The VISTA Science Archive

N.J.G. Cross[1], R.S. Collins[1], R.G. Mann[1], M.A. Read[1], E.T.W. Sutorius[1], R.P. Blake[1], M. Holliman[1], N.C. Hambly[1], J.P. Emerson[2], A. Lawrence[1], and K.T. Noddle[1]

[1] Wide-Field Astronomy Unit, Institute for Astronomy, School of Physics and Astronomy, University of Edinburgh, Royal Observatory, Blackford Hill, Edinburgh EH9 3HJ
[2] Astronomy Unit, School of Physics & Astronomy, Queen Mary University of London, Mile End Road, London E1 4NS

Accepted XXXX. Received XXXX ; in original form XXX

ABSTRACT

We describe the VISTA Science Archive (VSA) and its first public release of data from five of the six VISTA Public Surveys. The VSA exists to support the VISTA Surveys through their lifecycle: the VISTA Public Survey consortia can use it during their quality control assessment of survey data products before submission to the ESO Science Archive Facility (ESO SAF); it supports their exploitation of survey data prior to its publication through the ESO SAF; and, subsequently, it provides the wider community with survey science exploitation tools that complement the data product repository functionality of the ESO SAF.
This paper has been written in conjunction with the first public release of public survey data through the VSA and is designed to help its users understand the data products available and how the functionality of the VSA supports their varied science goals. We describe the design of the database and outline the database-driven curation processes that take data from nightly pipeline-processed and calibrated FITS files to create science-ready survey datasets. Much of this design, and the codebase implementing it, derives from our earlier WFCAM Science Archive (WSA), so this paper concentrates on the VISTA-specific aspects and on improvements made to the system in the light of experience gained in operating the WSA.

Key words. astronomical databases – surveys: infrared – stars: general – galaxies: general – cosmology: observations

1. Introduction

One of the clearest trends in observational astronomy over the past two decades has been the rise in the importance of systematic sky surveys. When coupled with good archives, sky surveys provide homogeneous, re-usable data products facilitating a range of research programmes and, in particular, enabling the large-scale statistical analyses required for many of the most important science goals in modern astronomy.

Probably the most prominent of this generation of surveys has been the Sloan Digital Sky Survey (SDSS; York et al. 2000) and its SkyServer archive system (Szalay et al. 2002) demonstrated the power of a survey archive based on a Relational Data Base Management System (RDBMS). Such systems can offer astronomers the ability to pose powerful analytical queries in Structured Query Language (SQL) against seamless, survey-wide source catalogues, thereby enabling survey science that would be impossibly cumbersome for an astronomer provided with nothing more than a repository of matching image and catalogue files for each of the thousands of pointings making up the survey dataset.

The success of the SDSS SkyServer was a strong influence on the design of the archive component of the VISTA Data Flow System (VDFS; Emerson et al. 2004). VISTA, the Visible and Infrared Survey Telescope for Astronomy, is described by Emerson et al. (2006). VDFS was designed as a two-phase project, with an initial goal of supporting near-infrared surveys to be conducted with the Wide Field CAMera (WFCAM; Casali et al. 2007) on the UK Infrared Telescope (UKIRT) and an ultimate objective of supporting surveys with VISTA. Within VDFS, the Cambridge Astronomy Survey Unit (CASU) run a night-by-night data processing pipeline, with the Wide-Field Astronomy Unit (WFAU) in Edinburgh generating further data products and providing science archive facilities.

The first-generation VDFS archive – the WFCAM Science Archive (WSA) – is described by Hambly et al. (2008) and serves catalogue and image data from the UK Infrared Deep Sky Survey (UKIDSS; Lawrence et al. 2007), as well as other, P.I.-mode, data taken with WFCAM. More than 1000 users are registered for authenticated access to proprietary UKIDSS data through the WSA, and it supports anonymous access by a larger community once the data are public. The phased approach adopted within the VDFS ensured that the design and development of the WSA progressed with scalability to the larger data volumes of VISTA kept explicitly in mind, along with the likely scientific usage patterns of the VISTA surveys. For example, Cross et al. (2009) describe the enhancements made to the WSA database schema to support time-series analysis of multi-epoch data, which was prototyped using observations from the UKIDSS Deep Extragalactic Survey, but motivated by the requirements for supporting variability analyses with VISTA.

The initial scientific programme for VISTA is mostly focussed on six ESO Public Surveys which deliver reduced images and derived catalogue data products to the ESO Science Archive Facility (SAF). Five of the six Public Survey consortia (see § 2), use the VDFS for the generation of these data products and employ the VISTA Science Archive (VSA) to manage their data, both for quality assurance analysis and preliminary exploitation prior to submission to the ESO SAF and, following its publication there, to provide the wider community with sophisticated science archive capabilities that complement the data product repository functionality of the ESO SAF.

This paper describes the VSA and its first public release of data from the five VDFS-supported Public Surveys. In Section 2

we discuss the VISTA telescope and the Public Surveys and compare to UKIRT-WFCAM, focussing on the essential differences that affect the VSA. In Section 3, we provide an overview of the VSA, discussing the table structure before we compare the VSA to the WSA in Section 4. Since the design of the VSA is derived so directly from that of the well-used WSA, we do not repeat the basic design background material from Hambly et al. (2008), but refer readers unfamiliar with the WSA to that paper and to Cross et al. (2009), as well as to the comprehensive online documentation provided on the VSA website[1]. We discuss changes to the image metadata, the catalogue parameters and the infrastructure in Sections 5–7 in the VSA compared to the WSA and new features that are common to both in Section 8. Section 9 provides examples of some of the different types of science queries that the VSA supports, while Section 10 provides details of the contents of the first VSA releases of the five VDFS-supported VISTA Public Surveys. We summarise this paper and discuss future work in Section 11, while several appendices provide technical details supplementing the main body of the paper.

2. Overview of VISTA and its Public Surveys

The Visible and Infrared Survey Telescope for Astronomy (VISTA; Emerson et al. 2006) is currently the fastest near-infrared survey telescope, with an étendue (area times instantaneous field-of-view) equal to $7.5\mathrm{m}^2\mathrm{deg}^2$. It has a 4m f/1 primary mirror, and a 1.2m secondary giving it a 1.65 degree diameter field-of-view (see Emerson & Sutherland 2010a,b). The VISTA Infra Red CAMera (VIRCAM; Dalton et al. 2010), has 16 2048×2048 pixel non-buttable Raytheon VIRGO HgCdTe detectors and has a quantum efficiency $> 80\%$ between 0.9μm and 2.4μm. The pixel scale is $0.34''$ and the instantaneously sampled field-of-view is 0.6 sq. deg (see Fig. 1). Compared to its nearest counterpart, the United Kingdom Infra Red Telescope with its Wide Field CAMera (UKIRT-WFCAM; Casali et al. 2007), the survey speed of VISTA is $\sim$ 6 times faster, having twice the sensitivity – increased throughput for a similar sized telescope – and 3 times the area per pointing.

ESO's Science Verification for VISTA started at the end of 2009 and the main science programme commenced in early 2010. VISTA's programme initially focuses on six ESO Public Surveys[2], nicely complementing the sub-surveys of UKIDSS in the northern sky.

These six surveys are:

- VHS: the VISTA Hemisphere Survey[3];
- VVV: the VISTA Variables in Via Lactea[4] (Saito et al. 2012);
- VMC: the VISTA Magellanic Cloud survey[5] (Cioni et al. 2011);
- VIKING: the VISTA Kilo-degree INfrared survey for Galaxies[6] (Findlay et al. 2012);
- VIDEO: the VISTA Deep Extragalactic Observations survey[7] (Bonfield et al. 2010);
- UltraVISTA[8] (McCracken et al. 2012, in preparation).

These surveys have a 'wedding cake' arrangement of galactic/extragalactic surveys (VHS, VIKING, VIDEO, UltraVISTA) with different depth/area combinations and two specialised stellar astronomy programmes (VVV, VMC), much like the UKIDSS surveys. The five surveys supported by the VDFS are VHS, VVV, VMC, VIKING and VIDEO. UltraVISTA makes some use of the CASU pipeline products, but is not currently archiving its data in the VSA.

VISTA data is calibrated on the natural VISTA photometric system (see Hodgkin et al. 2012, in preparation). All magnitudes (unless designated as AB mag) are on this Vega mag system.

2.1. Differences in the telescope and instrument VISTA & WFCAM

For detailed descriptions of VISTA and WFCAM, see Emerson et al. (2006) and Casali et al. (2007) respectively. In this section we will just discuss the salient differences which affect the VSA design compared to the WSA.

2.1.1. The VISTA focal plane: pawprints and tiles

VISTA is significantly different from UKIRT/WFCAM in several important aspects, which affect image processing and subsequent archive operations. The main differences of significance to VSA are the arrangement of the focal plane and the ability VISTA's alt-azimuth mount provides to observe the same piece of sky in any orientation with respect to the focal plane.

VISTA has 16 2k×2k Raytheon VIRGO detectors arranged in a *pawprint* with detectors spaced 90% (10.4′) of a detector apart in the X-direction and 42.5% (4.9′) apart in the Y-direction (see Fig. 1) whereas WFCAM has 4 2k×2k Hawaii 2 detectors arranged in a *pawprint* (see Fig. 2) with the same spacing of 94% (12.8′) in each direction.

The VISTA basic filled survey area is a *tile* made up of six pawprints, three in the Y-direction separated by 0.475 of a detector, and two in the X-direction separated by 0.95 of a detector. Except at two *strips* with just a single exposure, this tile has between twice and six times the exposure time of each pawprint at every pixel with a mode of two exposures, (see Fig. 3). There are two possible ways to achieve a required uniform minimum (two pawprints contributing) depth across a multi-tile survey. The most efficient in terms of observing time is for successive tiles to overlap the *strips* at top and bottom, coadding the data from the two separate tiles, and each will have an area of 1.636 sq.deg covered at least twice. However reaching the full depth by coadding these two *strips* can be complicated by varying sky conditions if the adjacent tiles are not observed under the same conditions (e.g. different PSFs and sky conditions). Indeed the same two effects can be a problem in making a tile from six pawprints (depending on how rapidly the seeing varies between pawprints). The other less efficient, but simpler, way to achieve constant depth across a multi-tile survey is to butt together regions of tiles that have reached the minimum double exposure, ignoring the singly exposed *strips* resulting in an area of 1.501 sq.deg covered at least twice.

In the case of WFCAM four pawprints are required to make a filled tile, and everything gets a single exposure and there are no edge *strips*, so large contiguous regions of the sky can be surveyed by simply overlapping subsequent pawprints as shown in Fig. 2. We note that in common with WFCAM, 'microstepping' is possible with VISTA, but its use is not recommended by ESO and the VISTA Public Surveys do not use the technique.

[1] http://surveys.roe.ac.uk/vsa

[2] http://www.eso.org/public/teles-instr/surveytelescopes/vista/surveys.html

[3] http://www.ast.cam.ac.uk/~rgm/vhs/

[4] http://vvvsurvey.org

[5] http://star.herts.ac.uk/~mcioni/vmc/

[6] http://www.maths.qmul.ac.uk/~wjs/VIKING

[7] http://star-www.herts.ac.uk/~mjarvis/video/

[8] http://www.strw.leidenuniv.nl/~ultravista/

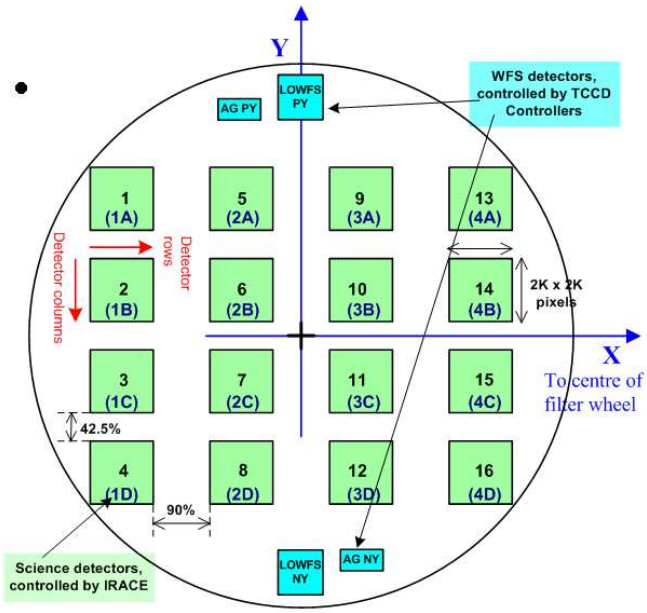

Fig. 1. The VISTA focal plane showing 16 2k × 2k detectors with 90% spacing in the x-direction and 42.5% in the y-direction. There are also two auto-guider (AG) and two low-order wave-front sensor (LOWFS) detectors.

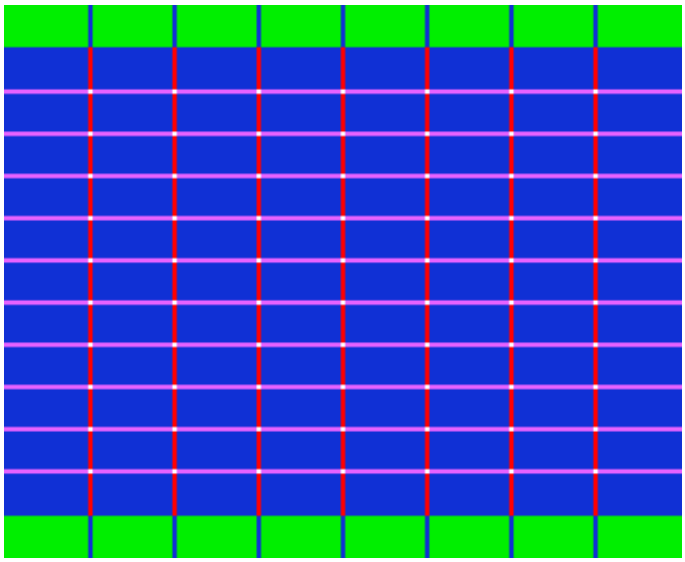

Fig. 3. An exposure map of a VISTA tile. The green *strips* at top & bottom have a single exposure. The majority of the area (blue) has two exposures, the pink has 3 exposures, the red 4 and the white 6. The doubly or more exposed area is 1.501sq.deg. The singly exposed green *strips* at top & bottom of the plot are each 1.475 deg×0.092deg=0.135sq. deg and can be overlapped by corresponding areas from adjacent tiles for many surveys. Assigning one of the two 0.092 deg overlap (top & bottom) to each of the adjacent tiles involved in an overlap, means that each tile, when part of a filled larger area, covers (1.017+0.092)×1.475=1.636sq. deg at least twice.

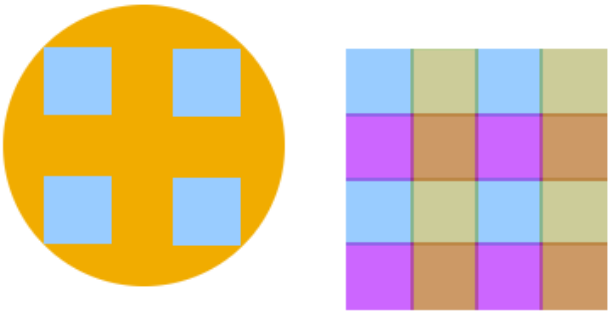

Fig. 2. Areas of sky can be efficiently surveyed by arranging 4 WFCAM pawprints (left) into the arrangement on the right. Each colour in the right hand image represents a different pawprint. There is a small amount of overlap at each edge.

2.1.2. Merging VISTA pawprints into tiles

The processing of raw pawprints frames into calibrated images is done by the VDFS pipeline at CASU on a nightly basis[9] and includes combining the pawprints into tiles. These pawprints and tiles are ingested into the VSA without any additional image processing. If any pointing is observed only once in a given filter in the survey design, the nightly pipeline–processed catalogue product is used to generate the merged `Source` (see § 3.1) table. For deeper surveys, however, stacking tiles observed multiple times typically involves observations over several different nights, and multi-night products are not the responsibility of CASU. WFAU creates tiles from multiple nights of data by first stacking the separate component pawprints and then combining the six stacked pawprints into a tile. Stacking the tiles rather than the component pawprints is very firmly not recommended.

While tiling pawprints together to form a tile is quite standard when working with visible images (Fruchter & Hook 2002), ground-based infrared sky subtraction is difficult because the sky is so much brighter than in the optical and dominates the flux of most objects. Furthermore, the VISTA camera's distortion across the field-of-view and the larger variation in both the sky and the point-spread function (PSF) within the duration of the observations needed to create a single tile make it necessary to do additional processing (Lewis et al. 2010). Tiles are processed using the following procedures[10]:

- Stack all components of each pawprint
- Extract the catalogue from each pawprint
- Recalculate the world coordinate system (WCS) of each pawprint, by comparing stars in the catalogue to the Two Micron All Sky Survey (2MASS) catalogue (Skrutskie et al. 2006)
- Calculate the photometric zero-points (VISTA system) of each detector in each pawprint and update the headers
- Filter each pawprint to remove large scale variations (> 30″ in the background), see (Irwin 2010) for details
- Mosaic all the pawprints (to produce a tile with large scale features present)
- Mosaic all the filtered pawprints (To produce a tile without large-scale background features)
- Extract the catalogue from the filtered tile
- Recalculate the WCS of the tile

[9] http://casu.ast.cam.ac.uk/surveys-projects/vista/technical

[10] http://casu.ast.cam.ac.uk/surveys-projects/vista/technical/tiles gives more details of the algorithms

– Calculate the photometric zero-point (VISTA system) of the tile catalogue
– 'Grout' the tile catalogue to find the correct PSF in each region of the tile and the correct offset for the Modified Julian Date. The grouting procedure takes information from the original pawprint catalogues and confidence images to compute differential aperture corrections at the location of each object. The differential aperture correction is the difference between the pawprint aperture correction and the median tile aperture correction. Aperture fluxes 1 - 7 are corrected for these differences.
– Reclassify the stars and galaxies in the tile catalogue
– Recalculate the photometric zero-point (VISTA system) of the tile catalogue
– Remove any temporary files, such as the filtered pawprints. Only the unfiltered tile is stored in the archive

Unsurprisingly the filtering also makes it impossible to produce accurate catalogue values for large extended sources, e.g. nearby galaxies or Galactic nebulae. The filtering scale used for the main VISTA Public Surveys removes structure on scales larger than 30″. This is a similar scale to the local background scale length (22″) which would in any case limit the accuracy of any photometry of larger objects.

2.1.3. Active Optics

The quality of VISTA images is maintained, as it observes at different elevations and its temperature changes, by updating the position of the secondary mirror, with corrections derived from look up tables and the low order wave front sensors. If a correction has not been applied recently enough, or it is bad for some reason, the image quality can be degraded. Some images show evidence for such problems, contributing, along with seeing variations between pawprints, and PSF variation across the field when in perfect alignment, to a greater variety of PSFs than in the case of WFCAM.

2.1.4. Telescope mounts

VISTA is an alt-azimuth mounted telescope, whereas UKIRT has an equatorial mount, so the focal plane in WFCAM remains in the same equatorial orientation, but the VISTA focal plane must rotate with respect to the telescope to keep the same orientation on the sky during an exposure. Since the focal plane can rotate, orientation is an additional degree of freedom. Different programmes can choose the orientation that best accommodates their survey design. Given the complex processing of image and catalogue data (see previous subsection), stacking two images of very different orientation was considered inadvisable. Thus we group data based on orientation as well as position/filter when stacking and tiling and to do this we have added an extra column, **posAngle** to the `RequiredStack` (see § 3.1) table. This is the orientation of the image x-axis to the N-S line. This means that if images in the same programme lie in the same position and filter but have very different orientations they will be processed separately. As a default, we have been using a tolerance of 15 deg, but it can be set programme by programme. In practice this situation has only occurred once in the Public Surveys: the VVV team had a small amount of data in the Science Verification stage which was orientated along equatorial RA axis, and later had data in the same region of sky which was orientated along the Galactic longitude axis, and so the tiles are 60 deg different in orientation.

3. Overview of the VSA

3.1. The basic structure of the VSA

WFAU receive VISTA data as multi-extension FITS files (MEFs; Pence et al. 2010): primarily these come from the CASU pipeline, which processes the data by Observing Block (OB), but WFAU can also handle 'external' images made outside VDFS, such as those produced by the survey consortia, as is currently done with deep VIDEO mosaics. The MEFs contain either images or catalogues of objects extracted from them: in VSA parlance the content of an MEF image file is a *multiframe* (consisting of images - frames - for all detectors) and a *detection* is an object extracted from a single image in a single filter. The metadata headers from these files are loaded into tables within an RDBMS, as are the catalogues.

Depending on the survey, further data products may be generated in a database-driven manner. For example, repeated observations of the same field in the same filter may be stacked to create deeper images, from which catalogues are then extracted, while information from single-filter catalogues is merged to create catalogues of *sources*, which are objects described by attributes in several filters. The metadata from the additional image data products are ingested into database tables, as are derived catalogues, while the images themselves are stored in FITS format on disk. Further information may then be derived from the database tables and stored in new tables: e.g. variability information may be derived from multi-epoch data, following the synoptic data model of Cross et al. (2009). The VSA comprises, therefore, a set of tables within an RDBMS, a collection of FITS files stored on disk and the interfaces that allow users to access these data.

Whole image files may be selected for download from the archive through web forms or via SQL queries on the image metadata tables, while image cut-outs may be created from these using a different web form. Other web forms exist to provide a basic level of access to the catalogue data, but the real power of the VSA comes from the ability to query RDBMS tables using SQL. To do this requires knowledge of the VSA database schema. The online VSA schema browser[11] provides detailed descriptions of every column in the hundreds of database tables in the VSA, but we summarise the five main classes of table in the remainder of this section.

Throughout this paper, we use a fixed-width font to refer to VSA database tables: e.g. `Multiframe` or `vvvSource`. Many of the tables are set up for individual programmes (e.g. an individual Public Survey), such as `vvvSource` for the VVV, and `videoSource` for VIDEO. When we are discussing generic properties of "Source" tables, we will abbreviate them as `Source`, rather than using programme-specific names. Individual columns within tables are referred to using a bold font: e.g. **multiframeID** or **aperMag3**.

3.1.1. Metadata tables

The following tables record metadata about images:

– `Multiframe`: This contains the main primary header keywords from the VISTA images and some additional derived quantities that are calculated for each multi-extension image.
– `MultiframeDetector`: This contains the main extension header keywords from the VISTA images except for astrom-

[11] http://surveys.roe.ac.uk/vsa/www/vsa_browser.html

etry related keywords and some additional derived quantities that are calculated for each detector.
– CurrentAstrometry: This contains the astrometric related extension keywords and some additional derived astrometric quantities.
– MultiframeEsoKeys: This contains subsidiary primary header keywords that are stored in the hierarchical ESO format (HIERARCH ESO).
– MultiframeDetectorEsoKeys: This contains subsidiary extension header keywords that are stored in the hierarchical ESO format.
– AstrometricInfo: This includes additional derived properties, for OB frames used in multi-epoch surveys.

3.1.2. Catalogue data tables

The following are the catalogue data tables used in the VSA. There is a different table for each programme, so they will each start with the programme acronym:

– Detection: This contains the extracted sources for each detection frame in MultiframeDetector (individual frame in a multiframe): the raw extraction attributes from the original FITS table, the calibrated positions and magnitudes and a few other derived quantities.
– Source: This is a merged filter catalogue from the deepest images in each pointing, and is made "seamless" (Hambly et al. 2008) to allow the user to find the most complete set of unique sources in the programme.
– SynopticSource: This is a merged filter catalogue made from detections in contemporaneous images. This is useful if colours of variable stars are needed. Only those programmes designed to have contemporaneous colours will have a SynopticSource table.
– Variability: This contains statistics for the light-curves of sources in multi-epoch programmes, allowing selection of variables based on different statistical quantities. VarFrameSetInfo is a useful supporting table for Variability and includes the fitted noise functions for each pointing.

3.1.3. Linking tables

The following are tables that link different types of catalogue data or metadata:

– MergeLog: For each pointing this lists the image frames in each filter from which the extracted detections were merged together to from the sources in the Source table.
– SynopticMergeLog: For each pointing at each epoch, this lists the image frames in each filter from which the extracted detections were merged together to from the sources in the SynopticSource table.
– BestMatch tables: These link the sources in the Source table to each epoch detection in multi-epoch surveys, to match epochs for light-curves. There are SourceXSynopticSourceBestMatch tables for "contemporaneous" filter data and SourceXDetectionBestMatch tables otherwise.
– Neighbour tables: These are simple tables where all sources in the master table are matched to all in the slave table within a specified radius. These can used for multiple purposes, such as to link with external surveys, e.g. vhsSourceXDR7PhotoObjAll links the VHS Source table to the Sloan Digital Sky Survey Data Release 7 PhotoObjAll table.
– TilePawPrints: This links tile image detections to the detections from pawprint images that make up the tile.
– Provenance: This links image frames to their components, e.g. a deep stack to each epoch stack frame that went into it, or an epoch stack frame to the raw images.
– ProgrammeFrame: This links image data to a programme and the programme requirements and is very important for programme curation. The same frame could be used in multiple programmes, for instance different PI programmes with the same PI in different semesters or an all hemisphere release containing data from VHS, VVV, VIKING and VMC.

3.1.4. External catalogues

The scientific goals of surveys tend to require external data (e.g. from surveys on other telescopes / instruments at different parts of the electromagnetic spectrum), in addition to data from VISTA itself. To support those analyses, the VSA contains copies of catalogues from a number of external surveys, which are listed in the online schema browser[12]. The list of these is updated in response to requests from the survey consortia, and new cross-match neighbour tables are added for different programmes, as these external surveys become available. The online documentation explains how these cross-neighbour tables can be used to perform effective cross-catalogue queries.

3.1.5. Curation tables

As mentioned above, the operations of the VSA are database-driven once the original MEFs have been ingested, with processing steps and data product provenance recorded automatically in the database. The VSA contains, therefore, a large number of tables that drive, and are derived from, these curation tasks. Many of these are only of relevance only to the VSA operations team, the following list do contain some pertinent information for users of the VSA:

– Programme: Basic programme information. This includes the programme dependent information used to create the SQL schema which drives most curation tasks.
– RequiredTile: The current expected tile product pointings and matching tolerance. In the case of VIDEO, for which we ingest mosaics provided by the survey team, the relevant table is RequiredMosaic.
– RequiredNeighbours: lists which neighbour tables that join surveys have been created and what are the matching radii.
– PreviousMFDZP: The photometric calibration history of each image extension.

4. Differences between WFCAM and VISTA Science Archives

While the design of the WSA was developed with ultimate application to VISTA in mind, there are some differences between the WSA and VSA structures.

[12] http://surveys.roe.ac.uk/vsa/www/vsa_browser.html

4.1. Tile and pawprint information in the VSA

In the VSA, we store catalogues from both the pawprints and tiles in the detection tables for each survey (e.g. `vhsDetection` for the VHS survey). The tile catalogues are needed to produce uniform catalogues to the full depth of each survey. However, the astrometric solution in tile catalogues is not quite as good as that in pawprint catalogues because the distortion is not as well represented by the tangent plane (TAN) projection which tiles are projected onto, as it is in the zenithal polynomial projection (ZPN) that can be used for the pawprints, (see Calabretta & Greisen 2002). Saturated stars also have better photometry in the pawprint catalogues which are not filtered. Producing pawprint catalogues does not add any additional overhead, since they must be produced as part of the production of tiles to allow the pawprints to be aligned correctly before mosaicking.

Having both tiles and pawprints has created the need for multiple layers of products and more complicated archive curation infrastructure (see § 7) to keep track of these and allow them to be used together. It also means that stack requirements need an additional constraint, the offset position. Each stack that goes into a tile has a different offset position, 0-5, which is a function of the difference (in arcseconds) of the centre (optical axis) of the pawprint from the centre of the tile. These offsets are stored as **offsetX**, **offsetY** in the `Multiframe` table. The offset position is not the same as the **offsetID** in `Multiframe`, which is simply the order that the offset was observed in and may differ in relative position on the tile from one epoch to the next (i.e. the order in which the pawprints are executed can be chosen by the observer but the relative positions of the 6 pawprints are fixed in the OBs). The **offsetPos** always refers to the same part of the tile. There can be considerable overlap between two pawprints, from different parts of two different tiles, but they will not be stacked together.

4.1.1. The Tile-PawPrint matching tables

To link the tile and pawprint catalogues together, we have created two extra tables: `TileSet` and `TilePawPrints`, which match each detection in a tile catalogue with detections in the same position in the pawprint catalogues. These tables are survey specific, so VHS, which has detections in `vhsDetection` will have tables `vhsTileSet` and `vhsTilePawPrints` to link the tile and pawprint catalogues. `TileSet` and `TilePawPrints` are designed along the lines of `SynopticMergeLog` and `SynopticSource` (Cross et al. 2009): `TileSet` links the frames together using the multiframe identifiers for the tile and pawprints, and `TilePawPrints` links the detections using the extension numbers (i.e. detector number) and sequence numbers (i.e. order that object was extracted in the frame). Table `TilePawPrints` is deliberately as narrow as possible, and simply includes the necessary linking information, with no additional attributes such as magnitudes, since it is expected that it will always be used to link, and could be used with a whole variety of attributes. By not including magnitudes, we also reduce the number of updates that are needed when recalibrating the photometry. We have put some examples of linking tile detections to pawprint detections in the VSA SQL cookbook[13].

In Fig 4 we show an entity-relationship model (ERM) for these new tables, showing how they are related to the current `Multiframe`, `MultiframeDetector` and `Detection` tables. ERMs show the relationship between entities (a collection of related data, such as metadata data in the `Multiframe` table) in the database by a series of solid and dashed lines representing one-to-one, one-to-many or many-to-many matches between the entities. A one-to-one match is a single line, a one-to-many match is a single line at one end and a "crow's foot" of 3 branching lines at the other end. A solid line means that there must be a match for every entry in the table, and a dotted line means there are matches for some entries. ERMs are discussed in more detail in Hambly et al. (2008) § 2.4 which also shows the fundamental relationships between entities in the VSA and WSA. ERMs are also shown for entities in the multi-epoch data model in Cross et al. (2009) § 2.

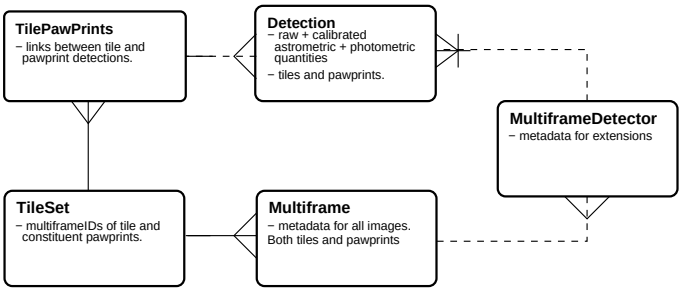

Fig. 4. The ERM for the tile-pawprint linking tables. `TileSet` links the 7 multiframes (1 tile and 6 pawprints). Each row in `TileSet` links to many associated rows in the `TilePawPrints` table, each of which refers to a different object. Each of these objects has between 1 and 7 detections. The complication comes from `TileSet` containing a mix of tile and pawprint frames which cover different areas on the sky and have different extensions in `MultiframeDetector`.

Tile and pawprint detections are matched within a fixed radius of 0.8″ in both VISTA and WFCAM, which is approximately the average seeing and is many times the typical astrometric error but less than the separation between neighbouring objects, so objects will only be matched to the same object on different images, not to neighbouring objects. The matching algorithm is the same as for sources in the `Source` table, see Hambly et al. (2008). Like `MergeLog`, `TileSet` includes all the associated frames as a frame set consisting of the tile frame (**tlmfID**) and the 6 pawprint frames (**o1mfID** – **o6mfID**) where **o1** is the pawprint with **offsetID**=1. `TilePawPrints` then contains the matched detections between the 7 frames, just like `Source` would have the matched detections between all the different filters. These tables should be used as linking tables to compare the tile and pawprint or even pawprint to pawprint between offsets.

There is an important difference in the way that frame sets are matched that causes some peculiar features to exist in the tables. Instead of merging *Z*, *Y*, *J*, *H* and *Ks* frames in a particular pointing, we merge the tile frame and the constituent pawprints. In the `SynopticMergeLog` frame sets are made up exclusively of tiles or pawprints matched on a detector by detector basis. In either case, each frame in the frameSet is similar to each other and the matching condition is a combination of multiframe identity and extension number. In the case of `TileSet`, frame sets are matched by multiframe identity only - all pawprint extensions are related to the same tile. This makes the assignment of default rows more complicated. Non-detections in a particular filter have a default entry in the `SynopticSource` table: if there is no detection in the J-band then **jSeqNum** is simply set to the standard integer default value (-99999999).

Frame sets in `TileSet` are made up of a tile and six pawprints; the tile is composed of and overlaps all 16 detectors in each pawprint. Thus frame sets are not matched by exten-

[13] http://surveys.roe.ac.uk/vsa/sqlcookbook.html#TilePawprints

sion, but by whole multiframe so it is not possible to have entries in the tile set identified by **multiframeID** and **extNum**, they must be identified by **multiframeID** only, and the links in `TilePawPrints` must include both the extension number and sequence number. If there is no detection in a particular frame, such as pawprint offset 1, we do not know off-hand which extension the detection should have been on (if any, since it may be in a gap between the detectors for this offset). We should set the default row for this missing detection as **o1ExtNum** = −9999, **o1SeqNum** = −99999999. However, there is no equivalent row in the `Detection` table because the foreign key constraint between the `Detection` table and `MultiframeDetector` forbids this, since there are not rows in `MultiframeDetector` with a good **multiframeID**, but a default **extNum**. Instead we set default rows in `TilePawPrints` to have extension numbers equal to 2 and default sequence numbers, so that they can match with default rows in the `Detection`. This allows us to match the tile detections in `Source` or `SynopticSource` and find all the pawprint detections with the correct numbers of rows returned, with default attributes in all the cases where there were no detections. We must emphasize that an entry in `TilePawPrints`, which has a key (**o1ExtNum**=2, **o1SeqNum**=-99999999) does **not** denote that there is overlap with extension 2. This just means that the object was not detected in the multiframe in question.

Most `TilePawPrints` rows will be entries where the tile and two pawprints have matched detections, some where the tile and just one (in the outer *strips*) or three four, five or (infrequently) six pawprints have detections, some where the tile only has detections (usually at the faint end). If there is no detection from a particular frame, a default entry is added: the extension number **extNum** and sequence number **seqNum** will both be default. Careful selection of what is default and what is not will optimise the use of these tables. Any attributes in the detection tables can be compared in this way, although it is necessary to match a new detection table for every frame in the table. Since this is done via the primary key (multiframeID, extNum, seqNum), the joined SQL queries are very efficient.

To match tile detections in a `Source` table or `SynopticSource` table to the detections on the constituent pawprints, it is necessary to remove the pawprint-only detections to leave a table with only the good tile detections and necessary defaults. If a query retains the pawprint-only detections, they are interpreted as defaults, so if a detection is missing in a particular filter it will be matched to every set of pawprint detections that are not linked to a tile-detection, a nonsensical result. To avoid this, we have created a view `TilePawTDOnly` that can be directly matched in the same way as `TilePawPrints`. However, for very large datasets, such as the VVV, queries work better if users use `vvvTilePawPrints` and add the necessary constraints into the where clause, see example queries found in the SQL cookbook. The ERM for the matching of the `Source` table to the pawprint detections is shown in Fig 5.

In the case of the VHS, which has a single epoch only, we have provided an additional table `vhsSourceXPawPrints` which is a neighbour table between `vhsSource` and the pawprint detections in `vhsDetection`. This simply matches all pawprint detections within a given radius to a source. A typical query is shown at the bottom of the tile-pawprints section of the SQL Cookbook. It is more difficult to do precision queries on particular offsets or extensions as with `TilePawPrints`, but it is possible to do faster queries and may be preferable if only the pawprint data are required.

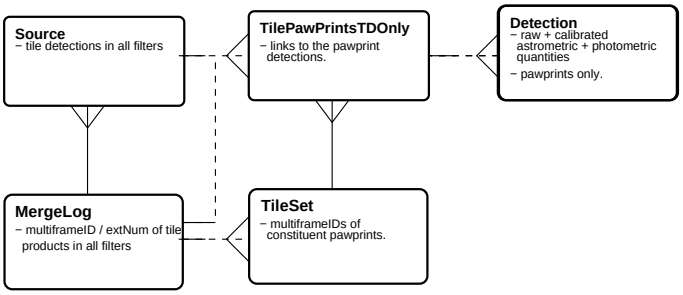

Fig. 5. The ERM for the tile-pawprint linking tables matched with the Source table. Detections in the Source table come from tiles. In each MergeLog row there will be at least one tile with a maximum equal to the number of filters used in the programme. Each of these is matched to a TileSet and each Source row to a similar number of rows in TilePawPrintsTDOnly, depending on whether there was a detection in each observed filter. The relationship between MergeLog and TilePawPrintsTDOnly is necessary because the multiframe extension number is stored in MergeLog and TilePawPrintTDonly.

5. Changes to image metadata

5.1. ESO attributes

VISTA data pass through an ESO quality control pipeline (whose modules are provided by VDFS) in Garching before being ingested into the VDFS data processing pipeline in Cambridge, while the VDFS-generated data products supplied to the ESO SAF must comply with ESO metadata standards. As a consequence, the VISTA data products present contain a quantity of standard ESO information not present in WFCAM data products. For example, the headers of image files contain a number of ESO hierarchical FITS keywords[14]. Those required for data processing, or judged to be scientifically useful, are propagated into the `Multiframe` or `MultiframeDetector` tables for keywords from primary or extension headers, respectively, while the remainder are recorded in `MultiframeEsoKeys` and `MultiframeDetectorEsoKeys` tables.

Initial quality control occurs when the data is checked at the telescope and then at Garching to determine if the data was taken within the required constraints. This results in additional quality control metadata created for VISTA, which is included in the VSA, and which are not found in the WSA. These include OBSTATUS (“Completed”, “Executed”, “Aborted”, “Pending” and “Undefined”), and ESOGRADE (“A”, fully within constraints; “B”, mostly - 90% - within constraints; “C”; “D”; “R”, rejected). If the OBSTATUS is not completed, the whole OB will be repeated later. Each OB is also quality assessed more generally when processed in Cambridge.

There are also requirements for the files that are imported into the ESO Science Archive Facility[15]. We generate the following required keywords for the FITS files that are sent to the ESO archives: ABMAGLIM, ABMAGSAT, MJDEND, TEXPSUM. These have been added into `MultiframeDetector` as **abMagLim** and **abMagSat** and `Multiframe` as **mjdEnd**, **totalExpTimeSum**. Currently the AB saturated magnitude is only calculated when images are released to ESO, so all the values are default in the archive. The image pixel data have also been scaled and converted to 32-bit integer from 32-bit floating point, as a requirement for the ESO archive. The calculation of the scaling parameter is shown in Appendix A.

[14] http://heasarc.gsfc.nasa.gov/fitsio/c/f_user/node28.html
[15] http://archive.eso.org/cms/

5.2. Deprecations

A number of new image deprecation codes have been introduced for the VSA. The different sub-surveys within UKIDSS followed the same deprecation policy, so it was possible to apply that uniformly across the whole WSA and not release any deprecated data. For VISTA, however, the Public Survey teams have defined different deprecation criteria, so it is not possible to define an analogous uniform policy for the VSA. Instead, it has been decided to release all data (deprecated or not) but to define additional deprecation codes to indicate whether or not particular images have been omitted in the creation of higher order products: deep stacks, tiles or mosaics, `Source` tables, `SynopticSource` tables, neighbour tables or multi-epoch/variability tables. The presence of most codes does imply exclusion from further use, but the following codes are more nuanced:

- 50: The frame is good enough for single epoch measurements, but should not be used in a deep stack;
- 51: The frame has a problem with intermittency problem with channel 14 (some early frames had this temporary issue in detector 6). During deep stack creation, channel 14 is set to zero weight in a temporary confidence image.
- 53: These are frames where the quality is marginal. Do not use the frames in deeps, or use the detections in the variability statistics, but do link them in the best match table, so the the survey team can do more tests to determine whether they are good enough to be used in future releases.
- 55: Aborted OB, if the science team decided they want to deprecate based on OBSTATUS;
- 56: Deprecated because of poor ESOGRADE, if the science team decided they wanted to deprecate based on this;
- 58: Deprecated because the catalogue could not be ingested. This happened for many very dense fields in early processing versions, but these have since been replaced. We have included the code in this paper for the sake of completeness and for users who use older team releases that still contain data annotated with this code. If a similar situation arises in the future, more data may be deprecated with this value.

6. Changes to catalogue attributes

Some of the catalogue attributes present in the VSA differ from those in the WSA, for several reasons. Differences in the VISTA and WFCAM detector properties mean that different effects need to be flagged, while the observing strategy changes described in § 2.1.1 led to requirements for different information. We have also introduced additional attributes in response to user demand and in the light of enhancements we have made, especially in the treatment of multi-epoch data.

We describe all these changes in the remainder of this section.

6.1. VIRCAM detector properties

The Raytheon VIRGO detectors used in VIRCAM have no detectable crosstalk and much lower persistence than the Rockwell Hawaii 2 detectors used in WFCAM. For this reason, detections do not have their quality bit-flags set for crosstalk (bit 19 of **ppErrBits**, see Hambly et al. 2008) in the VSA, which reduces processing time. However, the VIRCAM detectors have a narrower dynamic range, with non-linearity and saturation occurring at lower flux levels. Non-linearity is calibrated out as part of the VDFS pipeline at CASU, and WFAU have applied a saturation correction (Irwin 2009), where necessary, to the photometry in the VSA — note that users wishing to extract photometry for objects brighter than m $\sim$ 13 (though this limit is survey dependent according to the DIT used in the OBs) should use pawprint detections rather than tiles.

On the top half of detector 16, the quantum efficiency (QE) varies on short timescales making flat fields inaccurate. This is particularly noticeable at short wavelengths ($\sim$ 1μm e.g. Z and Y). Since tiles are produced from 6 pawprints, each with 16 detectors, we still create a tile even if one or more of the constituent detectors has problems. Each tile pixel comes from up to 6 pawprints, which may have different PSFs (this is why 'grouting' the tiles is done as described earlier). Tile detections that come partly from detector 16 in one or more pawprints are flagged using the post-processing error flag (**ppErrBits**) bit 12, so that users can select a data set without these detections or with them, whichever they prefer. Many of these detections have a low average confidence. We also added a flag for low confidence detections, bit 7.

Occasionally a particular pawprint detector is deprecated for one of several reasons (Hambly et al. 2008). The confidence of a deprecated detector is set to zero when making the tile, and this produces poorly defined extractor values (infinities and not-a-number), which are ingested as defaults into the database. These tile detections are flagged with bit 24.

The two *strips* at the top and bottom of the tile have half the exposure time of the other parts of the tile. We flag these with bit 23. This is partly for the users and partly so that these detections do not become primary sources in the `Source` table of the survey if they overlap with a full exposure region of another tile.

The list below is a summary of the new detection quality bit-flags developed for VISTA tile detections.

- Bit 7, Low average confidence (< 80) in default aperture.
- Bit 12, Source image comes partly from detector 16
- Bit 23, Source lies within a strip of the tile that has half the average exposure of the tile.
- Bit 24, Source lies within an underexposed region due to a missing or deprecated detector

These tile flags have not been applied to catalogues from external mosaics created for VIDEO.

6.2. Changes to attributes in `Detection` tables

The VDFS extractor (Irwin et al. 2004), which generates the raw catalogue parameters for all the VSA data (apart from VIDEO mosaic catalogues, which are extracted using Source Extractor; Bertin & Arnouts 1996) has had a few modifications so that there are slightly different output parameters for VISTA than WFCAM. In the original FITS catalogues produced by the VDFS extractor the *Parent_or_child* column (**deblend** in the WSA `Detection` tables) has been replaced by *average_conf* (stored as the **avConf** in the VSA `Detection` tables), while the Hall radius, Hall flux and Hall flux error have been replace by a half-light radius (**halfRad**) and flux and flux error (**halfFlux**, **halfFluxErr**) within an aperture twice the half-light radius.

6.2.1. Modified Julian Day

In WFCAM, we used the **mjdObs** in `Multiframe` as the time for each observation, which was used in light-curves. This attribute is inadequate in VISTA though, since tiles are made from overlapping pawprints which each have different mean observation times and each tile detection may come from a different

combination of pawprints. This is particularly important for surveys such as the VMC (Cioni et al. 2011), which require a significant fraction of an hour of integration time to reach the required depth at each epoch, but are looking for variables with periods of a few hours. In this case an accurate measurement of the mean observation time is fundamental to the science.

We have now added a new attribute **mjd** into the detection tables. This is the standard Modified Julian Date (MJD, in double precision days since midnight on Nov 17th 1858) and is calculated detection by detection in the case of tiles or extension by extension for pawprints. During 'grouting', the average MJD of each tile detection is calculated as the weighted average (weighted by the average confidence in aperture 3) MJD of the different pawprints that contributed to the tile detection. In the FITS file, this is expressed as a floating point value in minutes from the beginning of the day of the observation, as column *MJDOff*, with the beginning of the day given in the header as *MJD_DAY*. We have also calculated the mean MJD for each pawprint detector and added this to `MultiframeDetector` as **mjdMean**. This is the value that becomes the **mjd** in the detection tables for pawprints, not the **mjdObs**, which is the start time of the observation.

6.2.2. Half-light radii

As well as ingesting detection attributes calculated by the VDFS extractor produced in the FITS catalogue output, other attributes are calculated by the archive curation software. These include several half-light radius measurements, based on the aperture fluxes and the Petrosian flux, using the same method discussed in Smith et al. (2009). These new attributes are:

- **hlCircRadAs**, the half-light circular radius in arcseconds.
- **hlCircRadErrAs**, the error in the half-light circular radius in arcseconds.
- **hlGeoRadAs**, the geometric mean between the half-light radius along the semi-major axis and the half-light radius along the semi-minor axis in arcseconds.
- **hlSMnRadAs**, the half-light radius along the semi-minor axis in arcseconds.
- **hlSMjRadAs**, the half-light radius along the semi-major axis in arcseconds.
- **hlCorSMnRadAs**, the half-light radius along the semi-minor axis corrected for seeing, in arcseconds.
- **hlCorSMjRadAs**, the half-light radius along the semi-major axis corrected for seeing, in arcseconds.

The algorithms used to calculate the above attributes are given in Appendix B.

6.2.3. Magnitude corrections

Photometric calibrations, derived by the VDFS pipeline at CASU (Hodgkin et al. 2009), are applied in the archive curation software. As mentioned in § 6.1, we now make a saturation correction to the pipeline produced magnitudes of stars flagged as potentially saturated. We decided to include explicit columns that detail this and other source dependent corrections (those that are not simply field dependent), making it easy for users to understand and apply the corrections themselves.

The current corrections that are applied to the magnitudes by WFAU[16] in the VSA are:

- **illumCorr**, the illumination or scattered light correction that is calculated for fields on a month by month basis.
- **distortCorr**, the radial distortion correction, which depends on the distance from the optical axis and the filter only.
- **saturCorr**, the saturation correction, that is applied to the 1 arcsecond radius aperture magnitude (**aperMag3**) of bright stars only (those that are flagged as potentially saturated).
- **deltaMag**, the sum of the exposure time correction ($2.5 \log_{10}$ **expTime**), the atmospheric extinction correction ((0.5(**amStart** + **amEnd**) − 1)**extinctionCat**), the illumination correction and the radial distortion correction. The saturation correction is not included, because it is only applied to **aperMag3**. A user can calculated their own magnitudes on the VISTA photometric system for objects by measuring a flux in any way they like and applying the zero-point and the adding **deltaMag**.

The aperture corrections are not included since they are only applied to specific magnitudes and are the same for all objects on one detector. The values for these are included in the `MultiframeDetector` table.

WFAU had several requests for aperture magnitudes without the point-source aperture correction (i.e. for extended sources). Therefore we have included these values for the 7 apertures (1–7) which aperture corrections have been applied as standard. These are named **aperMagNoAperCorr1**, **aperMagNoAperCorr2** to **aperMagNoAperCorr7**. These magnitudes can be used for extended sources if required.

6.3. Changes to attributes in the Source tables

Several of the attributes in the `Detection` table have been propagated through to the `Source` table or `SynopticSource` table. Of the new `Detection` attributes, we have propagated **hlCorSMjRadAs** and the non-aperture corrected aperture magnitudes into the `Source` table. We do not propagate **mjd** though, since the sources in the `Source` table come from the deepest data, stacked across multiple epochs where the time of observation is not particularly useful. Since the `SynopticSource` matches data with a specific epoch, and is particularly useful for variability work on point sources, **mjd** is propagated but **hlCorSMjRadAs** and **aperMagNoAperCorrX** are not.

We still only produce one `Source` table, from the highest product layer (see § 7). Producing one for tiles and one for pawprints breaks the idea of a single master source list (see Cross et al. 2009). Instead, we have a master source list produced from tiles which is linked to pawprints using the `TilePawPrints` table, see § 4.1.1.

7. Infrastructure

There have been several changes to the science archive infrastructure that improve curation of the surveys, but can also be useful for scientists who want to make the best use the VSA. Some of these changes have been incremental and have been documented in Collins et al. (2009); Cross et al. (2009, 2010). The main changes to the VISTA Public Surveys from the UKIDSS Public Surveys are listed below:

- Set up all the requirements, the database schema and curation tables contents using available data and basic programme properties from the `Programme` table. This was also done for the UKIDSS-DXS and WFCAM PI programmes.

[16] calculated from calibration work at CASU (Hodgkin et al. 2009)

– Manage multiple layers of products: pawprints, tiles and mosaics, including external products automatically.
– Have a more sophisticated setup for multi-epoch products, specified by the **synopticSetup** string in `Programme`.

We set up all the requirements, the database schema and curation tables contents for a survey, when we start preparing a static release (e.g. VHS-DR1), using a combination of the programme requirements in the `Programme` table and the available data. We have made the infrastructure and processing the same for all surveys, unlike in UKIDSS where the wide shallow surveys (GPS, GCS and LAS) were processed differently from the deep surveys (DXS and UDS). This makes it easier for the operators who run tasks, and makes it much simpler if programmes evolve in the future. For instance, if the VHS decided to add in a second epoch in any filter, this would be automatically accommodated.

7.1. Stack, tiles and mosaics

Requirements for stack, tile and mosaic products are set up by grouping the data into different pointings by position, position angle and offset (if pawprints). The requirements for a particular release are stored in `RequiredStack`, `RequiredTile`, and `RequiredMosaic`. The stacking software uses the definitions to create the deepest stack possible for each product in `RequiredStack` from the pawprints. The tiling software creates tiles at each location using these stacks, so the tile requirements must be linked to the pawprint requirements.

The different layers of products (pawprints, tiles and mosaics) can be linked to each other using the `ProductLinks` table: e.g. tile productID 1 in `RequiredTile` in the VHS may be composed of pawprints with productIDs 1, 3, 5, 7, 9 & 11 in `RequiredStack`. `ProductLinks` links the requirements whereas `Provenance` links the image metadata from the actual files. From one release the values in `ProductLinks`, `RequiredTile` and `RequiredStack` may stay the same (although this is not guaranteed), but a product which initially contained 2 epochs worth of image data may be replaced by one containing 5 epochs worth of image data and will therefore link to different multiframes in `Multiframe` and `Provenance`.

External (made outside VDFS) products, e.g. VIDEO (or UKIDSS-UDS in the WSA) mosaics, which are created by the survey team and imported into the VSA, are set up via the `ExternalProducts` table, which contains the programme, product type, release number and information about who created the mosaic and where on the file system it can be found.

The required products and the actual image frames are now linked to each other via the `ProgrammeFrame` table which includes **programmeID**, **productID** and **releaseNum** and links to the image metadata tables via **multiframeID**. The release number for products is a running number from when WFAU first started producing releases for the science teams, so the products in the first public releases of the VISTA Public Surveys have a variety of release numbers depending on the programme. `ProgrammeFrame` is essential for keeping track of what images are related to each requirement. This makes it much easier for scientists and VSA support staff to keep track of what has been created and whether anything is missing. This infrastructure is crucial for the automated curation (Collins et al. 2009) of VSA products, where decisions are made about what tasks need to be run based on the requirements and what has already been completed.

When OB frames are recalibrated in multi-epoch programmes, the OB tiles are compared to the deeps and the zero-points adjusted accordingly. A change to the zero-point of the tile is propagated to the constituent pawprints. The code to propagate the zero-point differences was not developed until very recently, so most datasets in the first release will not include this propagation; at the time of writing only the VVV dataset will include this. The pawprint zero-points for the other multi-epoch public surveys (VIKING, VMC and VIDEO) will be correctly updated for all recalibrations in the data releases that contain data from semester P87 and beyond.

7.2. Multi-epoch tables

We have introduced the **synopticSetup** string attribute in `Programme`, to control the production of more than one Best Match table, i.e. both a `SourceXDetectionBestMatch` and a `SourceXSynopticSourceBestMatch`. In surveys such as the VVV, many scientists would like colour information for variable stars, so the colours must come from near-contemporaneous observations. This information is in the `SynopticSource` table whereas the colours in the `Source` table come from the deepest images which are stacked from several epochs of data and are certainly not contemporaneous. However, this survey will take many tens of epochs, mostly in one filter, K_s, so a `SynopticSource` table that covers the full time range would be inefficient - the *Z*, *Y*, *J* and *H* band attribute columns would contain mainly default values. A more efficient way is to specify the `SynopticSource` over a short time range and specify that the statistics in the `Variability` should come from data in `SourceXDetectionBestMatch`, which covers the whole time range. It is still necessary to link the `SynopticSource` table with all the other tables, so a `SourceXSynopticSourceBestMatch` table is required too.The **synopticSetup** attribute is a string, with the following value in the VVV: `BOTH:VAR-UNC:COR,SV,P87`, which can be parsed to give the following information: create both Best Match tables; use the uncorrelated (`SourceXDetectionBestMatch`) when calculating the variability statistics, and only use data between the beginning of the Science Verification period (SV) and the end of ESO Period 87 (P87) in the correlated table (`SourceXSynopticSourceBestMatch`).

8. Other enhancements since the launch of the WSA

In addition to the above changes necessary for processing VISTA data, we have made various changes to improve overall curation of WFAU products. These extend the database design described in Hambly et al. (2008) and Cross et al. (2009).

8.1. Improvements to multi-epoch data model and calculations

8.1.1. Creating the BestMatch tables

In Cross et al. (2009) § 9.2, we discussed possible improvements to checking missing observations to correctly create the `BestMatch` tables. One method that we discussed possibly implementing was the half-space method (Budavári et al. 2010). We define 16 half-spaces for each single epoch image. These 16 half-spaces come in 4 sets, one 2 pixels outside each image edge, one 2 pixels inside each image edge, one 2 pixels outside the edge of the jittered region where the exposure time per pixel

goes from the total exposure time to some fraction of it and one 2 pixels inside this edge. With 4 edges, this gives 16 half-spaces. We found that defining a half-space using 3 points: the two ends of an edge and the midpoint, the edge could usually be described to an accuracy of around one pixel, so a 2 pixel margin each side would encompass all points which we were unsure about. Each half-space is described using 4 numbers, a 3-D Cartesian vector, normal to the plane of the half-space and a constant that gives the offset of the plane from the centre of the sphere. All the half-space info is stored in a new table, `AstrometricInfo` (one for each multi-epoch programme, e.g. `vmcAstrometricInfo`), to help with the curation of multi-epoch data. As well as the half-space info, we have included place-holder columns for attributes to describe small adjustments to the astrometric solution of each image that will improve fitting to the proper motion, when we start calculating proper motions in VISTA data (see Collins & Hambly 2012, for the description of the method used for wide area UKIDSS surveys).

The half-spaces are used to check frames which do not have an expected match to a primary source in the `Source` table, (see Cross et al. 2009, § 9.2). If there is no detection, this may be for one of several reasons: the frame does not overlap with the part of the sky containing the source; the source is within the jitter regions, where the integration time is less and there may be a gradient in the integration time across the object; the source is too faint to be detected on a single exposure; the source is usually bright enough, but has faded below the detection threshold; the source is blended with another; the object has moved sufficiently far from the expected position.

Using the half-spaces allows us to flag the first two possibilities. Checking whether a detection should be within the image or within the jitter section is trivial and each calculation is extremely quick. Most importantly, since the half-space describes an edge accurately within a pixel or two, very few objects need the more careful test that use the WCS to calculate the exact position of the object on the frame. Using the half-spaces we are able to reduce the number of slow tests to only those objects within two thin shells, each 4 pixels wide, at the image edge and the edge of the jitter region. The half-space information is stored so that archive users can use it too, to rapidly determine whether an object is within the frame.

8.1.2. Expected noise model

We have made changes to the calculation of the expected noise. The expected noise is still based on fitting a function, in this case, a Strateva function to the RMS versus mean magnitude data (see Cross et al. 2009). In early team data releases (before version 1.1 data was released), the expected noise was simply the value of this function at the mean magnitude of the source. However, we found that the actual magnitude limit was often quite a bit brighter than the expected magnitude limit, especially when the field is very dense, see Fig 6. When this happens the RMS versus magnitude plot turns over and simply using the fit will underestimate the RMS; indeed sometimes the expected RMS will be negative. To mitigate against this, we have made the following changes:

- Calculate the turnover point: the maximum RMS as a function of magnitude if the function does turn over.
- For all magnitudes fainter than the turnover point or the maximum range of magnitudes in the fit (whichever is brightest), set the expected RMS to the value at this point, see Fig 6

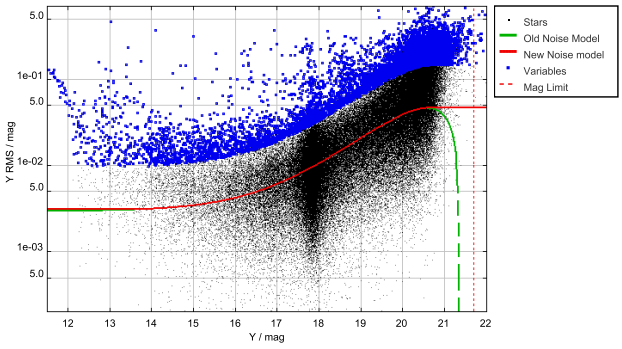

Fig. 6. Magnitude-RMS plot of a dense pointing in the VMC, where the fitting function turns over before the expected magnitude limit (dotted vertical line). The old expected RMS as a function of magnitude is plotted as the dashed line, and the new expected RMS as a function of magnitude is plotted as the solid line. Outside of the fitting range, or beyond the turnover point, the RMS is a constant. Non-variable stars are shown as dots, and variables are shown as open squares. At the bright end, $Y < 12.5$ mag, the RMS increases due to saturation effects, but this has not been included in the model yet.

- For all magnitudes brighter the minimum range of magnitudes in the fit, set the expected RMS to the value at this point.

The astrometric fit has also been updated. Instead of calculating a simple clipped mean, we calculate a weighted mean position. We calculate the expected astrometric noise in each filter, in the same way as we calculate the expected photometric noise, using the angular separation between each epoch position and the median absolute deviation clipped median position for stationary stars and fitting a function to these median values as a function of magnitude. This function describes the locus of the astrometric uncertainty for non-moving point sources, much as the equivalent fit for the magnitude RMS described the photometric uncertainty as a function of magnitude for non-variable point sources. This calculated uncertainty as a function of magnitude will be used to weight the position.

A particular pointing may only be observed once in one filter, and in some programmes a filter is only observed once in each pointing, so it is often not possible to fit a noise model for each pointing and filter. For photometric variability statistics, this is not a problem, since all the values are default if there is only one epoch, but when it comes to the astrometric fit, we would like to use all the data in all filters together, to improve the fit.

To estimate the errors on these frames, we calculate a default noise model in each filter which has a calculable noise in at least one pointing. We take all the calculated noise models, and calculate the mean RMS of these models in each of a set of nine bins across the magnitude range, and fit the noise model to the mean RMSs. This model is used in any pointing in this filter where there is only one epoch.

For filters where there is only ever one epoch, we cannot directly measure the noise as a function of magnitude, so we make the assumption that the behaviour in any filter is similar to the others (which is born out by experience of surveys where multiple observations are taken in all filters). We expect that the limit reached at the bright end will be the same and that the increase in noise toward the faint end depends on the depths of the exposure, which depends on the total exposure time or expected magnitude limit. Moreover, any other differences, such as the effects of sky

brightness or residual non-linearity are likely to be a function of wavelength, so we choose the nearest filter in wavelength that has enough epochs for a fit to be made to the RMS as a function of magnitude. We take the default model in this filter and adjust for the difference in expected magnitude limit, e.g. using the Strateva model, we will calculate new values for b and c as follows

$$
\begin{aligned}
<\zeta(m)> &= a + b\,10^{0.4m} + c\,10^{0.8m} \\
\Delta m &= m_{\rm l} - m'_{\rm l} \\
a &= a' \\
b &= b'\,10^{-0.4\Delta m} \\
c &= c'\,10^{-0.8\Delta m},
\end{aligned}
$$

where $m_{\rm l}$ is the magnitude limit of frames in this filter and $m'_{\rm l}$ is the magnitude limit of frames in the comparison filter.

The weighted mean position uses a 3σ clipped weighted mean in each of the three Cartesian coordinates, and then converts back to equatorial coordinates. The type of fit used is recorded in the `VarFrameSetInfo` table as **motionModel**. The model described above is a static weighted model: 'wgtstatic'. When we have VISTA data over several years we expect to fit for proper motion too.

8.2. Improvements to the interface

The VSA proprietary and public release databases can be accessed and queried via the web-browser based interface. The various access methods allow users to perform SQL queries on the science ready tables; extract image cut-outs and download entire image and catalogue files. In addition public releases will be accessible under the Virtual Observatory (VO). Releases will be discoverable in the VO registries. A Table Access Protocol[17] (TAP) interface to each data releases will allow users to perform SQL queries using the Astronomical Data Query Language (ADQL) from clients such as TOPCAT (Taylor 2005). Images will be accessible via Simple Image Access Protocol (SIAP[18]) services.

9. Illustrative science examples

9.1. Colours of VHS point sources

The optical-infrared colour-colour plot is a powerful classifier of different types of stars, with most stars lying along a narrow locus. However, extinction and poor photometry can widen this locus and prevent the separation of brown dwarfs, QSOs and compact galaxies. The following selection will select point sources in the VHS, which are matched to point sources in the SDSS and are not flagged for poor quality. We select the colours and positions, but only for stars in areas of low Galactic extinction.

```
SELECT s.sourceID, s.ra, s.dec,
/* select position colour and magnitude
information */
(sdPho.psfMag_g-s.jAperMag3) AS gmjPnt,
jmksPnt, ksAperMag3
/* from vhsSource, SDSS DR7 PhotoObjAll,
neighbour table */
```

[17] http://www.ivoa.net/Documents/TAP/

[18] http://www.ivoa.net/Documents/SIA/

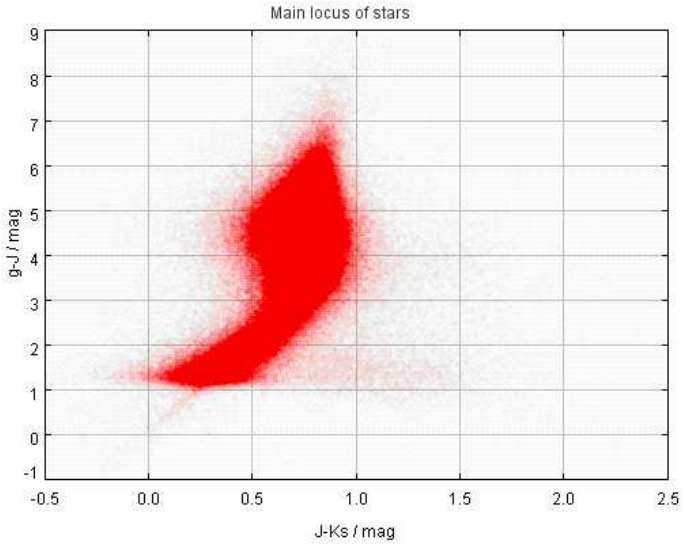

Fig. 7. The main locus of stars matched between the VHS and SDSS selected in (g-J) versus (J-Ks).

```
FROM vhsSource as s,
vhsSourceXDR7PhotoObjAll as x,
BESTDR7..PhotoObjAll as sdPho
/* join the tables */
WHERE s.sourceID = x.masterObjID AND
x.slaveObjID=sdPho.objID
/* find matches within 2 arcsec */
AND x.distanceMins<=0.0333
/* that are nearest matches */
AND x.distanceMins IN (
/* sub query to find minimum distance
for a match to this sourceID */
SELECT MIN(distanceMins)
FROM vhsSourceXDR7PhotoObjAll
WHERE masterObjID=x.masterObjID)
/* select SDSS primary objects and stars */
AND x.sdssPrimary=1 and x.sdssType=6
/* objects with no flags in VHS */
AND jppErrBits=0 AND ksppErrBits=0
/* stars or probable stars in VHS */
AND mergedClass IN (-1,-2)
/* Not default magnitudes in SDSS or VHS */
 AND sdPho.psfMag_g>0. AND s.jAperMag3>0.
AND s.ksAperMag3>0.
```

These data can be plotted using TOPCAT (Taylor 2005), and the main locus can be found, as seen in Fig. 7. The gradient of the locus in the colour-colour plot can be measured and then rare objects can be further selected.

9.2. Flare stars

The following query selects objects that could be flaring stars or some type of cataclysmic variable. To do this, we select sources that have a minimum K_s magnitude that is at least 2 magnitudes brighter than the median magnitude and further, it has at least 2 measurements that are brighter than the median by 0.5 magnitudes. This second constraint should remove sources with one point that has escaped flagging. We also want at least 5 good K_s detections for a reasonable light curve. The following query was performed on the VVV survey.

```
 SELECT v.sourceID, s.ra, s.dec,
 /* select some useful attributes, pointing info,
number of observations, min, medium, maximum,
variable class, and star/galaxy class */
v.framesetID, ksnGoodObs, ksMinMag, ksMedianMag,
ksMaxMag, variableClass, mergedClass,
(ksMedianMag-ksMinMag) as ksFlareMag,
COUNT(*) AS nBrightDetections
/* from vvvVariability and vvvSource */
FROM vvvVariability as v,vvvSource as s,
vvvSourceXDetectionBestMatch as b,
vvvDetection as d
/* first join the tables */
WHERE s.sourceID=v.sourceID AND b.sourceID=
v.sourceID AND b.multiframeID=d.multiframeID
AND b.extNum=d.extNum AND b.seqNum=
d.seqNum AND
/* select the magnitude range, brighten than
Ks=17 and not default. */
ksmedianMag<18. and ksmedianMag>0. AND
/* at least 5 observations */
ksnGoodObs>=5 AND ksbestAper=5 AND
/* Min mag is at least 2 magnitudes brighter
than median mag(but minMag is not default) */
(ksmedianMag-ksminMag)>2. and ksMinMag>0. AND
/* Only good Ks band detections in same
aperture as statistics are calculated in*/
d.seqNum>0 AND d.ppErrBits IN (0,16) AND
d.filterID=5 AND d.aperMag5>0 AND
d.aperMag5<(ksMedianMag-0.5)
/* Group detections */
GROUP BY v.sourceID, s.ra, s.dec,
v.framesetID, ksnGoodObs, ksMinMag, ksMedianMag,
ksMaxMag, variableClass, mergedClass
HAVING COUNT(*)>2
/* Order by largest change in magnitude first.*/
ORDER BY ksMedianMag-ksMinMag DESC
```

We plot the Ks-band light curve of one of these objects in Fig 8. The majority of the detections are 14^{th} magnitude, but there is a flare of almost 2.5 mag followed by fading of 1 mag before the star returns to K_s = 13.9 mag.

9.3. Global properties of VIKING-SDSS galaxies

The following selection uses IR photometry and sizes from VIKING combined with optical colours and redshifts – both spectroscopic and photometric – from SDSS. In this query we use neighbour tables to join VIKING and SDSS. We use the SQL command "UNION" to combine the query which matches to galaxies with spectra to the query for those with only photometric redshifts. Users who are worried about completeness can use the "UNION" command to combine further queries, such as those that select for VIKING galaxies without SDSS matches or ones for SDSS matches but neither spectroscopic or photometric redshifts. The two combined queries must have the same number of columns, with the same names. In the cases where one query has columns which the other has no entry for (e.g. redshift, z) we can fill this column with a default number, just as we demonstrate below with the redshift status and spectroscopic type.

```
SELECT
/* select information necessary to create
bi-variate brightness distribution,
```

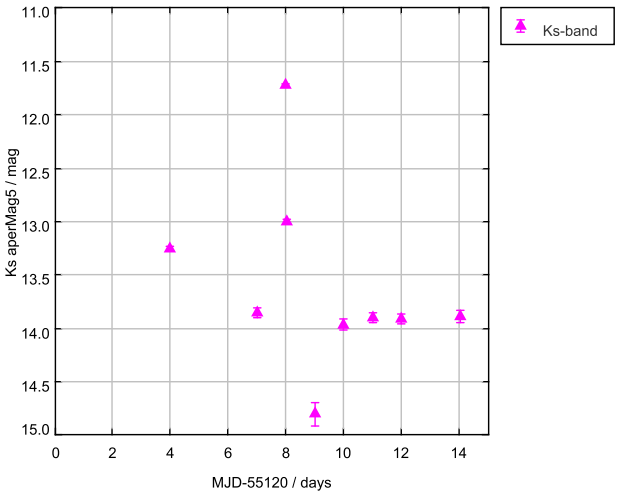

Fig. 8. The light curve of a flare star in the VVV selected using the query above.

```
extinction corrected Petrosian magnitudes
put into AB system and seeing corrected,
semi-major axis size, PLUS SDSS colours
and redshifts (spectroscopic and
 photometric)
*/
s.sourceID,s.ra,s.dec,s.frameSetID,
(s.hPetroMag-s.aH+fh.VegaToAB) AS hPetroAB,
s.hHlCorSMjRadAs,(s.hPetroMag+
2.5*log10(2.*3.14159*s.hHlCorSMjRadAs*
s.hHlCorSMjRadAs)-s.aH+fh.VegaToAB) AS
hSurfBright, (s.ksPetroMag-s.aKs+
fks.VegaToAB) as ksPetroAB,
s.ksHlCorSMjRadAs,(s.ksPetroMag+
2.5*log10(2.*3.14159*s.ksHlCorSMjRadAs*
s.ksHlCorSMjRadAs)-s.aKs+fks.VegaToAB)
AS ksSurfBright, dr7spec.objID as sdssID,
((dr7spec.modelMag_u-dr7spec.extinction_u)-
 (dr7spec.modelMag_g-dr7spec.extinction_g))
as umgModel,z,zErr,zConf,zStatus,specClass
/* from vikingSource, Filter (one for each
filter for VegaToAB), SDSS-DR7 neighbour
table, SDSS SpecPhoto table */
FROM vikingSource AS s,Filter AS fh,Filter
AS fks, vikingSourceXDR7PhotoObjAll AS xdr7,
BESTDR7..SpecPhotoAll as dr7spec
/* First join tables, */
WHERE xdr7.masterObjID=s.sourceID AND
fh.filterID=4 AND fks.filterID=5 AND
dr7spec.objID=xdr7.slaveObjID AND
 /* select VIKING primary sources matched to
 SDSS primary sources */
 (priOrSec=0 OR priOrSec=frameSetID) AND
sdssPrimary=1 AND
dr7spec.sciencePrimary=1 AND
/* within 2" of nearest match */
 xdr7.distanceMins<0.03333 AND
 xdr7.distanceMins IN (
  SELECT MIN(distanceMins)
  FROM vikingSourceXDR7PhotoObjAll
  WHERE masterObjID=xdr7.masterObjID AND
```

```
   sdssPrimary=1) AND
/* for objects classified as galaxies or
probable galaxies in VIKING */
mergedClass IN (1,-3) AND
/* h and ks size is 0.7<sma<=10. arcsec */
ksHlCorSMjRadAs>0.7 AND hHlCorSMjRadAs>0.7
AND ksHlCorSMjRadAs<=10.0 AND
hHlCorSMjRadAs<=10.0 AND
/* good quality data in VIKING h and ks */
hppErrBits=0 AND ksppErrBits=0 AND
 /* ks extinction corrected AB mag < 20.5 */
 (ksPetroMag-aKs+fks.VegaToAB)<20.5 AND
/* ra and dec range to restrict to where
SDSS is */
s.ra>100. AND s.ra<250. AND s.dec>-5. AND
/* z>=0.002 */
 dr7spec.z>=0.002
 /* Add in ones which do not have SDSS spectra
using UNION */
UNION
SELECT
/* select information necessary to create
bi-variate brightness distribution,
extinction corrected
Petrosian magnitudes put into AB system and
seeing corrected, semi-major axis size AND SDSS
matches to PhotoObj table and photoz table */
s.sourceID,s.ra,s.dec,s.frameSetID,
(s.hPetroMag-s.aH+fh.VegaToAB) AS hPetroAB,
s.hHlCorSMjRadAs,(s.hPetroMag+
2.5*log10(2.*3.14159*s.hHlCorSMjRadAs*
s.hHlCorSMjRadAs)-s.aH+fh.VegaToAB) AS
hSurfBright, (s.ksPetroMag-s.aKs+fks.VegaToAB)
as ksPetroAB,s.ksHlCorSMjRadAs,(s.ksPetroMag+
2.5*log10(2.*3.14159*s.ksHlCorSMjRadAs*
s.ksHlCorSMjRadAs)-s.aKs+fks.VegaToAB) AS
ksSurfBright, dr7phot.objID as sdssID,
((dr7phot.modelMag_u-dr7phot.extinction_u)-
(dr7phot.modelMag_g-dr7phot.extinction_g))
as umgModel, photz.z as z,photz.zErr as zErr,
-9.9999 as zConf,-9 as zStatus,-9 as specClass
/* from vikingSource, Filter (one for each
filter for VegaToAB), SDSS-DR7 neighbour
table, */
FROM vikingSource AS s,Filter AS fh,Filter AS
fks,vikingSourceXDR7PhotoObjAll AS xdr7,
BESTDR7..PhotoObjAll as dr7phot,
BESTDR7..photoz as photz
/* First join tables, */
WHERE xdr7.masterObjID=s.sourceID AND
fh.filterID=4 AND fks.filterID=5 AND
dr7phot.objID=xdr7.slaveObjID AND photz.objID=
dr7phot.objID AND dr7phot.objID NOT IN (
  SELECT dr7spec.objID
  FROM BESTDR7..SpecPhotoAll as dr7spec
  WHERE dr7spec.objID=xdr7.slaveObjID AND
  dr7spec.sciencePrimary=1) AND
  /* select VIKING primary sources matched to
  SDSS primary sources */
  (priOrSec=0 OR priOrSec=frameSetID) AND
  sdssPrimary=1 AND
   /* within 2" of nearest match */
   xdr7.distanceMins<0.03333 AND
   xdr7.distanceMins IN (
```

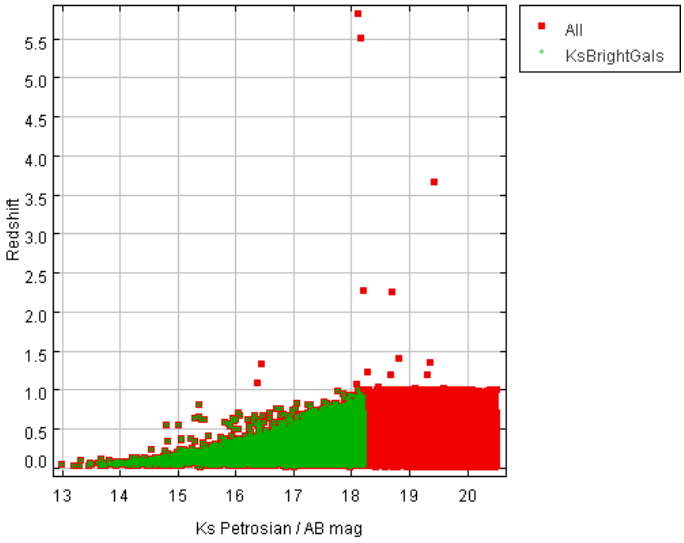

Fig. 9. Ks-band AB Petrosian magnitude versus redshift. This plot shows very few redshifts greater than $z = 1$, shown in red. The galaxies with $z > 1$ are likely to have spectroscopic redshifts, whereas the photometric redshifts are limited to $0 < z < 1$. Galaxies with $K_{AB} < 18.2$ tend to have $z < 1$, so we have selected a sample in TOPCAT which have $K < 18.2$ and $z < 1$, which are shown in green.

```
   SELECT MIN(distanceMins)
   FROM vikingSourceXDR7PhotoObjAll
   WHERE masterObjID=xdr7.masterObjID AND
   sdssPrimary=1) AND
   /* for objects classified as galaxies or
probable galaxies in VIKING */
mergedClass IN (1,-3) AND
/* h and ks size is 0.7<sma<=10. arcsec */
ksHlCorSMjRadAs>0.7 AND hHlCorSMjRadAs>0.7 AND
ksHlCorSMjRadAs<=10.0 AND hHlCorSMjRadAs<=10.0
/* good quality data in VIKING h and ks */
AND hppErrBits=0 AND ksppErrBits=0 AND
/* ks extinction corrected AB mag < 20.5 */
(ksPetroMag-aKs+fks.VegaToAB)<20.5 AND
/* ra and dec range to restrict to SDSS */
s.ra>100. AND s.ra<250. AND s.dec>-5. AND
/* z>=0.002 */
photz.z>=0.002
```

We plot the magnitude against the redshift for these galaxies, Fig 9, and find that there is an artificial selection at $z < 1$. The most likely explanation is that the photometric redshifts are limited to this range, since the SDSS optical colours do not give reliable photometric redshifts outside this range. By selecting a subsample at $K_s < 18.2$ and $z < 1$, we have a more complete sample with reliable redshifts. We use this sample to look at the colour-magnitude plot and the surface brightness magnitude plot of galaxies, Figs 10 & 11. The surface brightness, colour, magnitude and redshift are all fundamental for classifying galaxies.

9.4. Extragalactic variables in VIDEO

In deep extragalactic surveys, such as VIDEO, with many epochs over several months or years, it is possible to find a range of AGN, and very occasionally supernovae. There are also a few foreground stars that show variability. We select point-source variables VIDEO survey which show a range in magnitudes greater than 0.1 mag in any filter. Since the filters in this

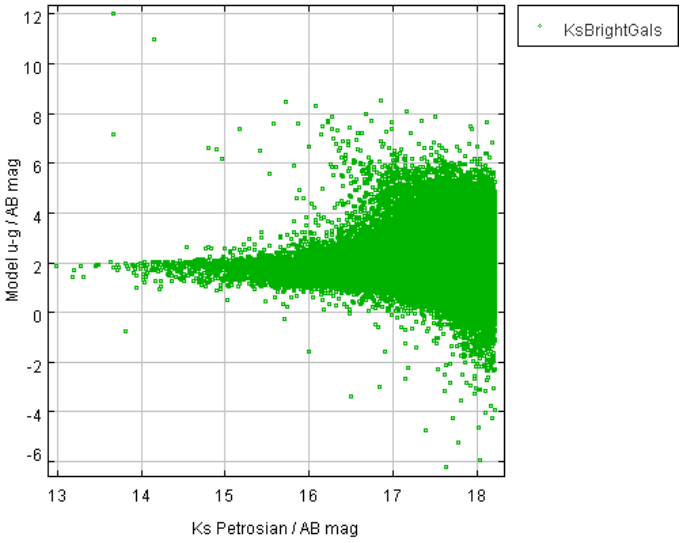

Fig. 10. Ks-band AB Petrosian magnitude versus u-g model mag colour for the $K_{AB} < 18.2$, $z < 1$ sample selected in Fig 9. At the bright end, the red-sequence is clear, but at fainter magnitudes, galaxies will be at higher redshift, so the observed colours are less meaningful.

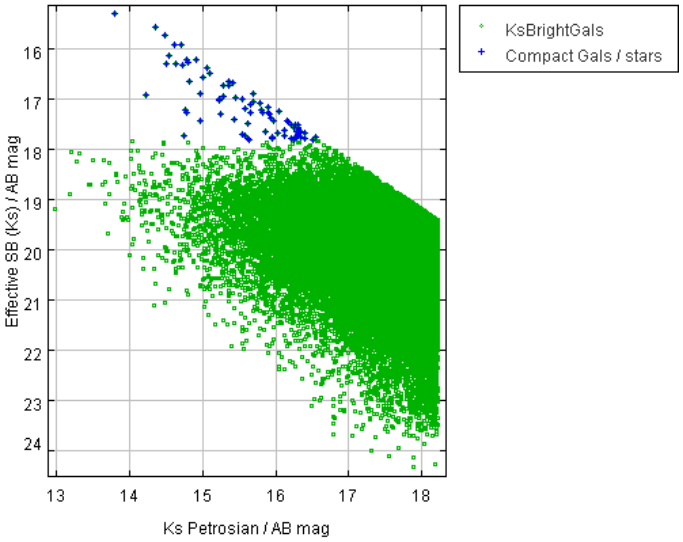

Fig. 11. Ks-band AB Petrosian magnitude versus effective Ks surface brightness for the $K_{AB} < 18.2$, $z < 1$ sample selected in Fig 9. The hard limit at the top-right hand side shows the size limit of 0.7″. The main galaxy population seems to have a high surface brightness limit of $\mu_{Ks} = 18.$ mag arcsec^{-2}, with a small group at higher surface brightnesses (shown as blue crosses) - either compact galaxies or stars that have managed to avoid all the selection criteria. Galaxies have surface brightnesses as low as $\mu_{Ks} = 24.0$ mag arcsec^{-2}, but fainter galaxies would need to be selected before the VIKING surface brightness became a limiting factor.

survey are not taken simultaneously, and AGN show sporadic variability, variations may only be seen in one filter.

```
SELECT s.sourceID,s.ra,s.dec,v.frameSetID,
v.zMedianMag,v.zMagRms,v.znGoodObs,v.zSkewness,
(v.zMaxMag-v.zMinMag) AS zRange,v.yMedianMag,
v.yMagRms,v.ynGoodObs,v.ySkewness,
(v.yMaxMag-v.yMinMag) AS yRange,v.jMedianMag,
 v.jMagRms,v.jnGoodObs,v.jSkewness,
 (v.jMaxMag-v.jMinMag) AS jRange,v.hMedianMag,
```

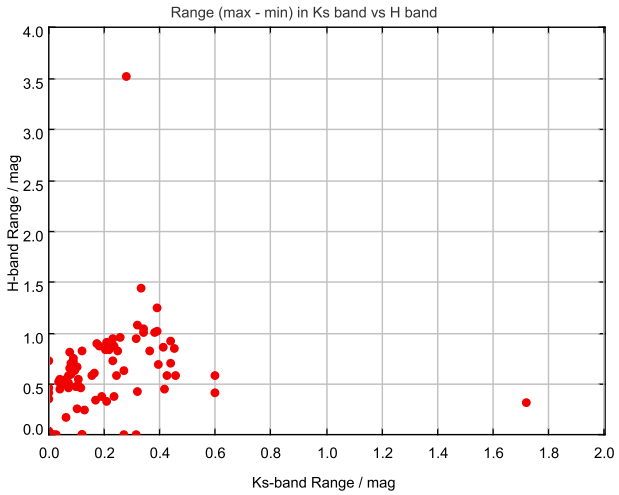

Fig. 12. Ks-band magnitude range versus H-band magnitude range. While there is a correlation between the two ranges, the H-band range seems greater than the Ks-band on average. There are some objects which have no discernible variation in Ks or H, but do so in one or more of the other filters.

```
v.hMagRms,v.hnGoodObs,v.hSkewness,
(v.hMaxMag-v.hMinMag) AS hRange,v.ksMedianMag,
v.ksMagRms,v.ksnGoodObs,v.ksSkewness,
(v.ksMaxMag-v.ksMinMag) AS ksRange
FROM videoVariability AS v,videoSource AS
s /* join tables */
WHERE v.sourceID=s.sourceID AND
/* point source variables */
 s.mergedClass IN (-1,-2) AND
v.variableClass=1 AND
 /* delta mag in > 0.1 in ANY filter, with
at least 5 good obs in that filter */
(((zMaxMag-zMinMag)>0.1 AND zMinMag>0.
AND znGoodObs>=5) OR ((yMaxMag-yMinMag)>0.1
AND yMinMag>0. AND ynGoodObs>=5) OR
((jMaxMag-jMinMag)>0.1 AND jMinMag>0. AND
jnGoodObs>=5) OR ((hMaxMag-hMinMag)>0.1
AND hMinMag>0. AND hnGoodObs>=5) OR
((ksMaxMag-ksMinMag)>0.1 AND ksMinMag>0.
 AND ksnGoodObs>=5))
```

We can plot some of the variability statistics, such as the range in the K_s band against the range in the H band, see Fig 12 or the RMS against the skewness in the K_s band, Fig 13. These types of plots help to classify different types of variable and to pick out odd objects.

We can then select one of these objects, e.g. the object in Fig 13 with a K_s band RMS > 0.4 mag, which is more than twice the RMS of any of the other objects and plot the light curve. To do this, we do a second query, below:

```
SELECT
/* Select time, filter, magnitude, magnitude
error and flags */
d.mjd,d.filterID,d.aperMag3,d.aperMag3Err,
 d.ppErrBits
/* From BestMatch table to link all
 observations of the same source and
videoDetection for each observation */
```

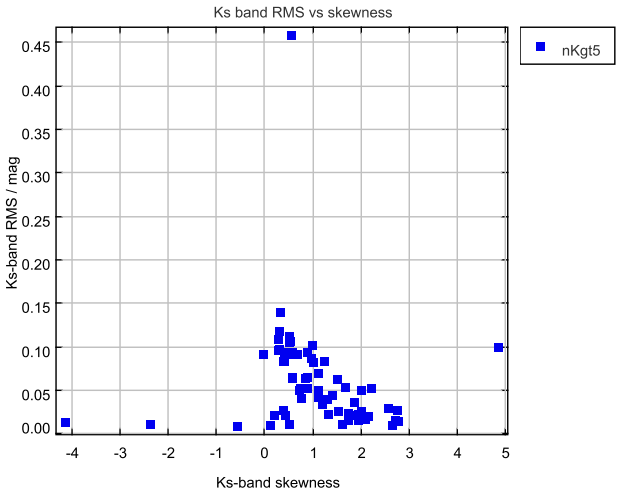

Fig. 13. Ks-band Skewness versus RMS for objects with at least 5 good Ks-band observations. Most of the objects selected have positive skews in the Ks-band. The skew decreases as the RMS increases.

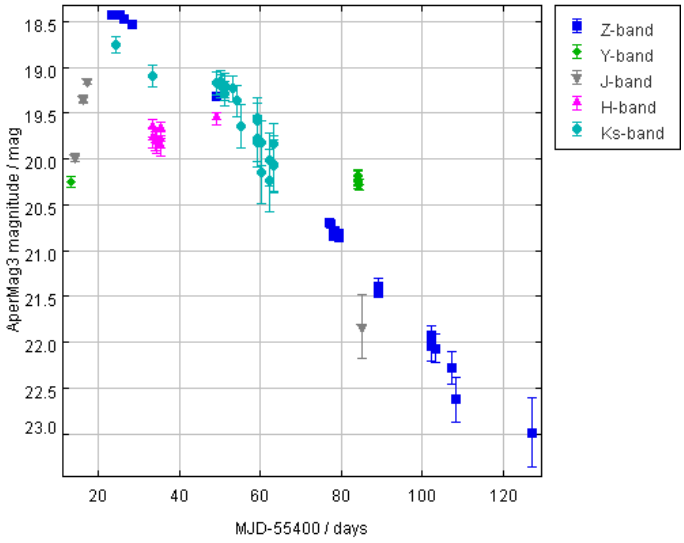

Fig. 14. Light-curve of variable selected in text. This variable brightens by ~ 2 magnitudes in less than 10 days and then fades by almost 5 magnitudes over the next 100 days. This is the expected behaviour of a supernova Type 1a.

```
 FROM videoDetection as d,
videoSourceXDetectionBestMatch as b
 /* First join tables */
 WHERE b.multiframeID=d.multiframeID AND
b.extNum=d.extNum AND b.seqNum=d.seqNum
 /* then select only detections and sourceID
 equal to object in previous selection
which has a Ks-band RMS>0.4 mag */
AND d.seqNum>0 AND b.sourceID IN (
SELECT s.sourceID
FROM videoVariability AS v,videoSource AS s
 /* join tables */
 WHERE v.sourceID=s.sourceID AND
/* point source variables */
s.mergedClass IN (-1,-2) AND
v.variableClass=1 AND
/* delta mag in > 0.1 in ANY filter, with
at least 5 good obs in that filter */
 (((zMaxMag-zMinMag)>0.1 AND zMinMag>0.
AND znGoodObs>=5) OR ((yMaxMag-yMinMag)>0.1
AND yMinMag>0. AND ynGoodObs>=5) OR
 ((jMaxMag-jMinMag)>0.1 AND jMinMag>0. AND
 jnGoodObs>=5) OR ((hMaxMag-hMinMag)>0.1 AND
hMinMag>0. AND hnGoodObs>=5) OR
 ((ksMaxMag-ksMinMag)>0.1 AND ksMinMag>0. AND
ksnGoodObs>=5))
 /* Ks-band RMS >0.4 mag */
 AND ksMagRms>0.4)
 /* order by time */
 ORDER BY d.mjd
```

We can use TOPCAT[19] to plot the light-curve, see Fig 14. The light-curve is very interesting, showing a short phase of brightening followed by a longer phase of fading, characteristic of an exploding star, probably a Type 1a SNa, with a maximum brightness of $Z = 18.5$ mag. The position and time of this object match SN2010gy (Chornock et al. 2010). The discovery team found that it has a redshift, $z = 0.06$, but could not find a likely host galaxy. The thumbnail of this source from the deep K_s band mosaic is shown in Fig 15.

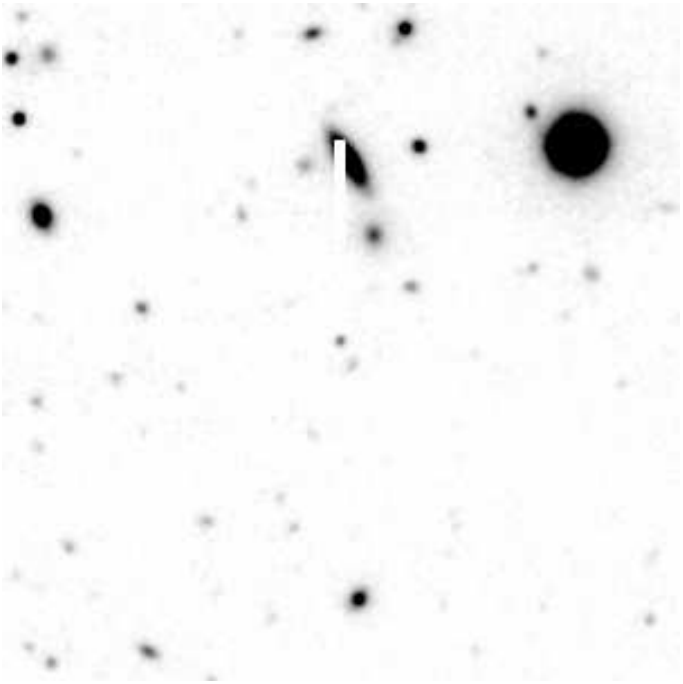

Fig. 15. Thumbnail of point-source variable which has light curve shown in Fig 14. Thumbnails can be shown by selecting the attributes **ra**, **dec** and **frameSetID** from the Source table and clicking on the link.

10. First Public Data Releases

The first public releases of VISTA Public Survey data through the VSA are intended to match the DR1/2 datasets published for each survey in the ESO SAF[20]. Thus, they will cover the data up to the end of ESO semester P85 (i.e. up to 30th September

[19] http://www.star.bris.ac.uk/~mbt/topcat/

[20] http://www.eso.org/sci/observing/phase3/data_releases.html

2010). Some of the surveys have released data from P86 as well, but with the following additional constraints:

- VMC: the data released will only be those fields where the whole set of epochs is complete, i.e. the 30 Doradus field (5h37m40s,-69° 22′18″) and the Gaia South Elliptical Pole field (5h59m23s,-66°20′28″).
- VIKING: the data will only be released in the following fields: GAMA09 (33 fields overlapping with the Galaxy and Mass Assembly 09h field; GAMA, Driver et al. 2011), CFHTLS-W1 (6 fields overlapping with the Canada France Hawaii Legacy Survey W1 field[21] and 9 fields in the Southern Galactic Pole region.
- VIDEO: the data will only be released in the pointings which VIDEO mosaics have been created in (XMM3 field, 2h26m18s, -4°44′; ES1-North field, 0h37m49s, -43°30′).

We have cropped all the tables in the surveys to match the pointings specified by the PIs. The excluded data will be released in future releases. Table 1 summarises the contents of the releases. Unlike UKIDSS each survey will be released into a separate database. The sixth public survey UltraVISTA is not using VDFS processing, except for the initial pawprint pipeline calibration, so we are not releasing data from this survey.

11. Summary and future work

The VSA was designed as the main access point to the VISTA survey data, allowing users to carefully select the data they need, rather than to bulk download all the data, a difficult and time consuming job in the era of billion row catalogues. As we have shown in the Illustrated Examples (§ 9), the VSA is designed to allow users to select on a wide range of attributes and to work with external data, such as the Sloan Digital Sky Survey, WISE, Glimpse, OGLE etc. The VSA is based on the WFCAM Science Archive but has VISTA specific features and various improvements based on our experience of WFCAM data and archive processing.

In the future we plan several enhancements. In the near future, we are working on improvements to our interface, including a MyDB (Li & Thakar 2008) style access, where users can combine queries with Python scripts, to produce a powerful work environment. We are also improving our plotting tools to more easily show very large datasets with a combination of density maps where the number of points is huge, and individual points where the density drops below a threshold.

12. Acknowledgements

Many people have contributed to the development of the VISTA Science Archive. We would like to thank the PIs and members of the VISTA Public Survey teams who have made requests, spotted mistakes, and checked the early releases carefully. We would also like to thank users of the WFCAM Science Archive. We thank Steve Warren for providing Figure 2.

We thank Microsoft also for providing software via the Microsoft Development Network Academic Alliance, and eGenix for providing a free, non-commercial licence for Python middleware.

Financial resources for VISTA/WFCAM Science Archive development and operations were provided by the UK Science and Technology Facilities Council.

[21] http://www.cfht.hawaii.edu/Science/CFHLS/

References

Bertin, E. & Arnouts, S. 1996, A&AS, 117, 393
Binney, J. & Merrifield, M. 1998, Galactic Astronomy, ed. Binney, J. & Merrifield, M.
Blanton, M. R., Dalcanton, J., Eisenstein, D., et al. 2001, AJ, 121, 2358
Bonfield, D. G., Jarvis, M. J., & consortium, V. 2010, in Bulletin of the American Astronomical Society, Vol. 36, American Astronomical Society Meeting Abstracts 215, 604.06
Budavári, T., Szalay, A. S., & Fekete, G. 2010, PASP, 122, 1375
Calabretta, M. R. & Greisen, E. W. 2002, A&A, 395, 1077
Casali, M., Adamson, A., Alves de Oliveira, C., & et al. 2007, A&A, 467, 777
Chornock, R., Huber, M., Foley, R. J., et al. 2010, Central Bureau Electronic Telegrams, 2414, 1
Cioni, M.-R. L., Clementini, G., Girardi, L., et al. 2011, A&A, 527, A116+
Collins, R., Cross, N., Sutorius, E., Read, M., & Hambly, N. 2009, in Astronomical Society of the Pacific Conference Series, Vol. in press, Astronomical Data Analysis Software and Systems XVIII, ed. D. A. Bohlender, D. Durand, & P. Dowler
Collins, R. & Hambly, N. 2012, in Astronomical Society of the Pacific Conference Series, Vol. in press, Astronomical Data Analysis Software and Systems XXI, ed. , , &
Cross, N., Collins, R., Sutorius, E., et al. 2010, ArXiv e-prints
Cross, N., Driver, S. P., Couch, W., et al. 2001, MNRAS, 324, 825
Cross, N. J. G., Collins, R. S., Hambly, N. C., et al. 2009, MNRAS, 399, 1730
Dalton, G. B., Sutherland, W. J., Emerson, J. P., et al. 2010, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 7735, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series
Disney, M. J. 1976, Nature, 263, 573
Driver, S. P., Hill, D. T., Kelvin, L. S., et al. 2011, MNRAS, 413, 971
Driver, S. P., Liske, J., Cross, N. J. G., De Propris, R., & Allen, P. D. 2005, MNRAS, 360, 81
Emerson, J., McPherson, A., & Sutherland, W. 2006, The Messenger, 126, 41
Emerson, J. & Sutherland, W. 2010a, The Messenger, 139, 2
Emerson, J. P., Irwin, M. J., Lewis, J., et al. 2004, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 5493, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, ed. P. J. Quinn & A. Bridger, 401–410
Emerson, J. P. & Sutherland, W. J. 2010b, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 7733, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series
Findlay, J. R., Sutherland, W. J., Venemans, B. P., et al. 2012, MNRAS, 419, 3354
Fruchter, A. S. & Hook, R. N. 2002, PASP, 114, 144
Golub, G. & Reinsch, C. 1970, Numerische Mathematik, 14, 403, 10.1007/BF02163027
Graham, A. W. & Driver, S. P. 2005, PASA, 22, 118
Graham, A. W., Driver, S. P., Petrosian, V., et al. 2005, AJ, 130, 1535
Hambly, N. C., Collins, R. S., Cross, N. J. G., & et al. 2008, MNRAS, 384, 637
Hodgkin, S. T., Irwin, M. J., Hewett, P. C., & Warren, S. J. 2009, MNRAS, 394, 675
Irwin, M. 2009, UKIRT Newsletter, 25, 15
Irwin, M. 2010, UKIRT Newsletter, 26, 14
Irwin, M. J., Lewis, J., Hodgkin, S., et al. 2004, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 5493, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, ed. P. J. Quinn & A. Bridger, 411–422
Kormendy, J. 1977, ApJ, 218, 333
Lawrence, A., Warren, S. J., Almaini, O., et al. 2007, MNRAS, 379, 1599
Lewis, J. R., Irwin, M., & Bunclark, P. 2010, in Astronomical Society of the Pacific Conference Series, Vol. 434, Astronomical Data Analysis Software and Systems XIX, ed. Y. Mizumoto, K.-I. Morita, & M. Ohishi, 91
Li, N. & Thakar, A. R. 2008, Computing in Science and Engineering, 10, 18
Liske, J., Lemon, D. J., Driver, S. P., Cross, N. J. G., & Couch, W. J. 2003, MNRAS, 344, 307
Pence, W. D., Chiappetti, L., Page, C. G., Shaw, R. A., & Stobie, E. 2010, A&A, 524, A42
Peng, C. Y., Ho, L. C., Impey, C. D., & Rix, H.-W. 2002, AJ, 124, 266
Saito, R. K., Hempel, M., Minniti, D., et al. 2012, A&A, 537, A107
Skrutskie, M. F., Cutri, R. M., Stiening, R., & et al. 2006, AJ, 131, 1163
Smith, A. J., Loveday, J., & Cross, N. J. G. 2009, MNRAS, 397, 868
Szalay, A. S., Gray, J., Thakar, A. R., et al. 2002, eprint arXiv:cs/0202013
Taylor, M. B. 2005, in Astronomical Society of the Pacific Conference Series, Vol. 347, Astronomical Data Analysis Software and Systems XIV, ed. P. Shopbell, M. Britton, & R. Ebert, 29
York, D. G., Adelman, J., Anderson, Jr., J. E., et al. 2000, AJ, 120, 1579

Table 1. Summary of VISTA Public Survey DR1 VSA releases.

Survey	N(pointings)	Filters and depth[a] AB (mag)	N(epochs) (typical)	N(sources)
VHS	1260	$Y \sim 20.9$, $J \sim 20.9$, $H \sim 20.7$, $K_s \sim 20.2$	1,1,1,1	1.2×10^8
VVV	350	$Z \sim 21.9$, $Y \sim 20.9$, $J \sim 20.8$, $H \sim 20.2$, $K_s \sim 19.3$	1,1,1,1,6	5.0×10^8
VMC	2	$Y \sim 21.9$, $J \sim 21.9$, $K_s \sim 22.1$	4,4,12	1.8×10^6
VIKING	48	$Z \sim 23.0$, $Y \sim 21.9$, $J \sim 21.8$, $H \sim 21.4$, $K_s \sim 21.3$	1,1,2,1,1	5.3×10^6
VIDEO	2	$Z \sim 25.6$, $Y \sim 25.1$, $J \sim 25.2$, $H \sim 24.8$, $K_s \sim 24.5$	33,31,38,40,50	7.5×10^5

[a] for multi-epoch filters, this is calculated from the catalogue, but for a single epoch, it is estimated from the exposure time.

Appendix A: Image scaling

For single epoch OB stacks,

$$bscale = \frac{1}{NDIT\ \sqrt{NJITTER}}, \tag{A.1}$$

and for tiles,

$$bscale = \frac{1}{NDIT\ \sqrt{2\,NJITTER}}, \tag{A.2}$$

where $NDIT$ is the number of readouts during the integration of a raw image and $NJITTER$ is the number of jitter positions in the jitter pattern to create a single pawprint stack. For deep stacks, we would scale by the number of epochs, in the same way as the number of jitters, but since $NDIT$ and $NJITTER$ can vary from stack to stack in the same programme, pointing and filter, we calculate *bscale* for deep stacks as:

$$bscale^{\rm deep} = \frac{1}{\sqrt{\sum_i \frac{1}{bscale_i^2}}} \tag{A.3}$$

For deep tiles, to take account of different integration times in each offset and deprecated detectors in some OB stacks, we compute the *bscale* values for each detector as above in each overlap and average over all overlaps, just as we calculate the total exposure time for a tile.

Appendix B: Half-light radii

The sizes of extended sources are difficult to measure for various reasons:

- The outer parts of a galaxy eventually merge into the sky background, so it is difficult to know how much of the galaxy is lost in the sky. For intrinsically low surface brightness galaxies or high redshift galaxies, the majority of the light may well be lost in the background and any measurement is a significant underestimate (Disney 1976; Cross et al. 2001, e.g.)
- The profile is not always smooth or axisymmetrical, e.g. grand design spirals, irregular galaxies, interacting galaxies.
- Nearby objects can make it difficult to get accurate measurements of the total luminosity and extent of a galaxy. The measurement of the background level is sometimes incorrect and this can affect the curve-of-growth measurement.
- Galaxies have various inclinations to the line of sight. Different researchers may want to use measurements that correct for inclination or do not.
- The light of galaxies and all objects is convolved with a point-spread function that will particularly effect small objects.
- The method must be robust enough and quick enough to be applied to the VISTA detection tables. We exclude the VVV and VMC since these are in extremely dense regions where contamination from nearby objects is almost guaranteed and the vast majority of sources are stellar which are point-sources. Even so, the catalogues that measure sizes will contain $> 10^8$ sources, and maybe 10^9 sources.

Our measurements of the size of galaxies try to take into account all the above effects as much as possible. We define our basic size measurement as the radius containing half of the flux of the galaxy – the half-light-radius – a measurement that has been used extensively before (Kormendy 1977; Cross et al. 2001; Blanton et al. 2001). The main difficulty with this measurement is measuring the total flux.

To take into account the missing light from the outer parts of the galaxy, we use the Petrosian flux (Blanton et al. 2001; Graham et al. 2005), which is generally insensitive to the effects of surface brightness, i.e. if you keep the galaxy profile the same (the relative flux as a function of radius) but reduce or increase the average surface-brightness, then the Petrosian flux measurement will return the same flux each time. This breaks down eventually: if you reduce the surface brightness enough the galaxy won't even be detectable, and close to this limit the total flux and size will become difficult to measure with any accuracy.

The Petrosian flux however gives different results for different profiles, which is an issue. Blanton et al. (2001) showed that while only 0.7% of the flux of an exponential disk galaxy was typically missed by the Petrosian, 22% of the flux of a de Vaucouleurs' profile elliptical galaxy was missed, and 5% of the flux of a PSF dominated profile was missed, although Graham & Driver (2005) shows that there are slightly different results for a standard Petrosian definition compared to the SDSS Petrosian that Blanton used. Small galaxies, close to the seeing limit will be dominated by the point-spread function. Galaxies close to the surface-brightness limit of the survey could be missing much more of the light. To try to take into account the missing light, we assume that all galaxies are missing 10%. This will be an overestimate in some cases and an underestimate in others, but to try and calculate a correction for each galaxy would require an iterative procedure which would take much longer, and in any case, it would be better to fit profiles for all objects (Peng et al. 2002, e.g.). The light profiles of galaxies are often well fit by Sérsic profiles (Graham et al. 2005), a more general function, that includes exponential ($\beta = 1$), de Vaucouleur ($\beta = 4$) and Gaussian ($\beta = 0.5$),

$$I(r) = I_{r_{hl}} \exp\left\{-k\left[\left(\frac{r}{r_{hl}}\right)^{1/\beta} - 1\right]\right\} \tag{B.1}$$

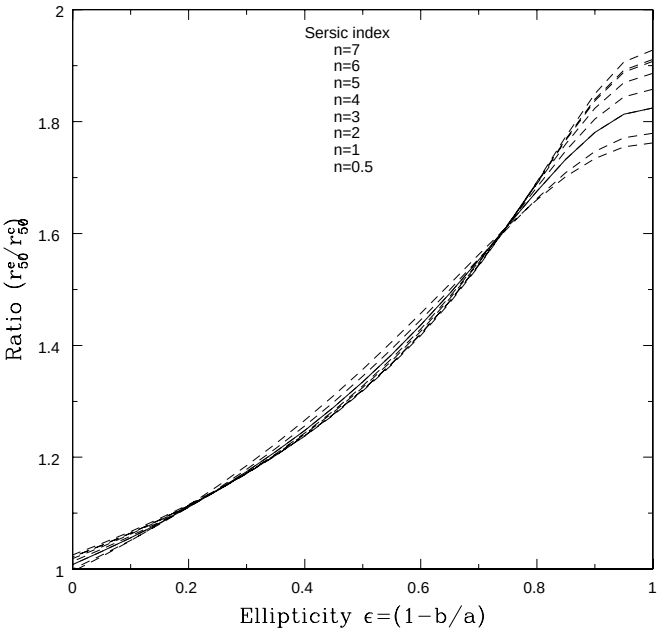

Fig. B.1. Figure of the ellipticity versus ratio of half-light semi-major axis to half-light radius for different Sérsic profiles. The lines are the best fit Moffat profiles in each case.

where r_{hl} is the half-light radius, β is the Sérsic index. To save time, we use the existing catalogue products to calculate the half-light radii, rather than going back to the images (Liske et al. 2003, e.g.). The half-light radii are calculated using the circular aperture radii measurements of the flux, which give a curve-of-growth. We use the 13 aperture fluxes measured by the VDFS extractor, at radii of ($0.5, 0.5\sqrt{2}, 1, \sqrt{2}, 2, 2\sqrt{2}, 4, 5, 6, 7, 8, 10, 12$ arcseconds). To calculate the half-light radius, we first find the aperture flux closest to half the total-flux and then use the five apertures centred on this (2 before, 2 after and the aperture in question). Using these 5 aperture fluxes, we fit a quadratic, which removes any small bumps in the curve, using the singular value decomposition method (Golub & Reinsch 1970). We find the root of the quadratic that gives the half-light radius (r^c_{hl}, **hlCircRadAs**) . We use the covariance matrix to calculate the error in the half-light radius ($\sigma_{r^c_{hl}}$, **hlCircRadErrAs**), adding in another half-pixel in quadrature, to take into account the granularity of the data.

The 13 aperture fluxes are all circular apertures, so the half-light radius calculated assumes a circular symmetry. However, most galaxies are elliptical in shape, either being triaxial spheroidal systems or inclined disks or a combination of the two. Therefore a semi-major axis size (Driver et al. 2005), to take out the inclination of disks, or geometric mean size for elliptical galaxies (e.g. see Binney & Merrifield 1998) are more useful. Figure B.1 shows the ratio of half-light semi-major axis to half-light circular radius as a function of ellipticity for a range of Sérsic profiles ($\beta = 0.5$ to $\beta = 7$): profiles which are a good fit to the vast majority of virialised galaxies. As can be seen, the variation between these profiles is around 1 – 2% for all ellipticities < 0.9, where it rises to 10%. In addition these curves are well fit by Moffat functions,

$$f(\epsilon) = \frac{a}{(1 + (\frac{1-\epsilon}{b})^2)^c} \tag{B.2}$$

where a, b & c are found from fitting the data for each profile.

Thus, we can convert our circular half-light radii r^c_{hl} to a semi-major axis size r^{smj}_{hl} by using the $\beta = 2$ function ($a = 1.8243, b = 0.30914, c = 0.24304$). We choose $\beta = 2$, since most galaxies will be ellipticals ($\beta = 1$), de Vaucouleurs ($\beta = 4$), or dominated by the PSF ($\beta = 0.5$), and $\beta = 2$ is nicely in the middle, but as Fig B.1 shows, there is very little difference.

The conversions to the half-light semi-minor axis, r^{smn}_{hl}, and half-light geometric mean, r^{geo}_{hl}, are easily computed from the geometry of an ellipse:

$$r^{smn}_{hl} = (1 - \epsilon) r^{smj}_{hl} \tag{B.3}$$

$$r^{geo}_{hl} = \sqrt{r^{smn}_{hl} r^{smj}_{hl}} \tag{B.4}$$

Finally, we take into account the effects of seeing. We use the method of Driver et al. (2005) to subtract the measured seeing, assuming a Gaussian PSF,

$$r^{smj,see}_{hl} = \sqrt{r^{smj^2}_{hl} - c_{see}\Gamma^2}, \tag{B.5}$$

where Γ is the full-width half maximum of stars in the image and c_{see} is a constant, 0.5 for a Gaussian PSF. By experiment we found values of $c_{see} \sim 0.45$, but with quite a large uncertainty, so we took the theoretical value $c_{see} = 0.5$.