5. deployment and maintenance;

this sequence usually being iterative as scope/specifications change and feed back to modify the system requirements.

WFCAM (see Appendix A.1) is a new camera for the 3.8m United Kingdom Infrared Telescope. This large format camera will have an unprecedented data rate. Ultimately, successful science exploitation of WFCAM will depend on user access to the large data volumes generated by this instrument. Data volumes are far in excess of those that users can expect to hold and process on their own facilities. This leads to the concept of pipeline processing and the establishment of a centralised 'science archive'. The project to develop the WFCAM Science Archive (hereafter WSA) is outlined in Appendix A.2 and references therein.

This 'science requirements document' (SRD) details the basic requirements for the WSA, and represents item 1 in the above sequence. The intention is to state the top–level science requirements being placed on the WFCAM Science Archive (WSA) as a whole; give science usage examples of the WSA; and finally to discuss in more detail those requirements pertaining to the WSA in order to produce a specification for its design. The approach taken in this document is to distil the external, top–level requirements and the usage examples, through analysis and implication, to an explicit statement of the WSA contents and functionality. Hence, the SRD is structured as follows:

- Section 2 references the external requirements;
- Section 3 restates the 'top–level', 'contents & functions', and 'security requirements', along with analysis and notes and further refines these above, along with the externally specified 'detailed requirements', to provide a basic list of the detailed requirements;
- finally, Section 4 summarises the WSA contents and functionality in a more concise and explicit way for the perusal of interested parties and to enable WSA developers to progress the design.

Subsequently, we intend to follow the sequence above and undertake a design for the WSA including documentation of Data Products, Data Flow, Hardware Architecture and Software Architecture. The detailed specification for the WSA will be developed further in these following documents rather than in the SRD.

It is not intended that the requirements and usage examples are set in stone at this stage. Both are 'living' documents in the sense that they are online on the web (URLs are given in the next Section) and are subject to small alterations as the WFCAM and UKIDSS projects progress. The intention of this document is to take these inputs as they are at the time of writing (Q4 2002) and analyse them in order to have something to work to in the WSA development project.

The WFAU WSA development homepage is at http://www.roe.ac.uk/~nch/wfcam.

2 REQUIREMENTS AND USAGE EXAMPLES

The top–level requirements and usage examples are reproduced in Appendices A.3 and A.4 where more details can be found. These are the external, top–level requirements placed on the science archive by the 'customers', viz. JAC and UKIDSS.

3 REQUIREMENTS ANALYSIS

In the following analysis, we discuss the top–level requirements referenced in the previous Section in more detail. Each item has an associated *Rationale*, *Implications* which discuss the implications for the WSA design, an optional *Note*, and finally a concise statement of the requirement to be developed in later Sections. It is intended that the requirements cannot be changed without consultation (primarily with JAC and UKIDSS).

3.1 Top–level requirements

T1:
Science archive shall provide the maximum possible potential for capitalizing on the UKIDSS surveys.
Rationale: UKIDSS will absorb the greater fraction (75%) of *all* WFCAM time on UKIRT and so is the top priority for WSA usage.
Implications: The UKIDSS programme must be the prime science driver for the WSA. Archive development needs to be an open process, with as much UKIDSS involvement as possible. Hence, full and up-to-date documentation needs to be available in web–browsable form as well as hardcopy. The tight schedule for WFCAM, the competition from CFHT's WIRCAM, and the need for timely release of data for competitive and high–impact science place a correspondingly tight schedule on delivery of the WSA. Resource/time constraints imply a phased approach to WSA development, with a committment to producing a basic working archive system by instrument first light, followed by development to a fully functioning archive system thereafter. To expedite delivery of the WSA, design should be based on existing archive solutions and code where appropriate.
Note: WFCAM is currently due for delivery by Q4 2003; UKIDSS survey operations will likely begin in earnest in Q1 2004.
Requirement:
A basic working science archive (hereafter 'Version 1.0') *must* be in place at Q4 2003. A fully functioning archive system (hereafter 'Version 2.0'), as defined by the requirements herein, must be available as soon as possible after WFCAM first light, and no later than 1 year after survey operations begin in earnest.

T2:
Science Archive must contain and serve *pipeline processed* data (processed pixels, object catalogues and housekeeping data) from both UKIDSS and other usage (e.g. open time, commissioning time).
Rationale: Even small PATT programmes (for example) may produce large amounts of data that are problematic for the user's home institute resources. Moreover, non–survey data will be a valuable datamining resource (see later).
Implications: WSA data accumulation must take into account non–survey usage. Database schema design must be flexible to allow for non–survey data. Proprietary rights need to be protectable in the WSA.
Note: Pipeline processing and subsequent archiving cannot be undertaken for frames taken in non–standard observing modes. For non–survey data that *are* taken in standard modes, limited standardised schemas will be set up and the data will be archived; it will not be possible to develop individual shcemas on a case–by–case basis.
Requirement:
Science Archive (all Versions) must contain and serve *pipeline processed* data (pixels, object catalogues and housekeeping data) from both UKIDSS and other usage (e.g. open time, commissioning time).

T3:
Science Archive must be flexible to cope with alterations to UKIDSS survey design over time.
Rationale: The UKIDSS observing allocation and programme are subject to change by the Board on a 2 yearly rolling review.
Implications: WSA design must not preclude changes in design of the major surveys. Again, database design must be sufficiently modular and flexible to cope with this.
Note: Following the initial Board review in May 2002, twice–yearly reviews are expected in mid–2004 and every two years thereafter.
Requirement:
Science Archive (all Versions) will match UKIDSS survey requirements as they are currently specified, but will be flexible enough to follow changes in survey design.

T4:
Science Archive design must facilitate usage from 'Grid clients' and inclusion in the Virtual Observatory (VO).
Rationale: Given the legacy aspect of the UKIDSS surveys (especially the LAS and GPS) it is expected that the WSA will form a substantial element in the 'datagrid' of the VO (indeed, WFCAM is a prime science driver in the UK's AstroGrid project).
Implications: WSA access tools, data product formats and transfer protocols must conform to internationally agreed VO standards.
Note: The AstroGrid Phase A report is now available (October 2002) for information concerning VO development prototypes.
Requirement:
Version 1.0 Science Archive will conform to *existing* standards and will be designed such that new standards can be easily incorporated, but must not be delayed by waiting for new developments to crystalize. Ultimately, the Science Archive must conform to internationally agreed VO standards in access tools, data product formats and transfer protocols.

T5:
Science Archive must allow, for example, *simple* and *complex* queries, with appropriate interfaces.
Rationale: Many users will query the WSA, from the Grid–client 'power user' to the casual, non–expert interactively browsing astronomer. Both are important from the science exploitation point of view.
Implications: Different levels of user interface will be needed for the WSA, from interactive web forms through remote–client GUIs to Grid–enabled clients.
Requirement:
Version 1.0 Science Archive will allow *simple* (see later) queries. Version 2.0 Science Archive will allow usages at varying levels of complexity (as defined later).

T6:
Science Archive must be simple to use for PR purposes.
Rationale: UKIDSS is the next development in the UK's Wide Field programme. High profile science will emerge from UKIDSS, and as the first point of contact with the data, the WSA must be designed appropriately.
Implications: Again, the WSA must be user–friendly to the casual, browsing user. 'Aesthetic' data products (e.g. pseudo–colour images) must be available, in addition to 'serious' science products.
Note: The SDSS has good examples of entry points for PR purposes (URL) as well as scientist access points (URL). However, while the production of individual images as a requirement of the WSA, the responsibility of designing and maintaining a 'gallery' website of publicity images lies elsewhere (eg. with JAC and/or UKIDSS).
Requirement:
Science Archive (all Versions) must have interfaces that are open to simple, intuitive use by the non–expert.

T7:
Science Archive must allow access to survey data before all observations are complete, and must not be disrupted by regular ingest of new survey data.
Rationale: Rapid exploitation requires immediate access. The full UKIDSS programme will take up to 6 years or more, and users will want to undertake preliminary analysis after months of data accumulation rather than wait until the full survey datasets are released.
Implications: WSA design must allow for constant data ingest and regular data releases (e.g. interim survey products). WSA must allow for updates to calibrated quantities. WSA must allow for archiving of catalogues from 'reruns' of the processing pipeline, as well as catalogues from previous runs, over pixel datasets in the event of bug fixes and/or enhancements of processing algorithms.
Note: The approach taken with the WFAU's SSS database is to locally mirror the entire released dataset so that two versions are held: a static online version, and another online (but inaccessible from the outside) version for updates. At a release point, the update version becomes the network online version, is copied back to mirror the latest updates, and the whole procedure is so cycled.
Requirement:
Version 1.0 Science Archive must be operable in time for WFCAM first light. Interim survey products must be released to the community on timescales determined by WFCAM observing periods (i.e. a survey 'release' will occur as soon as possible after each observing period, and before the end of the following period).

T8:
Science Archive must allow requests for arithmetic operations, and options from an advanced processing toolkit, on pixel data.
Rationale: Pixel data volumes will be too large for efficient transfer to users home institute for manipulation.
Implications: WSA needs sufficient online storage for pixel data, and sufficient CPU, temporary storage and appropriate software toolkits for pixel manipulation.
Note: Astronomy community in general, and CASU, Subaru for example, are developing pixel processing algorithms. Not all routines will need coding from scratch.
Requirement: Version 2.0 Science Archive must allow requests for arithmetic operations, and options from an advanced processing toolkit (see later), on pixel data. (no requirement on the Version 1.0 Science Archive to allow this advanced functionality, since we do not anticipate any demand for this immediately after first light).

T9:
Science Archive must be scalable to VISTA data volumes.
Rationale: The WFCAM and VISTA cameras (and science programmes being pursued with them) are similar enough that it makes sense to produce a scalable solution from WFCAM to VISTA for cost effectiveness.
Implications: WSA developments must be open to scrutiny by, and must receive input from, the VISTA project.
Note: VISTA first light is currently scheduled for Q4 2006.
Requirement:
Despite the need to expedite delivery of the WSA, development will be made *at all times* with due regard to scalability to VISTA data volumes.

T10:
Science Archive must be able to merge reduced frames taken in non–photometric conditions with other data from the same survey.
Rationale: Rapid progress may require acceptance of sub–optimal observations in lieu of better, later repeated observations.
Implications: WSA must be able to cope with sub–optimal data and their subsequent displacement by better, repeat observations.
Requirement:
Science Archive (all Versions) must be able to cope with sub–optimal survey observations, and their subsequent displacement by better, repeated observations.

T11:
Science Archive must have some capability for the remote user to carry out data exploration and interaction in real time.
Rationale: The UKIDSS programme contains many instances (e.g. see the specific usage examples) where the remote user will want to manipulate and visualise large amounts of data quickly (i.e. without transfering the large dataset to their own machine).
Implications: Remote client GUI tools will need to be developed for the WSA to enable such interactive data exploration and manipulation. 'Real time' interaction has implications for WSA response time when trawling Tbyte–sized datasets. Clearly, $\sim$ 1000sec response time is unacceptable for interactive use, while $\sim$ 10sec response time is unrealistic given current technological and financial constraints (such a fast response time may be feasible with a very high degree of parallelism, with consequent complexity and cost implications). For these purposes, a figure of $\sim$ 100sec response time seems reasonable.
Note: Of course, for queries on *indexed* quantities (position, image class, brightness and other commonly used attributes), WSA response time will be fast but ultimately limited by factors beyond the control of WFAU (eg. user network connectivity).
Requirement:
Version 2.0 Science Archive must have some capability for the remote user to carry out data exploration and interaction in real time, where 'real time' is understood to mean a timescale of $\sim$ 100sec for wholesale trawls. No requirement on Version 1.0 Science Archive system to provide this speed; the ultimate goal should be a response time of $\sim$ 10sec.

3.2 Science archive contents and functions (minimum)

C1:
Contains calibrated object catalogues resulting from the pipeline, for both UKIDSS and open–time observations
Rationale: These are obvious, basic science archive functions.
Implications: Database schemas must be set up for various tables of object catalogues. Catalogue ingest software and procedures will be required. Software will be required for 'post–processing' type operations, for example, merging routines and recalibration routines.
Requirement:
Science Archive (all Versions) must contain calibrated object catalogues resulting from the pipeline, for both UKIDSS and open–time observations

C2:
Ingests and stores pipeline output frames for later online processing, generates compressed pixel images on the fly for rapid web–based access, carries out immediate cross–referencing with existing UKIDSS survey data and produces consolidated UKIDSS catalogue in a given field
Rationale: Again, basic science archive functionality.
Implicattions: Database schemas must be designed to track between object catalogue tables and pixel data files. Pixel manipulation software will be required.
Requirement:
Science Archive (all Versions) must ingest and store pipeline output frames, allow rapid web–based access to images, and produce merged UKIDSS catalogues in a given field.

C3:
Is able to recalibrate a given field or fields in the event of revised calibration information (specifically, photometric and astrometric), and allow database queries on the recalibrated quantities
Rationale: Changes in calibration information are frequently encountered in survey operations, and the science archive itself may lead to such changes.
Implications: Database schema must allow provision for recalibration – e.g. stores positions as pixel co-ordinates plus and astrometric solution (consisting of specified model and coefficients); stores photometry as flux measures plus calibration data. Calibrated quantities will also be required to be stored in tables, since inverting calibration models to translate queries in calibrated units to uncalibrated ones will be difficult in general. The archive must be able to replace calibrated quantities when new ones become available. Calibration version control within the archive is required.
Requirement:
Science Archive must be designed from the start to enable astrometric and photometric recalibration.

C4:
Is able to cross-calibrate photometric information using areas of overlap between processed frames, where available.
Rationale: This is not a sensible function of the pipeline, which is required only to produce results on a night–by–night basis. The science archive will have all photometric information and calibrations for all superframes, and is where this should happen.
Implications: Calibration tools will be required to homogenise photometry over surveyed areas using overlap information and photometric zeropoints.
Requirement:
Version 2.0 Science Archive must be able to cross-calibrate using areas of overlap between processed frames, where available (no requirement on Version 1.0 Science Archive to cross–calibrate).

C5:
Allows public access to subsets of survey data on a variety of different search criteria (specified below)
Rationale: Basic science archive functionality.
Implications: For versatility, SQL–like querying is required, even if this is transparent to the user (e.g. simple access via web–form interface).
Requirement:
Science Archive (all Versions) must be designed to allow public access to subsets of survey data on a variety of different search criteria (specified later).

C6:
Allows rapid on–line cross–referencing of search results with other catalogues.
Rationale: consistent with T1, this requirement is expanded on later.
Implications: The Science Archive must undertake to store commonly used catalogues locally for combination queries in a queryable database.
Requirement:
Science Archive (all Versions) must have available *commonly used catalogues* (see later) stored locally. Version 2.0 Science Archive may additionally hold SDSS (and other survey) pixel data for joint querying – see later).

C7:
Allows generation of finder charts via a web form
Rationale: Simple to provide and useful when observing at a site remote from the UK.
Implications: Software will be required for generation of pixel and/or ellipse plot finder charts. A web form will be required as the user interface.
Requirement:
Science Archive (all Versions) must allow generation of finder charts via a web form.

C8:
Holds housekeeping information for all archived data.
Rationale: It is essential to propagate all available data description (e.g. FITS header data) through to the Science Archive, to enable users to query those data
Implications: The Science Archive must be able to track between object catalogue records, image data files and the housekeeping data. For example, to protect proprietary data rights the Science Archive will need to validate queries against the source of any particular image subset (e.g. UKIDSS, PATT time, etc.)
Requirement:
Science Archive (all Versions) must hold housekeeping information for all archived data.

3.3 Security

A1:
Archived *data* must be accessible only by validated users
Rationale: The WSA will contain data resulting from internationally competitive science proposals. Proprietary rights of the UKIDSS consortium and open–time PIs/CoIs must not be compromised by data being freely available through the online archive.
Implications: The Science Archive must have security systems in place that prevent unfettered access by opportunistic users, but at the same time must not become so protected that access by valid users is hampered (e.g. by constantly asking for usernames/passwords). Security systems must be able to cope with various proprietary periods, and allow unfettered access after appropriate time intervals. All of this in turn implies user registration with username/password login and/or 'digital certification'.
Note: *Any* user (not just proprietors) should be able to derive information on what is in the archive without being given access to those data.
Requirement:
Science Archive data (all Versions) must be accessible only by validated users; archive *content* information should be available without restrictions.

A2:
Archived data must be uncorruptable by Science Archive users.
Rationale: Scientific exploitation will be compromised if data are corrupted.
Implications: Constant data ingest, recalibration of photometry/astrometry, and functionality enhancements imply a 'living' archive that is subject to change. This opens up the possibility of accidental corruption, especially by local archive managers with read/write access to filesystems. Archive design must minimise the possibility of accidental corruption, and also insure against data loss and minimise reconstruction times by invoking and appropriate backup policy.
Requirement:
Science Archive (all Versions) must be uncorruptable by Science Archive users.

A3:
Science Archive must allow data protection on the basis of proprietary data (per frame)
Rationale: Proprietary periods will be different for different observations (survey/non–survey).
Implications: Security systems must be able to cope with various proprietary periods, and allow unfettered access after appropriate time intervals.
Requirement:
Science Archive (all Versions) must allow data protection on the basis of proprietary data (per frame)

A4:
Science Archive must be quickly recoverable in the event of corruption by hardware/software faults etc.
Rationale: Clear need to ensure against data loss.
Implications: Science Archive will require backup on removable media and/or 100% redundant storage with data striping (i.e. fault tolerant hardware/software).
Requirement:
Science Archive (all Versions) must be quickly recoverable in the event of corruption by hardware/software faults etc.

3.4 Detailed requirements

The following requirements form the baseline for the WSA; they are an expansion of the top–level requirements above and items D in the 'Detailed Requirements'. Following T1 above, we have divided the requirements into those that must be in place for WFCAM first light and those that need fulfilling after a significant amount of data have accumulated. There are several reasons for this: i) the timescale for the delivery of WFCAM is short, so there is limited time for R&D concerning a large, scalable archive system; ii) such a system is not required at first light anyway since data volumes will be of limited size initially; iii) a phased approach means that the final large hardware purchase can be delayed as long as possible. So, we have grouped these into 'Version 1.0 requirements', and 'Version 2.0 requirements'; some requirements appear in the earlier version with limited scope, and in the later versions with full–blown functionality. We include some more long–term goals which may or may not be delivered, contingent on implementation and resource constraints, and delivery of appropriate tools/knowledge from related e–science projects (e.g. AstroGrid).

3.4.1 Version 1.0 requirements

T1/T7: The 'Version 1.0' working science archive *must* be in place in time for WFCAM first light (currently scheduled for September 2003).
T2: Science Archive must contain and serve pipeline processed data (pixels, object catalogues and housekeeping data) from both UKIDSS and other usage (e.g. open time, commissioning time).
T3: Science Archive will match UKIDSS survey requirements as they are currently specified, but will be flexible enough to follow changes in survey design.
T4: Science Archive will conform to any *existing* 'Virtual Observatory' standards and will be designed such that new standards can be easily incorporated, but must not be delayed by waiting for new developments to crystalize.
T5: Science Archive will allow *simple* (see below) queries.
T6: Science Archive must have an interface that is open to simple, intuitive use by the non–expert.
T9: Despite the need to expedite delivery of the WSA, development will be made *at all times* with due regard to scalability to VISTA data volumes.
T10: Science Archive must be able to cope with sub–optimal observations, and their subsequent displacement by better, repeated observations.
C1: Science Archive must contain calibrated object catalogues resulting from the pipeline, for both UKIDSS and open–time observations
C2: Science Archive must ingest and store pipeline output frames, allow rapid web–based access to images, and produce merged UKIDSS catalogues in a given field.
C3: Science Archive must be designed from the start to enable astrometric and photometric recalibration.
C5: Science Archive must be designed to allow public access to subsets of survey data on a variety of different search criteria (specified below).
C6: Science Archive must have available *commonly used catalogues* (see later) stored locally.
C7: Science Archive must allow generation of finder charts via a web form.
C8: Science Archive must hold housekeeping information for all archived data.
D1: Science archive must allow searching individual (or all) UKIDSS surveys on the following criteria (or combination of them):

- Position rectangle expressed in spherical co-ordinates: RA/Dec (J2000); l, b (Galactic) and λ, η (SDSS system)
- Circular sky patch within specified radius from given spherical co-ordinates: RA/Dec (J2000); l, b (Galactic) and λ, η (SDSS system)
- Circular sky patch within specified radius of a resolvable source name

D3: Science Archive must allow similar queries to be repeated for all objects in a user–supplied source catalogue.
D4: Science Archive must allow combinations of queries on UKIDSS data and the following other source catalogues:

- 2MASS
- SuperCOSMOS Sky Survey
- SDSS DR1 data release
- USNO–B
- FIRST source catalogue
- IRAS point source catalogue
- ROSAT All–Sky Survey catalogue

D6: Science Archive must have a simple interface for very quick searching on a given object name or position.
D8: Science Archive must return pixel images, confidence maps and catalogue data in gzipped FITS format, and must allow users to specify the output format of returned data as follows:

- FITS images with options for lossless and/or lossy compression
- ASCII (tab or comma–separated) or FITS table, for object catalogue and housekeeping data
- Space–separated ASCII with CDS–type descriptors for object catalogues
- VOTable – http://vizier.u-strasbg.fr/doc/VOTable – a proposed protocol for exchange of astronomical data embedded in XML.

D9: Science Archive must be able to return pixel data in any available passband, over a contiguous field up to one 'tile' (0.8°) across together with a matched catalogue.
D11: Science Archive must be able to generate and return stacked images given a user–selected list of input images and the standard stacking algorithm in the CASU basic pipeline.
D12: Science Archive must be able to generate and return merged multi–colour, multi–parameter catalogues with the best available photometric and astrometric calibrations.
D13: Science Archive must support federation with the source catalogues specified in D4 above
D14: Science Archive must be able to generate and return meaningful optical/IR colours for all objects in the overlap with the existing SDSS data where counterpart detections occur in the SDSS object catalogue.
D16: Science Archive must support the returning of only a subset of the entire possible array of object parameters.
D19: Science Archive must be able to produce a finder chart of size up to 10 arcmin for any region within which survey data exist, returning ellipse detection plot and/or a single colour pixel plot, as specified by the user.
D20: Science Archive must allow access to best or duplicate data for objects in overlapping survey data.
D21: Science Archive must allow general access to all housekeeping data – e.g. for a given survey area, what is currently available, how good it is, etc.
D22: Science Archive must store uncalibrated quantities, calibrated quantities and the calibration model/coefficients. Archive output must therefore include (in headers)

- Archive version identifier

- calibration version identifier

D23: Science Archive must allow a summary of data available to be generated for a given search region.
A1: Science Archive must be accessible only by validated users.
A2: Science Archive must be uncorruptable by Science Archive users.
A3: Science Archive must allow data protection on the basis of proprietary data (per frame).
A4: Science Archive must be quickly recoverable in the event of corruption by hardware/software faults etc.

User access is to be through web forms providing fill–in boxes and button clicks, and also via an SQL query form interface; a command–line interface for remote users to bypass interactive webforms will also be provided.

The summary in Section 4 gives an explicit statement of the Version 1.0 WSA contents and functionality.

3.4.2 Version 2.0 requirements

In addition to the Version 1.0 requirements:

T1: A fully functioning archive system, as defined by the requirements (and where possible, goals) herein, must be available as soon as possible after WFCAM first light, and no later than 1 year after survey operations begin in earnest.
T4: Science Archive must eventually conform to internationally agreed VO standards in access tools, data product formats and transfer protocols.
T5: Science Archive will allow usages at varying levels of complexity (as defined later).
T7: Interim survey products must be released to the community on timescales determined by WFCAM observing periods (i.e. a survey 'release' will occur as soon as possible after each observing period, and before the end of the following period).
T8: Science Archive must allow requests for arithmetic operations, and options from an advanced processing toolkit (see later), on pixel data.
T9: WSA solution must be scalable to VISTA data volumes.
T11: Science Archive must have some capability for the remote user to carry out data exploration and interaction in real time: the Science Archive response time should be $\sim$ 100sec for wholesale trawl–type queries.
C4: Science Archive must be able to cross-calibrate photometric information using areas of overlap between processed frames, where available.
C6: Science Archive must have the final SDSS catalogues (and, if possible, images) stored locally, in addition to the catalogues specified for the Version 1.0 Science Archive.
D1: Science Archive must allow searching individual (or all) UKIDSS surveys on the following criteria (or combination of them):

- Search positions specified at arbitrary equinox and time system, and additionally ecliptic and super–Galactic systems
- Source colour in any linear combination of those colours available for any given survey
- Source parameter ranges

D2: Science Archive must allow searching within open–time programme data using the same criteria as D1 (where possible), returning whatever data are available.
D4: Science Archive must allow combinations of queries on UKIDSS data and the following other source catalogues:

- most recent SDSS data release, as per availability
- User supplied catalogue for complementary imaging (at any wavelength) for any of the UKIDSS sub–surveys.
- any general user–supplied catalogue at any wavelength (eg. GLIMPSE, ASTRO–F)

D5: Science Archive must allow arithmetic functions to be used in setting up complex queries (e.g. for a colour index not stored in survey catalogue tables)
D6: Science Archive must have a remote GUI application for formulating queries (e.g. an interface analogous to the SDSS Java–based query tool).
D7: Science Archive access GUI must allow plotting of returned parameters, in selected (X,Y) pairs or histograms, and also provide basic fitting routines.
D10: Science Archive must be able to generate (on–the–fly) and return larger (than D9) areas from survey data traversing survey tile boundaries, blocked down as specified by the user, in formats specified in D8.
D11: Science Archive must be able to generate and return stacked images using user–specified (see later) stacking algorithm options.
D12: Science Archive must be able to generate and return merged multi–colour, multi–parameter catalogues with the best (or previous as specified by the user) photometric and astrometric calibrations.
D13: Science Archive must support federation with the source catalogues specified in D4 above
D14: Science Archive must be able to generate and return meaningful optical/IR colours for all objects in the overlap with the SDSS, whether or not detected in the SDSS data (i.e. it must be possible to place an aperture in and measure the flux from SDSS image data given the position of an IR source detection).
D15: Science Archive must support ANDing of one query with another, where both have already been executed.
D17: Science Archive must allow trial–and–error searches (e.g. return the number of source hits rather than the output results), for any valid query
D18: Science Archive must allow repetition of queries using previous versions of astrometric and photometric calibrations.
D19: Science Archive must be able to produce a finder chart for any region within which survey data exist, returning a colour pixel plot, as specified by the user, generated from available single–passband images of the same field.
D20: Science Archive must allow access to best or duplicate data for objects in overlapping survey data, and must contain proper motion measures for objects where multi–epoch position measurements exist.

Section 4 gives an explicit statement of the Version 2.0 WSA contents and functionality.

3.4.3 Goals

T11: Science Archive response time should be $\sim$ 10sec for wholesale trawl–type querying.
C6: Science Archive will, insofar as external developments allow, be integrated into the 'Virtual Observatory' (VO) as a general solution to rapid, online cross–referencing with any published astronomical catalogues that are also contained within the VO.
D1: Science Archive may recast web services as 'Grid services' (a Grid–based solution to user access) in collaboration with AstroGrid.
D4/13: Science Archive may allow combinatorial queries with catalogues anywhere on the 'data–Grid', i.e. may allow database federation across the grid.
D7: Science Archive will aspire to the mantra 'ship the results, not the data', i.e. may allow remote procedure calls to advanced manipulation tools and may allow user upload of analysis codes.
D10/11: Science Archive may ultimately support advanced visualisation tools, e.g. large area, panoramic pseudo–colour images with panning in real time; three–dimensional catalogue parameter plotting and rotation.

4 SPECIFICATION OF CONTENTS AND FUNCTIONALITY

At its meeting on 2002 November 25, the UKIDSS Consortium met and discussed the requirements and usages along with the WSA development plan. The Consortium suggested several changes along with some issues for discussion. The results of these discussions have been folded into this document, yielding the following specification (in as much detail as is possible at this time) for the WSA functionality and contents at Versions 1.0 and 2.0 (note: this specification will be developed in later documents). The V2.0 requirements can be considered 'goals' of V1.0.

4.1 Version 1.0

WSA Version 1.0 is deliverable at WFCAM first light (currently scheduled for September 2003). In addition to the following, WFAU undertakes to apply UKIDSS–specified algorithms, and import UKIDSS–supplied catalogues, to the WSA in lieu of automatic tools for such functionality (see Version 2.0).

4.1.1 Contents

The V1.0 WSA will contain the following information in a relational DBMS:

1. *Observations Information* containing details of observations contained in the archive and their generic properties;
2. *Image Information* containing details of all images (stored as flat files) in the archive along with housekeeping data (from stripped FITS headers);
3. *Observations Catalogue Information* containing the object catalogues, generated by the CASU standard pipeline, associated with each image, and list–driven source catalogues between the different passbands in any given field;
4. *Merged Catalogue Information* for each of the accumulating UKIDSS subsurveys LAS, GPS and GCS (merged in the sense that the 'same' objects observed in different colours and/or at different times will be merged into one multi–colour, multi–epoch record);
5. Catalogues for 2MASS, SDSS DR1, SSS, USNO–B, FIRST, IRAS and ROSAT–ASS surveys;
6. A *Survey Progress Catalogue*, containing for each of the 5 UKIDSS subsurveys information on observations taken to date;

and also image data (pixels with confidence maps; default stacks for the) deep surveys; and difference images for the GPS K band) in flat files, along with a large reserve (scratch) workspace for use during querying. The V1.0 WSA will also contain online documentation and 'cookbook' style worked examples to aid users.

4.1.2 Functionality

The V1.0 WSA will have the following access points:

1. A web interface allowing searching of individual (or all) UKIDSS survey catalogues on the following criteria:
   - position rectangle expressed in spherical co-ordinates: RA,Dec (J2000); l, b (Galactic); and λ, η (the SDSS spherical co-ordinate system)

- circular sky area within specified radius from given spherical co-ordinates: RA,Dec (J2000); l, b (Galactic); and λ, η (the SDSS spherical co-ordinate system)
- circular sky patch within a specified radius of a resolvable source name (using the CDS/NED name resolver)

and additionally the same searching functions on a user–specified ASCII (space separated) of centres (sexagesimal or decimal degrees) and search radii (i.e. a batch mode search). This interface will also produce ellipse plots for use as finder charts. For an example of such an interface, see WFAU's SuperCOSMOS Sky Survey access page http://www-wfau.roe.ac.uk/sss, particularly the 'Get a CATALOGUE' interface.

2. A web form interface allowing *querying* of individual (or all) WSA catalogues (e.g. UKIDSS survey catalogues, housekeeping data, details of archived images) via Structured Query Language (SQL), with push–button options for the format of output data:

   - ASCII (space, tab or comma–separated);
   - FITS binary tables;
   - VOTable format;
   - uncompressed or lossless compression (e.g. gzip);

   combinatorial queries with the 2MASS, SSS, SDSS–DR1 and USNO–B catalogues will be provided for. For an example of such an SQL interface, see WFAU's 6dF access interface at URL http://www-wfau.roe.ac.uk/6dFGS/SQL.html.

3. A web form interface that returns pixel data (images and confidence maps) given an arbitrary input position (as in 1 above) and size up to 0.8° (one WFCAM tile) as follows:

   - mosaiced across any frame boundaries as necessary;
   - FITS format, with user–specified options for lossless or lossy compression;
   - with corresponding *merged* catalogue (supplied as FITS binary extension).

   For an example of such an interface, see WFAU's SSS page (URL above), particularly the 'Get an IMAGE' facility.

'Remote server' functionality for web–based browsing tools (e.g. SkyCAT/GAIA/Aladin) will be provided for some of the above image/catalogue servers, along with a command line interface for remote user non–interactive web access. Archive response time for catalogue queries will be rapid for indexed quantities as follows: position, magnitude, colour, and image class.

4.2 Version 2.0

Version 2.0 is deliverable no later than one year after survey operations begin, and will include more 'database driven' products and features. In addition to contents and functionality provided in V1.0, the following specifies the V2.0 contents and functionality.

4.2.1 Contents

The V2.0 WSA will additionally contain:

1. Externally provided catalogues and pixel data (UKIDSS complementary imaging, and SDSS data release as avialable at that time);
2. A database of open–time observations;
3. Enhanced UKIDSS catalogues containing derived information (e.g. proper motions, dereddened colours, catalogue parameters from placing apertures on SDSS pixels at WFCAM detection positions) where possible using available data.

4.2.2 Functionality

In addition to the simple access tools provided in V1.0, one (or more) advanced GUI(s) will be provided that have the following functionality:

1. User–specified options for stacking pixel data, i.e. select images to be stacked, and the stacking algorithm from a choice of: i) unweighted; ii) sensitivity weighted; iii) psf matched ...; iv) ... further toolkit options ...;
2. Arbitrary sized, mosaiced images (across tile boundaries), blocked down as appropriate, with a multi–colour option;
3. Source extraction options on any specified subset or bespoke stack of pixel data: i) CASU standard source extraction; ii) SExtractor; iii) mutiple simultaneous profile fitting (i.e. DAOphot–like); iv) ... further toolkit options ...;
4. Data exploration/interaction facilities: simple XY plotting; histogram plotting; simple model fitting routines (generalised least–squares with robust outlier rejection);
5. Automatic user–supplied catalogue ingest facility for joint querying with existing catalogues;
6. Enhanced output format options to include any new Virtual Observatory standards available at that time;
7. Ability to analyse archive pixel data (both WFCAM and other, e.g. SDSS) at arbitrary positions defined by an input list of positions, apertures and/or profiles types/models (ie. list–driven photometry for *any* data);
8. Generalised difference imaging (and subsequent source analysis)
9. Persistence of multi–stage usage/query; storage of intermediate user–generated results sets

Additionally, the web–based access tools in V1.0 will be supplemented with a 'web service' interface (eg. a non–interactive access tool employing XML format data transfered using Simple Object Access Protocol) to provide, where appropriate, non–interactive access to pixel and catalogue data. Archive response time is to be $\sim$ 100sec for wholesale catalogue trawls on non–indexed quantities.

4.3 Later Versions

At this time, we make no explicit statement concerning the functionality of subsequent WSA versions.

A APPENDICES

A.1 Background

WFCAM (see http://www.roe.ac.uk/atc/projects/wfcam/index.html) will enable the next generation wide–angle sky survey to be undertaken in the UK. It follows on from the hugely successful UK Schmidt photographic surveys of the last decades of the twentieth century, the major difference between the old and the new being the data rates and volumes that will be produced. WFCAM employs 4 2k×2k Rockwell devices and has an instantaneous field–of–view of 0.21 square degrees. WFCAM is expected to be on–telescope for the greater fraction of all available UKIRT time, and will have average/peak data rates of 100/230 Gbytes per night. It will commence operations in the final quarter of 2003. VISTA, on the other hand, is a dedicated survey telescope with an IR camera employing 16 2k×2k devices in a 0.44 square degree FOV. The data rate for VISTA will be $\sim$ 400 Gbytes per 10 hour night, and this facility is expected to begin operations in the third quarter of 2006. In terms of both timescale and scope, WFCAM therefore represents a natural 'stepping stone' to VISTA, and in the overall scheme of UK wide–field astronomy the WFCAM project can be thought of as 'VISTA phase A'.

There is, of course, a clear need for 4m survey facilities in the era of 8m–class telescopes; the relative performance of WFCAM (as measured by its 'grasp', or information gathering product AΩ) shows (see, for example, the original VISTA science case, available from http://www.vista.ac.uk/) that it is amongst the world's leading IR survey instruments, even when including other non–dedicated survey facilities such as VLT–IRMOS. The combined science case (for complete details, follow the URL http://www.ukidss.org/sciencecase/sciencecase.html) proposed by the UKIDSS consortium for WFCAM, for example, details a programme that is unrivalled in terms of depth, field of view and therefore survey volume. UKIDSS proposes a nested series of surveys ranging from the Large Area Survey ('LAS', 4000 sq. deg. to K=18.4), the Galactic Plane Survey ('GPS', 1800 sq. deg. to K=19), the Galactic Clusters Survey ('GCS', 1600 sq. deg. to K=18.7), the Deep Extragalactic Survey ('DXS', 35 sq. deg. to K=21) to the Ultra–Deep Survey ('UDS', 0.8 sq. de.g. to K=23). The image data alone for these amounts to $\sim$ 50 Tbytes of data, while the object catalogue and ancillary information are likely to be many Tbytes in size. VISTA survey data volumes will likely be more than 5× those of WFCAM.

A.2 The need for science archives.

The question naturally arises as to how science exploitation of such large datasets will be undertaken. Data volumes will simply be too large for users to download and keep their own copies. Raw data processing is likely to be complicated, while calibration procedures will evolve as cameras are better characterised and more calibration data are obtained. Reprocessing of substantial amounts of pixel data may be necessary in the light of improved algorithms or for specific 'non–standard' science goals. Once data are reduced using standardised pipeline procedures, the establishment of a centralised 'science archive' offers the greatest potential for full science exploitation (see the paper presented by Lawrence et al. at the 2002 SPIE meeting in Kona, Hawaii; available online at http://www.roe.ac.uk/~nch/wfcam/misc). Again, calibration procedures can be more easily developed and applied in a controlled manner to data in a central repository – it makes sense to solve data–specific reduction and calibration problems once, yielding an optimal solution. Early community access to well calibrated data will facilitate timely science exploitation. A well constructed science archive will enhance greatly the scope of research that can be done with the survey data; in fact, many science applications will only be feasible via a sophisticated science archive. For example, much of the science that will be done with the UKIDSS LAS will rely on complementary data from the SDSS and other non–IR wavelength surveys. Given the volume of all of these datasets, some thought needs to go into the design of the archive to enable full exploitation.

WFCAM Science Archive Science Requirements

The WFCAM science Archive must provide rapid and straightforward access to WFCAM data from both the UKIDSS surveys and generalised open–time usage. This document describes the expected inputs (output from the CASU pipeline), some of the expected archive usage modes, and gives more specific requirements in tabular form.

Top–level requirements

T1	Provides the maximum possible potential for capitalizing on the UKIDSS surveys. *Rationale: UKIDSS amounts to 80% of WFCAM time and rapid exploitation is crucial to the UK's future strategy. Since exploitation is through the science archive, the need for T1 is clear.*
T2	Contains data from both UKIDSS surveys and open–time usage *Rationale: WFCAM data are likely to be far larger than the average UK University site can hold. Standard pipeline products held centrally are the solution.*
T3	Is flexible and copes with alterations in the UKIDSS survey design over time *Rationale: UKIDSS has an initial two–year allocation followed by review. This review may result in changes in both priority and design of the surveys. It is imperative that the science archive should not be a blockage in this process.*
T4	Does not preclude usage from the GRID and inclusion in Virtual Observatory *Rationale: obvious, and clearly relates to T1*
T5	Allows simple and complex queries with appropriate interfaces for both *Rationale: a single complex interface suitable for data federation would be excessive when all the user wants is a K frame around a given target.*
T6	Is simple to use for PR purposes *Rationale: UKIDSS is a truly world–beating undertaking. The science archive will be the obvious point from which to draw PR material and should not be a hindrance to this.*
T7	Allows access to survey data before all observations are complete, and must not be disrupted by regular ingest of new survey data. *Rationale: This is consistent with rapid exploitation (and was a PDR requirement). Access to interim releases of the data will clearly be needed.*
T8	Must allow arithmetic operations and options from the advanced processing toolkit on pixel data to be specified in requests *Rationale: multicolour imaging is perfectly possible given standard pipeline output and should not require the data to be transferred to the users home site just to stitch three colours together.*

T9	Must be scalable to VISTA data volumes *Rationale: WFCAM and VISTA represent the UK's deep infrared survey capability. To use the same science archive design for both will ensure maximum exploitation of both.*
T10	Must be able to merge reduced frames taken in nonphotometric conditions with other data from the same survey.
T11	Has some capability for the remote user to carry out data exploration and interactive analysis tasks. *Rationale: Many situations (eg KX selection of quasars) are anticipated where the remote user will want to manipulate and visualize large amounts of data without wanting to download large amounts of data to his/her local machine.*

Science archive contents and functions (minimum)

C1	Contains calibrated object catalogues resulting from the pipeline, for both UKIDSS and open−time observations *Rationale: obvious, basic science archive function.*
C2	Ingests and stores pipeline output frames for later online processing, generates compressed pixel images on the fly for rapid web−based access, carries out immediate cross−referencing with existing UKIDSS survey data and produces consolidated UKIDSS catalogue in a given field *Rationale: again basic science archive functionality. Compression by at least a factor of order 10 is necessary in the early stages, whether or not the need disappears later due to enhanced network speeds. It is assumed (but not required) that this will be lossy compression.*
C3	Is able to recalibrate a given field or fields in the event of revised calibration information (specifically, photometric and astrometric), and allow database queries on the recalibrated quantities *Rationale: changes in calibration information are frequently encountered in survey operations, and the science archive itself may lead to such changes.*
C4	Is able to cross−calibrate photometric information using areas of overlap between processed frames, where available. *Rationale: this is not a sensible function of the CASU pipeline, which is required only to produce results on a night−by−night basis. The science archive will have all photometric information and calibrations for all superframes, and is where this should happen.*
C5	Allows public access to subsets of survey data on a variety of different search criteria (specified below) *Rationale: basic science archive functionality.*
C6	Allows rapid on−line cross−referencing of search results with other catalogues *Rationale: consistent with T1, this requirement is expanded on later.*

C7	Allows for generation of finder charts via a web form *Rationale: simple to provide and useful when observing at a site remote from the UK*
C8	Holds housekeeping information for all archived data

Detailed requirements

Consistent with the above, the Science archive:

D1	Must allow searching individual (or all) UKIDSS surveys on at least the following criteria (or combination of them) – Positional rectangle RA, Declination (J2000 coordinates) – Positional rectangle Galactic Coordinates (l^{II},b^{II}) – Circular sky patch within a specified radius from a given RA, Dec – Circular sky patch within a specified radius from a given Galactic Coordinate – Circular sky patch within a specified radius of a resolvable source name – Source colour index (e.g. YJHK magnitudes and any linear combination thereof) between any of the available survey colours for the particular survey – Source parameter ranges (using any of the available source parameters) *Rationale: These are all basic science archive functions*
D2	Must allow searching within open−time programme data on the same criteria (where possible), returning whatever data are available (may only be image frames) *Rationale: open−time data become UKIRT science archive data as normal, and in many instances will be as searchable as data from the UKIDSS survey.*
D3	Must allow similar queries to be repeated for all objects in a given user−supplied source catalogue, or those in any catalogue also stored locally to the UKIDSS science archive. *Rationale: the user may have their own catalogue generated at a different wavelength, and may wish to obtain UKIDSS catalogue or image data at this list of field centres.*
D4	Must allow combinations of queries on a number of different source catalogues *Rationale: this requirement anticipates compound queries such as "return all UKIDSS image data around infrared sources with axial ratios greater than 1.5 and which have no detection in the SDSS". An arbitrary number of source catalogues should be allowed, subject only to their local availability. This is not the place to specify which catalogues should be available, just that the system shouldn't preclude this type of query; perhaps goals would be the SDSS and Schmidt catalogues.*
D5	Must allow functions [methods] to be used in setting up complex queries (i.e. for colour index if not stored separately) *Rationale: for consistency with T9.*

D6	Must have a web–based user interface to the object catalogue database implemented in an SX–like manner, and pixel database access implemented with in–house software. Ideally a simple interface for very quick searching on a given object name or position would be provided separately. *Rationale: clearly a web interface is needed, at least in the short term, and the SX similarity requirement follows from the desire to maximise the power of matching with Sloan for the LAS. For very quick returns of UKIDSS data around a given object, either the basic web interface must be straightforward enough to make this very quick, or a separate form should be provided.*
D7	Must allow plotting of returned parameters direct from the user interface, in selected pairs (x–y plots) or histograms (in the case of a single parameter)
D8	Must return pixel images, confidence maps and catalogue data (where reduction recipe generates it) to the user in (at least) gzipped FITS format
D9	Must be able to return pixel data in any available passband, over a contiguous field up to [TBD but one tile across seems a sensible minimum], together with matched object catalogue
D10	Must be able to generate (on–the–fly) and return larger areas from survey data, blocked down as required
D11	Must be able to generate and return stacked images using a variety of stacking algorithms, where survey data are available
D12	Must be able to generate and return merged multi–colour multi–parameter catalogues with best available astrometric and photometric calibration from survey data
D13	Must support database federation with other source catalogues (initially SDSS and Schmidt sky survey)
D14	Must be able to generate and return meaningful optical/IR colours for all objects in the overlap with SDSS, whether or not detected in the Sloan data
D15	Must support ANDing of a given query with another, where both have already been executed
D16	Must support the returning of only a subset of the entire possible array of object parameters
D17	Must allow trial–and–error searches (e.g. return number of source hits rather than entire catalogue), for any valid query
D18	Must allow repetition of queries using previous versions of astrometric and photometric calibration
D19	Must be able to produce a finder chart for any region within which survey data exist, returning both object shapes and positions (ellipse map) and a colour pixel image if suvey data in more than one colour are available.
D20	Must allow access to best or duplicate data for objects in overlapping survey data (e.g. to generate light curves and/or proper motions from duplicate observations of the same object).
D21	Must allow general access to all housekeeping data – e.g. for a given survey area, what is currently available, how good it is, etc.
D22	Calibrated quantities must be specified through coefficient values and calibration scheme
D23	Must allow a summary of data available to be generated for a given search region

Security

A1	Archived data must be accessible only by validated users
A2	Archived data must be uncorruptable by science archive users and managers
A3	Allows data protection on the basis of proprietary date (per frame)
A4	Science archive must be quickly recoverable in the event of corruption by hardware/software faults etc. Rationale: Clear need to insure against data loss (e.g. backup on removable media and/or 100% redundant storage with data striping)

Usage examples [UKIDSS–specific examples to follow]

SIMPLE QUERY

User requests LAS survey data within 20 arcminutes of (say) NGC 4565.
Science archive returns (i) image data for all available colours in the LAS to date, (ii) source catalogue data for the field (see C2).

User requests one arcminute JHK colour images of all Gliese nearby stars fainter than 10^{th} magnitude.
Science archive returns true–colour JHK images, one FITS file per object

COMPLEX QUERIES

User requests a list of UKIDSS colour indexes for all QSOs in a source catalogue provided by them.
Science archive returns the list, including only the colour index parameters.

User requests a cross–reference between, for example, all double radio sources from a radio survey with UKIDSS J–K greater than a given value. User further specifies that only source identification and axial ratio information is required.
Science archive returns a merged list of axial ratios and identifications. User then clicks for a histogram of axial ratio.

INPUT FROM CAMBRIDGE PIPELINE

The CASU pipeline output will be the only input to the science archive.

Relevant definitions:

Processed frame	Result of reducing the data from a single array which arise from a complete microstep sequence (pointing) or jittered series of pointings (or other currently unforeseen sequence)
Tile	Result of reducing the data from all four arrays after completion of four microstepped sequences at positions which fill in the gaps between arrays.

Inputs to the archive will be as follows. Observing modes are detailed later:

P1	FITS image of each processed frame (one per array, N(arrays) per pointing; this applies in observing modes O1 to O3).
P2	FITS line–emission processed frame (in observing mode O8).
P3	Photometric calibration information and confidence maps for each processed frame
P4	Catalogue of objects detected in the processed frames (one per array; this applies in observing modes O1 to O3). Catalogued image properties are detailed elsewhere.
P5	For tiled observing modes, both the products detailed in P1–P4 as appropriate, plus catalogue and confidence maps for the completed tile will be supplied.

DQ information

The CASU pipeline will also provide the following QC information via the FITS headers of the output images. Some of these will not be relevant but they should be easy to provide:

F1	Sky brightness and sky noise measurement (per filter)
F2	Mean stellar image diameter (determined by fitting within the processed frames, filed as a calibration after conversion to K–band equivalent).
F3	Instrumental zero point determined from 2Mass stars in the survey processed frames (or the difference between this and the expected value). Stored as a calibration and made available to the QT.
F4	"Locally photometric" flag (determined from variations in detected image properties in the frames in the microstep sequence).
F5	"Globally photometric" flag (determined from all previous observations of standard fields on the current night).
F6	Limiting magnitude (5σ) in the processed frame.
F7	Mean axial ratio of suitable stellar images.
F8	Error on astrometry in catalogues derived from reduced frames.
F9	Nightly photometric zero point from observations of standard fields

For information, the pipeline operates on the following inputs

O1	1x1 microstep
O2	2x2 microstep sequence
O3	3x3 microstep sequence
O4	Complete tile of four 1x1 microsteps
O5	Complete tile of four 2x2 microsteps
O6	Complete tile of four 3x3 microsteps
O7	All of modes 1–6 may be repeated immediately in a series of filters. It is not anticipated that filter changes will be made within the space of one microstep sequence, nor that any Observation will consist of Integrations in multiple colours.
O8	All of modes 1–6 may be carried out in narrow–band filters.

O9	2x2 microstep or microstep/mesostep sequence on a standard field (mesostep is an offset sequence large enough to allow determination of the vignetting function).
O10	Large extended field (complete tile interspersed with jitter frames on a remote sky patch)
O11	Crowded field: 2x2 or 3x3 microstep sequence with equal number of interspersed sky observations

Usages of the WFCAM Science Archive

Revision date: Jan 9th 2003

Nigel Hambly and Ian Bond

Here are some example "usages" of the WFCAM science archive. We use the term "usage", rather than "query" (used, for example, in a similar exercise undertaken by the SDSS archive team in specifying the SQLServer implementation of the SDSS Science Archive) because it seems that a lot of the applications that the user would want are going to involve a combination of interactive analysis and on-the-fly processing (or indeed feeding the results of one application into another) as well as formulating SQL queries. Furthermore, these usages concentrate on science applications rather than Science Archive/DBMS implementation details. Note that we prefer the term "usage" to "use case" since the latter is commonly used in Universal Modelling Language to describe the general, high-level interaction between a system and an "actor", whereas the following examples are specific functions of the Science Archive.

As a starting point for generating these usages, we have used the UKIDSS proposal, specifically the 11 "key science goals" (proposal Section 1.5). We then detail other usages extracted from science goals within the subsurvey components of UKIDSS, and finally give some more general examples of data exploration and data mining. The usages have been supplemented with some discussion and analysis (appended) following review by the UKIDSS Consortium.

U1: *Count the number of sources in the LAS which satisfy the colour constraints (Y-J) > 1.0, (J-H) < 0.5 where SDSS i,z flux limits at the same position are less than 2-sigma. User then refines the query as necessary to give a reasonable number of candidates. When satisfied, the user requests a list, selecting output attributes from those available for the LAS, and finder charts in JHK for each object.*

This is to search for the nearest and faintest substellar sources (first UKIDSS key science goal).

U2: *List all star-like objects with izYJHK SDSS/UKIDSS-LAS colours consistent with the colours of quasars at redshifts 5.8 < z < 7.2 or z > 7.2 (user specifies cuts in colour space). Return plots of (i-z) v. (z-J) and (i-Y) v. (Y-J) with these sources plotted in a specified symbol type, with 1 in every 10,000 other stellar sources plotted as points.*

This is to search for the highest redshift quasars, and break the z=7 QSO barrier (second and third UKIDSS key science goals).

U3: *For a given cluster target in the UKIDSS GCS, make a candidate membership list via selection of stellar sources in colour-magnitude, colour-colour and proper motion space. Cross-correlate the candidate list against a user-supplied catalogue of optical/near-infrared detections in the same region.*

Allows the user to determine the substellar mass function for the cluster (fourth UKIDSS key science goal).

U4: *From the UKIDSS LAS, provide a list of all stellar objects that have measured proper motions greater than 5x their estimated proper motion error; additionally give a count of all stellar objects that are unpaired between the two epochs of the LAS observations with specified conditions on image quality flags. User then refines these conditions to produce a manageable list of very high proper motion candidate stars. Return finder charts in JHK for all candidates.*

Investigate potential Population II BDs (fifth UKIDSS key science goal) and also very cool, helium-atmosphere Halo WDs (if they exist; LAS Section 2.2) via their high proper motion.

U5: *From the UKIDSS DXS & UDS, construct galaxy catalogues. User selects all non-stellar sources satisfying quality criteria. User also requires the spatial sampling of this catalogue. Cross-correlate the galaxy catalogues against user-supplied optical catalogue in the same region.*

Construct basic galaxy catalogues as a tool for further study (e.g. for z=1 and 3; measure the growth of structure from z=3 to the present day and clarify the relationship between QSOs, ULIRGs and galaxy formation: sixth, seventh and ninth UKIDSS key science goals)

U6: *From the UKIDSS LAS, construct a galaxy catalogue for all non-stellar sources satisfying K < 18.4 and given quality criteria; return full photometric list from SDSS & UKIDSS: ugrizYJHK. User also requires the spatial sampling of this catalogue.*

Construct an infrared-selected galaxy catalogue for local studies (seventh UKIDSS key science goal).

U7: *From the UKIDSS UDS, select a sample of galaxies with colours and morphology consistent with being elliptical galaxies. Provide a spatial mask to enable determination of sample characteristics. Provide a measure of the half-light radius for each galaxy.*

Constructs a candidate elliptical galaxy sample at high redshift. This is the first basic step in studying the epoch of spheroid formation (eighth UKIDSS key science goal; see also UKIDSS proposal Sections 6.2 and 6.4 for details).

U8: *From the UKIDSS GPS, provide star counts in 10 arcmin cells on a grid in Galactic longitude and latitude; also provide a list of cells where there is any quality issue rendering that cell's value inaccurate.*

Enables mapping of the Milky Way through dust via infrared star counts (tenth UKIDSS key science goal).

U9: *From the UKIDSS GPS, provide a list of all sources that have brightened by a given amount in the K band.*

Provides the means to increase the number of known YSOs (including rare types such as FU Orionis stars) by an order of magnitude (eleventh UKIDSS key science goal).

U10: *Provide a plot of g-J vs J-K for all point-like sources detected in the UKIDSS/LAS survey subject to quality constraints. User interacts with the plot to fit a straight line (g-J)=a+b(J-K) to the main sequence stars. Then find all UKIDSS/LAS sources with g-J>a+b(J-K), 4>g-J>-1, and 3>J-K>-1.*

This is how a remote user may select quasars using the "K-excess" method. This is an example where remote interactive analysis is required (LAS KX-selected quasars; UKIDSS Section 2.6).

U11: *Construct H_2 - K difference image maps for all CCD frames within a specified subregion surveyed by the GPS.*

These difference image maps are used to study "macrojets" (UKIDSS GPS Section 3.2)

U12: *Find all galaxies with a de Vaucouleurs profile and infrared colours consistent with being an elliptical galaxy in the Virgo region of the UKIDSS LAS.*

e.g. UKIDSS LAS Section 2.4 Virgo (4); to study the IR morphology of specific galaxy types in the Virgo cluster.

U13: *Given input co-ordinates and a search radius (arbitrary system and reference frame) provide a list of all WFCAM observations ever taken that contain data in all or part of the specified area.*

U14: *Provide a list of point-like sources with multiple epoch measurements which have light variations > 0.1 magnitudes in J, H or K.*

U15: *From any UKIDSS data, where multiple epoch measures exist for the same object, provide a list of anything moving more than X arcsec per hour.*

U16: *Provide a list of star-like objects that are 1% rare for the 3-colour attributes.*

This involves classification of the attribute set and then a scan to find objects with attributes close to those of a star that occur in rare categories.

U17: *For a given device in a tile, give me all images from the UDS corresponding to that frame, stacked in 10 day bins.*

U18: *Give me a true colour JHK image mosaic using frames in the LAS centred at given co-ordinates (arbitrary reference frame and system) with 2 degree width and rebinned so that the entire mosaic is returned as a 2048x2048 pixel image.*

U19: *Find all detected sources from all UKIDSS sub-surveys within 3x the error boxes of a user supplied list of X-ray transient sources.*

Here we are searching the archive for counterparts to transient X-ray sources. One could also run this query to request pixel data.

U20: *For all sources in a user-supplied radio catalogue of HII regions in the GPS, return the Br-gamma surface brightness in an aperture of X arcsec*

This usage example specifically requires analysis of H_2 - K difference image pixel data with a user-specified list of aperture positions and radii.

Discussion and analysis

Additional usage examples, along with rationales, have been provided by UKIDSS. These were discussed at the 2002 November 25th Consortium meeting and a summary of those deliberations is presented here. In general, the Science Archive should be viewed as a system for managing and analysing large amounts of data and *enabling* science exploitation; however it cannot be all things to all people, and will not actually *do* the required science. Hence, the question of GUI functionality is necessarily restricted to those tasks that it would be impossible for the user to achieve (for example, requiring transfer and analysis large data volumes). Features requested for exploring and analysing small datasets are not a priority for the WSA design.

Large Area Survey (LAS):

No further usage examples were forthcoming from the LAS.

Discussion:

The key issue for the LAS is the cross-match with the SDSS optical survey. The question "*is absence from the SDSS catalogues sufficient, or does the LAS require flux measurements at arbitrary positions in pixel data?*"; was posed, i.e. for example in U1 above, it may be necessary for the Archive to be able to query the LAS catalogue in conjunction with SDSS pixel data (as opposed to simply the SDSS object catalogues). In the ensuing debate it became clear that there was a strong requirement for this. As for timescales, the Consortium was content to state that "it would be *nice*" to be able to do this at V1.0. For the purposes of the WSA, we will, therefore, undertake to look into this as a goal for V1.0 and a requirement for V2.0, but final implementation will depend on feasibility and resource constraints.

Galactic Plane Survey (GPS):

GPS 1: *Return list of all bright and dark nebulae within 10 arcmin of a specified galactic or equatorial coordinate, with size, average surface brightness (darkness) at JHK, and best elliptical fit to shape.*

GPS 2 (following U8): *Return J:K and H:K ratios of star counts in 10 arcmin cells in the GPS, on a grid in Galactic Longitude and Latitude.*

This gives information about extinction through the plane and detects clusters in distant spiral arms.

GPS 3: *Return list of all stellar clusters in a specified rectangular galactic or equatorial coordinate box, with number of projected members, cluster radius (from King profile) and cluster contrast (perhaps defined as ratio of cluster stellar density to average local stellar density at the same galactic latitude)*

GPS 4: *Return K band extinction and individual solutions to distance modulus, true colour, dereddened apparent magnitude, spectral type and luminosity class for all point source J+H+K detections projected within a specified GPS cluster. Each solution must have a confidence rating or distance error attached and ambiguous solutions must have relative probabilities.*

This type of data can be used to work out the distance to a cluster, either from the mode of solutions with >90% confidence or by main sequence fitting in a colour-magnitude diagram. This contributes to the 3-D atlas.

GPS 5 (following U11): Make J-K difference image map in magnitudes for a user specified rectangular region.

Provides extinction info for a nebulous star formation region.

GPS 6 (following U4, U1): *Return list of GPS stars with detected proper motion (5-sigma) and J-H<0.5.*

Searches for nearest and faintest brown dwarfs and white dwarfs.

Additional GPS comment:

U8 asks the archive to return star counts in 10 arcmin grid cells in the GPS. Stellar density does not require extra measurement but it would be worth calculating for each array field and/or for each square arcminute and putting the result in the archive. Similarly, a cluster detection algorithm should be run in regions of high stellar density, e.g. fitting a King model. The GPS is likely to find a lot of new clusters.

In summary the GPS wants to be able to get from the archive:

i) extinction (A(J), A(H), A(K)) for 3 colour detections

ii) solution(s) for true colours, dereddened apparent magnitude, sp. type, luminosity class and distance modulus, with probabilities and confidence attached.

iii) automatic update when multi-epoch data arrives - faint sources with proper motion being confirmed as dwarfs rather than giants.

Discussion:

The generalised querying enabled with SQL makes possible much of GPS 1-6 (for example, new columns of data can be generated from existing columns given a mathematically-expressed algorithm). For the purposes of the WSA, we undertake to work with the UKIDSS Working Groups to apply any supplied algorithm to catalogue tables to store required quantities. Ultimately, we aspire to making available an automatic system of upload of user-supplied codes/algorithms to enable general processing of catalogues and/or images.

Galactic Clusters Survey (GCS):

Following U3:

a) select all candidate cluster members which are within X arcmins of a brighter candidate cluster member to search for the incidence of brown dwarfs as wide companions to higher mass objects.

e.g. I would see if there is a significant number of these, then perhaps use them to test the brown dwarf ejection/formation scenario - one could then examine the 'primaries' to see if they were indeed binaries as this theory would indicate.

One might want to append something like:

b) For all candidate members derive offset stars/wavefront sensor stars for X spectrograph on Y telescope. Make finding charts for each object."

Discussion:

Again, we note the flexibility of SQL, which would enable part (a), and also given a list and a set of selection criteria for brightness and proximity of nearby guide stars, part (b).

Deep Extragalactic Survey (DXS):

DXS 1: *"I have a catalogue..."*

XMM, SWIRE, GALEX etc will all want to input a list of positions and spit out an id so we need a query tool that reads in:

position, 1sigma error radius and N for the number of 1sigma radii you want to search to

and returns:

UKIDSS source, offset, N 1sigma radii.

This catalogue importer MUST be flexible and allow a simple ascii table to be read in.

DXS 2: *"Find QSOs..."*

I suspect that most of the DXS interest will be in galaxies but we should be able to select on:

"FWHM"=stellar/unresolved .and. J-K>1.0

DXS 3: *"Find EROs..."*

The vast majority of the DXS work will be on the union of optical and UKIDSS data. This strikes me as the most difficult aspect to include in the archive. We must be able to import optical data into the system with a similar structure and then interogate them simultaneously to say:

R<26 .and. K<20.5 .and. non-stellar .and. R-K>5

Perhaps a folding into the UKIDSS data of the optical photometry is required (and vice-versa) where each UKIDSS source has a set of comparison magnitudes (2MASS, SDSS, CFHLS, VST etc) that we can query globally and are skipped if they aren't entered.

DXS 4: *"Find clusters...."*

I suspect that this is by necessity an off-line process but will require the preselection of all the galaxies in a given area with some magnitude and colour cuts:

In a rectangular region X"xY" return:

non-stellar .and. K<20.5 .and. J-K>1.1

The important factor will be writing out such a vast query onto disk. The logistics of extracting large parts of the database is something that needs thought through now.

DXS 5: *"Map me the source density..."*

The current queries don't allow for the ability to select objects then display their surface density as a greyscale. This would be a really useful tool for looking for clusters or groups of similar objects but also for checking the completeness across large areas (i.e. take the faintest objects and see whether they are evenly distributed).

So the query would look like:

select non-stellar K>20.5 map 5x5deg with 10' pixels

The majority of the discussion about optical data is wrt to SDSS which is not appropriate to the DXS. We will need to bring the deeper optical data in as both pixel and catalogue form (ideally) so including this in the full design is vital in my view.

Discussion:

Clearly the issue of import of catalogue (and ideally pixel) data is open-ended. At this stage, in the design of the WSA we will undertake to ensure sufficient space is available for import of UKIDSS-supplied complementary imaging catalogues given sufficient information concerning their size; if this information is unavailable then this will be done on a best-efforts basis.

Ultra-Deep Survey (UDS):

No further usage examples were suggested.

Discussion:

Some discussion centred around the availability of space (and ultimately tools) for catalogue import; as stated above, WSA design will allow for this.

Miscellaneous:

1: Ability to semi-automatically prepare an ORAC-OT program based on the selection of objects from the survey.

2: Facility for ORAC load up of archive data at UKIRT while doing follow-up work would be useful.

Discussion:

WSA output formats will be standard, and remote server functionality will be provided to enable 2.
