Contents

Summary Science Case

1.1 INTRODUCTION

- Continuing importance of survey astronomy
- Relevance to STFC priorities
- Research data management policy framework
- Supporting multiwavelength research consortia

1.2 REPORT ON RECENT WORK: 2009 – 2012

1.2.1 Achievements: continuing successful operation of the WFCAM and VISTA Science Archives

1.2.2 Achievements: development of the OmegaCAM Science Archive

1.2.3 Achievements: development of the Gaia-ESO Spectroscopic Survey Archive

1.2.4 Achievements: development of a new, VO-enabled archive infrastructure

1.2.5 Achievements: UK involvement in LSST and Euclid

1.3 PROPOSED PROGRAMME OF WORK Q2 2013 – Q1 2016

1.4 KEY DELIVERABLES and MILESTONES

Milestones:

QN 201N: Milestone 1

QN 201N: Milestone 2

Deliverables:

QN 201N: Deliverable 1

QN 201N: Deliverable 2

1.5 RESOURCES REQUESTED

(a) *Staff.*
(b) *Travel and subsistence.*
(c) *Consumables:*
(d) *Maintenance:*
(e) *Equipment:*

1.6 ADDITIONAL REFERENCES

WP1: Imaging Surveys

Staff involved

A. Person
A. N. Other

2.1 INTRODUCTION

2.2 WP1.1: WFCAM SCIENCE ARCHIVE

2.2.1 Report on recent work: 2013 − 2015

Achievements: data releases

Since 2013 there has been one new science ready UKIDSS data release, DR 10 (3.7 TB). Initially released in Q1 2013, GPS catalogue data were added in Q1 2015 following re-processing. DR 10 incorporates database products from observations taken up to and including the end of Semester 11B: we emphasise that flat file images, quicklook jpegs and individual passband detection catalogues become available as soon as monthly pipeline processing is completed, so, at the time of writing, all data up to and including February 2015 are available through the WSA. Operations staff have also made survey-like prepared database releases for 31 private PI programmes (in addition to enabling flatfile access for all registered projects). Static database products (UKIDSS data releases and survey-like databases for PI programmes) are made world public once the proprietary periods (18 and 12 months, respectively) are passed. The latest world accessible UKIDSS releases are DR 8 (for GPS) and DR 10; World-accessible releases are also published to the international Virtual Observatory via infrastructure developed by the AstroGrid project and maintained at WFAU.

Achievements: archive usage

More than 1050 individual users are currently registered for proprietary (18 month) UKIDSS database access in the WSA. These users are distributed over 105 distinct institutions, more than half of which are outwith the UK. Furthermore, 312 private PI programme registrations allow access by small PI-led teams to their proprietary (12 month) private datasets.

Figure 2.1 illustrates the level of archive usage through the WSA website over the reporting period (NB: these statistics exclude all Edinburgh access to avoid testing activities skewing the results). The WSA provides two distinct access modes: traditional' flat file access to standard data products (i.e. pipeline-processed image/catalogue FITS files and JPEG compressed images) for casual browsing and external QC; and flexible Structured Query Language (SQL) access to tabular datasets (mainly seamless, merged source lists) in prepared, static database products. The top panels in Figure 2.1 show flat file numbers and volumes. In terms of data volume, a download rate of around 1 TB/month over the past year is observed. The lower panels of Figure 2.1 show SQL activity. On the left, the number of queries from freeform' SQL, crossID and menu-driven queries input into the WSA web forms, broken down by data release type, shows a sustained high level of usage with >50,000 queries in the past year. Most impressively, on the right those same queries but plotted in terms of rows returned shows that billions of rows of tabular data are being extracted by our users year on year. In addition to the usage shown here, WSA UKIDSS releases are also queried thousands

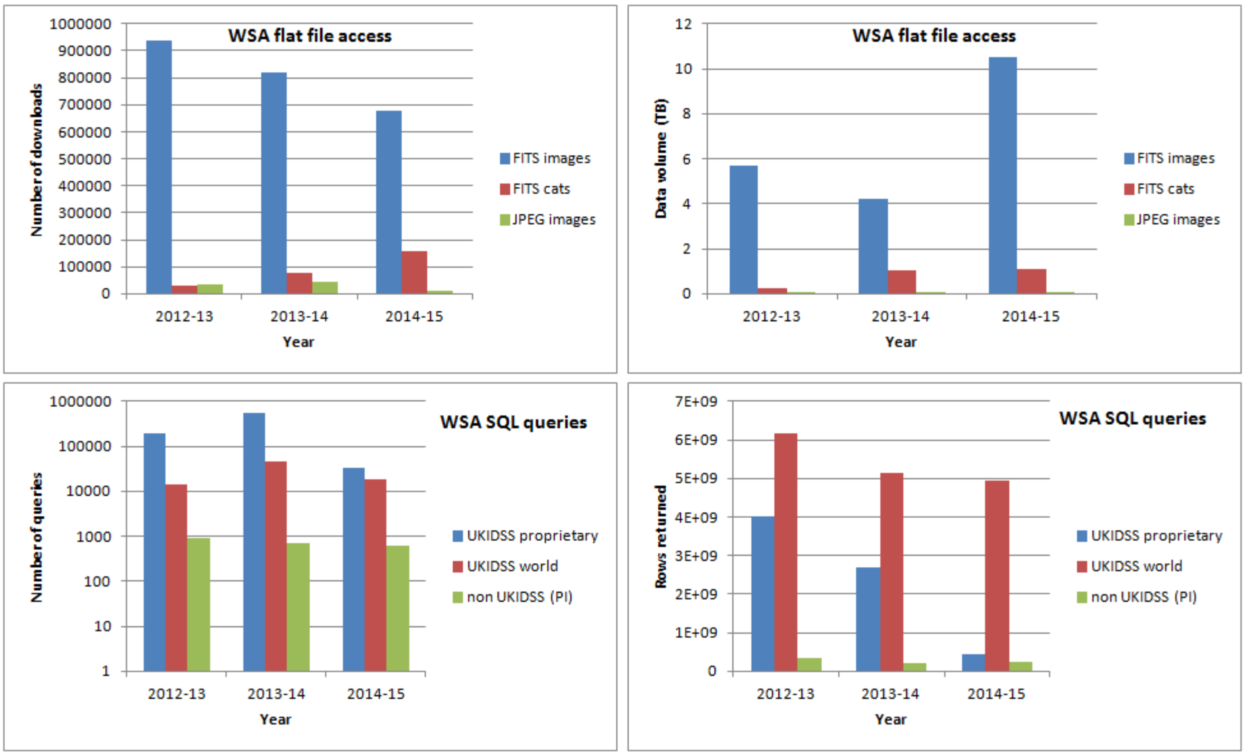

Figure 2.1: *WSA usage activity over the last three years. Top panels: flat–file access (image/catalogue files and compressed JPEG images); bottom panels: SQL queries.* UKIDSS proprietary *refers to queries performed by registered UKIDSS/ESO users,* UKIDSS world *shows queries to world accessible releases and* non UKIDSS *records queries to databases of private PI programmes.* N.B. *these statistics exclude* all *Edinburgh access to avoid testing activities skewing the results.*

of times a month from the Virtual Observatory.

Up to the end of 2012 when the consortium stopped maintaining records, UKIDSS had generated over 400 published papers. In the past 3 years 87 published papers have cited the reference WSA paper (Hambly et al., 2008, MNRAS, 384, 637).

A helpdesk system is operated to support users of the archive. Over the reporting period, $\sim$ 200 requests for support have been received. A response is normally made within a few hours of receipt during normal working hours. Any responses that are felt to be more widely useful are put on the static Q&A web pages while general SQL query solutions are included in the SQL cookbook maintained online. Sometimes users make requests for small changes and/or enhancements to archive user interface functionality. Maintaining the website and enhancing interface functionality have been further operational activities.

Administration of the IT infrastructure on which the WSA is built is a critical aspect of operations. During the reporting period, operations staff have maintained and expanded the large RAID arrays necessary for online storage; expanded and maintained the network infrastructure; maintained and upgraded the archive software including patches and updates to thirdparty software; and, finally, have operated the system backup policy. Maintenance of the overall hardware systems has also included periodic upgrading of PC servers and LTO tape devices. Archive downtime is recorded for the WSA - over the period Q4 2012 to Q3 2015 (1095 days) archive downtime is less than 20 days, i.e. archive availability has been $>$ 98%.

2.2.2 Proposed programme of work Q2 2016 – Q1 2019

WSA enhancements

Any further/final UKIDSS/UHS releases could be enhanced by implementing the following

1. include additional columns developed for the VSA/OSA in the detection tables, e.g. half light radii and separating the various photometric corrections.
2. improved error bit flagging e.g. around bright stars.
3. deep stacking of multi-epoch data e.g. LAS J epochs.
4. addition of HEALPix indices for improved coordinate searching and MOC usage.
5. improved user–interface based on the OSA prototype.
6. improved photometry of extended extragalactic sources.

The improved photometry of extended extragalactic sources is an issue because many extragalactic astronomers have been unhappy with the CASU processing of data in surveys such as LAS and DXS. The main issues are to do with sky-subtraction and extractor parameters. The UKIDSS-UDS team, VISTA-VIDEO team and VISTA-UVISTA team all make their own SWARP mosaics and extract using SExtractor, after a sky-subtraction process that involves object masking. We already have a pipeline that will run SExtractor on mosaics and convert mosaics to the WSA/VSA standards, and are designing a matched aperture pipeline to process mosaics across a wide wavelength range using the SExtractor dual image mode, see § **??** and § **??**. Additional components to this system would involve a sky-subtraction module, using object masking, and automation of the SWARP mosaicing software, or usage of similar data processing pipelines such as . . . to generate the mosaics. Storage can become an issue since the Terapix/Astromatic software such as SWARP and SExtractor do not work on Rice compressed images in the same way as CASU software.

2.3 WP1.2: VISTA SCIENCE ARCHIVE

2.3.1 Report on recent work: 2013 − 2015

Achievements: first achievement

WFAU have made consortium releases for VHS, VMC, VIKING and VIDEO up to the end of semester P93 (and half of P94 in the case of VIDEO). We have made releases for each of these surveys within a few months after the end of each semester, apart from VIDEO, where we have done releases when the VIDEO team have produced new deep mosaics.

WFAU have also produced releases for the VVV team. The much greater catalogue data volume (>6 times the data volume of VHS, the next largest) and the fact that most VVV data is taken over a few months in odd semesters has led us to aim for yearly releases. Substantial changes to the data structure and software have been necessary to more efficiently process the VVV, so that there been release of data up to P87 (VVVDR1), P89 (VVVDR2) and very shortly (October 2015) there will be a release up to P91 (VVVDR3) a 45TB database, one of the largest current astronical SQL databases.

These consortium releases are typically made publically available at the same time as the data are made public through the European Southern Observatory Science Archive Facility (ESO-SAF). We convert various tables (or a subset of the columns in some tables) of specified releases (typically one each year) to a set of FITS files formatted so that ESO can ingest them. These ESO releases contain only VISTA data - no cross-matched tables, and usually only contain tile data.

WFAU have also made data releases of several PI programmes. The releases are summarised in Table **??**.

During this time period there have been various improvements to the software to improve efficiency, correct errors and add features. This has included the splitting of the `vvvDetection` table into separate tables for each month during curation, improvements to the design of queries that have significantly increased the efficiencies of bulk outgests detections in the processing of variability data and copying to a static release database and the parallelism of various processes such as image processing and data outgests and ingests. These improvements have been necessary for the VVV and have speeded up the processing of other surveys too.

A substantial effort over the last few years has been the development of code to convert data in our SQL tables to FITS tables for transfer and ingest into the ESO-SAF. During this time there has been various changes to

Table 2.1: Summary future VISTA Public Survey VSA releases.

Database Name	Size (MB)	Release Date	Date Made Public
Public Surveys			
VHSDR1	951,235	22/02/2012	22/02/2012
VHSDR2	2,044,908	24/02/2014	24/02/2014
VHSDR3	4,174,747	10/04/2015	10/04/2015
VHSv20120926	2,278,399	26/09/2012	N/A
VHSv20130417	2,683,071	17/04/2013	N/A
VHSv20140409	3,142,575	09/04/2014	N/A
VHSv20150108	5,089,479	08/01/2015	N/A
VIDEODR2	169,044	28/03/2012	28/03/2012
VIDEODR3	198,326	12/07/2012	26/02/2014
VIDEODR4	518,977	27/10/2014	10/04/2015
VIDEOv20100127	3,874	27/01/2010	N/A
VIDEOv20100429	8,416	29/04/2010	N/A
VIDEOv20100513	15,237	13/05/2010	N/A
VIDEOv20111208	70,569	08/12/2011	N/A
VIDEOv20141027	518,977	27/10/2014	N/A
VIDEOv20150521	341,404	21/05/2015	N/A
VIKINGDR2	74,181	19/10/2011	28/03/2012
VIKINGDR3	410,795	05/12/2012	16/12/2013
VIKINGDR4	605,658	30/05/2014	10/04/2015
VIKINGv20100127	162,816	27/01/2010	N/A
VIKINGv20100429	19,429	29/04/2010	N/A
VIKINGv20110414	259,490	14/04/2011	N/A
VIKINGv20110714	265,062	14/07/2011	N/A
VIKINGv20111019	316,897	19/10/2011	N/A
VIKINGv20130417	476,788	17/04/2013	N/A
VIKINGv20140402	480,903	02/04/2014	N/A
VIKINGv20150123	720,912	23/01/2015	N/A
VIKINGv20150421	808,235	21/04/2015	N/A
VMCDR1	84,713	28/03/2012	28/03/2012
VMCDR2	87,523	25/02/2014	25/02/2014
VMCDR3	329,884	29/10/2014	29/10/2014
VMCv20110816	343,947	16/08/2011	N/A
VMCv20110909	446,738	09/09/2011	N/A
VMCv20120126	558,566	26/01/2012	N/A
VMCv20121128	697,396	28/11/2012	N/A
VMCv20130304	778,043	04/03/2013	N/A
VMCv20130805	1,170,697	05/08/2013	N/A
VMCv20140428	1,181,768	28/04/2014	N/A
VMCv20140903	1,669,029	03/09/2014	N/A
VMCv20141118	1,890,803	18/11/2014	N/A
VMCv20150309	1,935,294	09/03/2015	N/A
VVVDR1	9,616,068	29/05/2012	29/05/2012
VVVDR2	16,967,950	24/02/2014	29/10/2014
VVVv20100127	29,152	27/01/2010	N/A
VVVv20100531	101,518	31/05/2010	N/A
VVVv20110718	1,268,968	18/07/2011	N/A
PI programmes			
D284A5026Av20120229	932	N/A	N/A
D284C5034Av20130927	246,649	N/A	N/A
N088A0728av20150106	10,848	N/A	N/A
N089C0102v20130417	41,381	N/A	N/A
N089D0113v20131129	126,284	N/A	N/A
N090A0570v20150107	21,187	N/A	N/A
N090A0570v20150118	100	N/A	N/A
N091A0426v20141126	6,942	N/A	N/A
VSERV1v20150502*	197,323	N/A	N/A

All current VSA data releases. The sizes include the data, indices and statistics to help optimise searches and sometimes additional space from transaction logs and excess space in filegroups. * denotes a combined PI programme made from two projects over different semesters observed by the same PI.

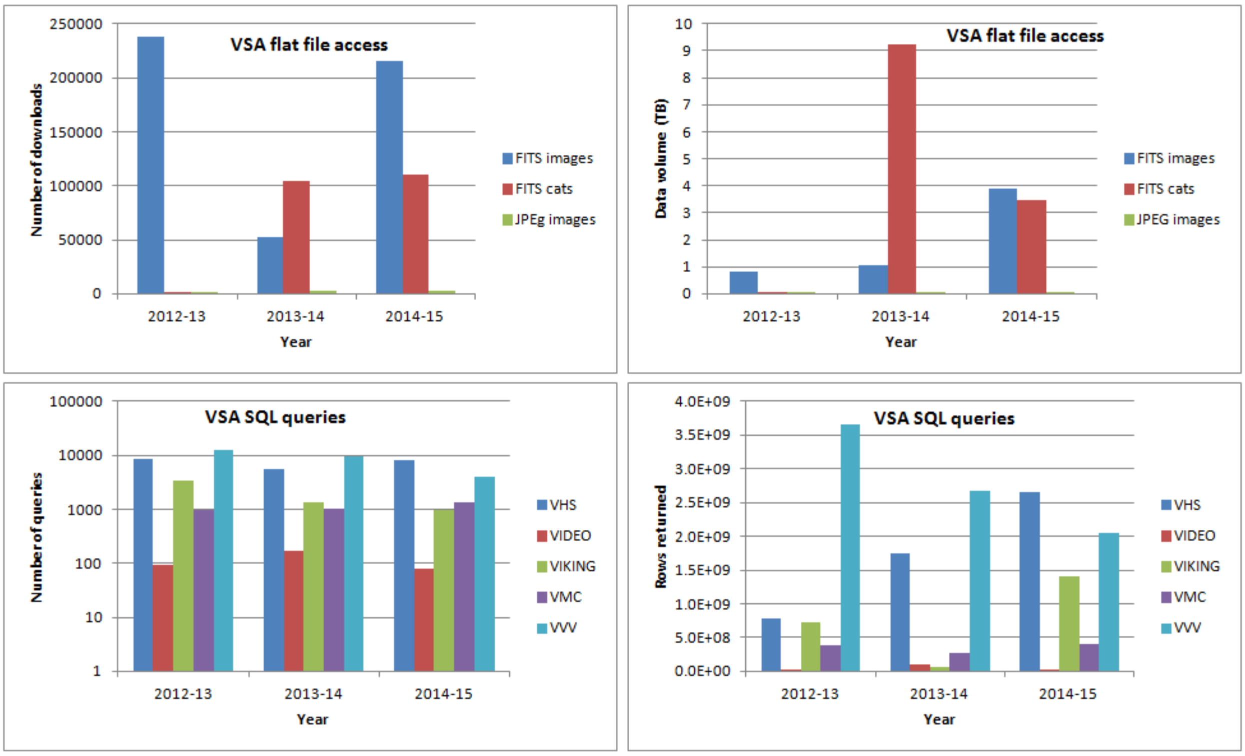

Figure 2.2: *VSA usage activity over the last three years. Top panels: flat–file access (image/catalogue files and compressed JPEG images); bottom panels: SQL queries.* N.B. *these statistics exclude* all *Edinburgh access to avoid testing activities skewing the results.*

the requirements, including temporary features, leading to much additional development. However, in the last year the requirements seem to be settled, so there should be less development, just what is required to make sure that both the VSA and OSA data are processed correctly and changes for new features such as new 'value added' data or specific requirements for new surveys in the recently announced VISTA extension.

The VISTA Science Archive has also started incorporating small amounts of team generated data ('value added') in addition to the images and catalogues produced by the CASU nightly pipeline and the WFAU generated products. These have included PSF and variable star catalogues generated by the VMC team ([?],[?]) and 3D extinction maps generated by the VVV team, [?]. The VMC products have been linked to the main source table via simple neighbour tables in the VSA and converted for release to ESO. The 3D extinction map has been incorporated so that they can be used in complex joins [?] and an infrastructure set up so that other 3D extinction maps can be easily incorporated in the future.

During this period, a generalised Matched Aperture Pipeline has been designed and much of the basic infrastructure has been tested [?]. This will be finished during the next grant cycle.

WFAU have also included Multi-Order Coverage Maps (MOCS, [?]) which allow users to overlay the footprints of surveys, both VSA and external surveys. We are also been incorporating more external surveys, such as WISE and AKARI.

The bottom panels in Figure **??** show flat file numbers and volumes. In terms of data volume, a download rate of around 1 TB/month over the past year is observed. The upper panels of Figure **??** show SQL activity. On the left, the number of queries from each survey shows a sustained high level of usage across all surveys, with. Most impressively, on the right those same queries but plotted in terms of rows returned shows that billions of rows of tabular data are being extracted by our users year on year. In addition to the usage shown here, VSA releases are also queried thousands of times a month from the Virtual Observatory.

2.3.2 Proposed programme of work Q2 2016 − Q1 2019

New VISTA Surveys

On 29th July 2015, ESO have announced a new call for Public Surveys on VISTA, with a deadline of 15th October 2015 for Letters of Intent (LoI). The decision will be made in December 2015 and the first observations will start from October 2016.

Given the timescale above, we cannot be sure of exactly who will be submitting LoIs and certainly cannot know what new surveys will be given time. Having said that, several of the existing VISTA Public Survey teams, such as VVV and VMC have contacted us about plans for extensions and have included WFAU in their data management plans. Others such as VIDEO and VIKING have expressed interest, and may well submit a LoI too. Some new teams have also been in contact and may well submit LoIs too, and would like CASU/WFAU to do the data management through VDFS.

Extensions pose more of a data management problem that new surveys since the releases are cumulative, i.e. VHSDR3 contains all the data in VHSDR2 and all the new data taken subsequent to VHSDR2. If the VVV team, who are planning a very substantial extension are given all of their time, this will have significant additional hardware costs and processing time costs. Some of these costs may be reduced by splitting the survey into regions, but this may not be viable since the expected final proposal will have a contiguous area.

From the proposals that we have seen so far we do not expect substantially additions to software for new surveys or extensions to new surveys. We also do not expect substantial changes to the CASU software that would require the data to be reprocessed and retransferred and reingested, apart from VVV tile catalogues where improvements to the pawprint zeropoints from modifications to the extinction corrections will lead to wholescale reprocessing over the next year or so.

VSA Enhancements

During the 2016-2019 period, we expect to make various enhancements to the VSA. These include:

- A Matched Aperture Photometry Pipeline
- Enhancements to the multi-epoch pipeline, especially variability statistics

The matched-aperture photometry pipeline has already undergone the main design phase and much of the basic single-survey source table option has undergone testing as have some parts necessary for multi-survey curation. However, much more testing and development is necessary to create a pipeline that will run both the CASU list-driven extractor and SExtractor dual-image mode in the following ways and present the data in the most useful formats:

- Matched aperture photometry across all filters within the survey. This is useful for Spectral Energy Distributions of galaxies and for low-signal-to-noise detections in some bands.
- Matched aperture light-curves. This has been requested by the VMC team.
- Matched aperture photometry across different surveys (e.g. optical-NIR) to get better SEDs.
- Using matched aperture photometry to improve photometry in a single filter where complex corrections need to be applied, e.g. VISTA tiles, VPHAS+
- User defined, to replace broken WSA-only on-demand user photometry tools, but will work in WSA/VSA and OSA.

We expect the total development time for the matched aperture pipeline to be 9 months to 1 year FTE. However, adding in more complicated features such as deconvolution or segmentation to use data with different seeing could increase this.

The enhancements to the multi-epoch pipeline are designed to either improve curation efficiency, or to improve the selection and classification of variable stars. These enhancements include:

- Rearrange multi-epoch processing pipeline so that detection data is only outgested once. This should make a significant performance improvement for the VVV.
- Add in indices specified in Fernandez Lopes & Cross (2015) and Fernandez Lopes & Cross in prep. The indices in the first paper have been shown to be much more efficient at discriminating real variables from noisy data.
- Multiple variability tables. These are useful in a few cases e.g. VVV, both aperture photometry and difference imaging photometry are available; UKIDSS-UDS/VIDEO where OB frames could provide very many epochs for bright sources and medium deep mosaics could provide fewer epochs for many fainter sources.
- Proper motion determination. This may include absolute proper motion determination in regions where reliable QSO catalogues of sufficient source density exist.
- Improved transient selection, i.e. improved indices to discriminate transients from noise when there are few data points.

We expect all the multi-epoch processing to take 8-10 months FTE.

In addition to developing software for new VSA products, we will also be ingesting an increased amount of team generated 'value added' data. In particular over the coming few years we expect to ingest:

- VVV Point spread function (PSF) and Difference Image Analysis (DIA) data. The single epoch multi-band PSF data will be used to calibrate the multi-epoch Ks light-curves. Over the last few months we have developed a draft design of the schema and the necessary processing steps for these data, which will reuse as much of our existing software as possible, such as the ingest code, quality bit flagging code, matching up of epochs, band-pass merging and variability statistics codes.
- VVV Proper motion and parallaxes. There are existing columns in the VVV variability table for proper motion and parallax, which we will try to use if possible.

We expect that the new value added data products will require new development which will take about 4 months FTE. There will also be additional storage requirements for difference images and catalogue data products, the latter of which could have a significant impact on the SQL database storage. The DIA catalogue data should have a much smaller data volume than the main detection data, since only objects that have changed between the master and epoch frames should be included. The PM/parallax data should also be relatively small, since these are a single solution for each source, but if the PSF data is expanded beyond one or two epochs it could increase the database sizes by at least 50% and perhaps more. The hardware requirements are described in more detail in WP4.

Required Releases

The VVV dominates all estimates of future release database sizes producing catalogue data at ¿6 times the rate of the VHS. We can estimate sizes of future release by extrapolating the rate of change of table sizes in existing releases and the estimated time of completion of each survey and for some of the extended surveys, the additional time that will be applied for. The existing surveys will all be reaching completion over the next grant: VVV 2016, VHS (2019), VMC (2018), VIKING (2017), VIDEO (2018). We have included extension information for VVV and VMC since they have made their plans known already. The other surveys may or may not have extensions. In Table **??**, we summarise the future public survey releases that we expect in the VSA in the period 2016-2020.

Given the new call for surveys, with surveys continuing until 2020 and releases every 6 months for the current surveys except VIDEO and VVV, which will be yearly, we have the following possibilities of increases beyond completing the existing surveys, which requires $\sim$ 220TB of storage space:

- Option 1: Completely new surveys. The average date range will be 2018 - 2021, about half the length of existing. These would be completely separate databases, so no cumulative data, and are more likely to be medium depth surveys, or have specialist filters. Total additional release data on top of first round completion is ¡30TB.

Table 2.2: Summary future VISTA Public Survey VSA releases.

Survey	Release	Estimated Date	Database Size (TB)		
			Original survey only	Extension	Extension + ValueAdded
VVV	P91	End 2015	45	—	—
VVV	P94	End 2016	55	—	—
VVV	P96	End 2017	65	70	100
VVV	P98	End 2018	75	80	120
VVV	P100	End 2019	—	90	135
VVV	P102	End 2020	—	100	150
VHS	P95	2016A	5.1	—	—
VHS	P96	2016B	5.7	—	—
VHS	P97	2017A	6.3	—	—
VHS	P98	2017B	6.9	—	—
VHS	P99	2018A	7.5	—	—
VHS	P100	2018B	8.1	—	—
VHS	P101	2019A	8.9	—	—
VHS	P102	2019B	9.5	—	—
VHS	P103	2020A	10.	—	—
VMC	P95	2016A	1.5	—	—
VMC	P96	2016B	1.7	—	—
VMC	P97	2017A	1.8	1.9	—
VMC	P98	2017B	2.0	2.2	—
VMC	P99	2018A	2.1	2.4	—
VMC	P100	2018B	2.3	2.7	—
VMC	P101	2019A	2.4	2.8	—
VMC	P102	2019B	2.6	2.9	—
VMC	P103	2020A	2.7	3.0	—
VIKING	P95	2016A	1.0	—	—
VIKING	P96	2016B	1.2	—	—
VIKING	P97	2017A	1.4	—	—
VIKING	P98	2017B	1.6	—	—
VIDEO	P96	2016B	0.3	—	—
VIDEO	P98	2017B	0.4	—	—
VIDEO	P100	2018B	0.5	—	—
VIDEO	P102	2019B	0.6	—	—
VIDEO	P104	2020B	0.7	—	—

The 3 columns reflect uncertainty due to expected additional extension data (because this time may or may not be awarded) and the VVV also has additional uncertainty because of a possible massive increase in team generated data in the form of new PSF/DIA photometry.

Release	Rel.Date	Semesters	Calibration	list-driven cats.
ATLASv20130304	2013-03-04	COMM	ESO standards	-
ATLASv20130426	2013-04-26	COMM - P89part	ESO standards	-
ATLASv20131029	2013-10-29	COMM - P90	ESO standards	-
ATLASv20131127	2013-11-27	COMM - P90	ESO standards	-
ATLASDR1	2013-12-17	COMM - P89	ESO standards	-
ATLASDR2	2015-03-05	COMM - P91	APASS	yes
ATLASDR3	2015 Q4	COMM - P93	APASS	yes

Table 2.3: OSA Releases

- Option 2: A mixture of extensions and new surveys. VVV likely to get an extension, e.g. half the time they request. Releases will include original VVV data too and the VVV extension would start earlier than some others, around 2016. New surveys ¡20TB in total. VVV+VVVX will add an extra ∼ 240TB from initial release in 2018. This includes some additional space required for difference imaging data, 1-2 epochs of multi-band PSF photometry
- Option 3 Worst-case scenario Like option 2, but VVVX gets all requested time. VVV+VVVX will add an extra ∼ 270TB from initial release in 2018.
- Option 4 Worst-case scenario, but with massive additional value added data. E.g. PSF photometry for all epochs. PSF photometry has a much higher source density than aperture photometry in regions which are confusion limited, but careful choice of columns can limit the increase to a 50% increase in database size. This would lead to a ∼ 400TB increase.

The most likely scenario in our opinon is either option 2 or option 3, so the total database hardware requirements in this period are ∼ 470TB (220TB to complete the existing surveys and another 250TB for the 2nd round of public surveys).

2.4 WP1.3: OMEGACAM SCIENCE ARCHIVE

2.4.1 Report on recent work: 2013 − 2015

The Omegacam Science Archive (OSA[1]) is the third archive included in the VISTA Data Flow System project (VDFS). Pipeline processing is handled at the Cambridge Astronomy Survey Unit (CASU), which adds in addition to the standard products now also list-driven photometric catalogues. The data is transfered and then ingested into a ingest database from which static release databases are created.

In the first processing version by CASU the data was calibrated using ESO standard fields observed by VST each night. Afterwards the data was recalibrated using the APASS (American association of variable star observers (AAVSO) All Sky Survey) catalogue which reduces the scatter between adjacent fields and gives a calibration comparable to Pan–STARRS. This forced a re-transfer of all data from CASU to WFAU and the creation of a new ingest database to have both calibration methods available for comparison.

Achievements: data releases

Beginning in Q1 2013, the OSA operations staff have prepared 6 science–quality data releases for the ATLAS consortium. In Q4 2015 two more releases are in preparation: one for the ATLAS consortium and a first data release to the ESO–SAF, since previous ESO–SAF releases were made from CASU without the inclusion of passband merged catalogues.

Achievements: inclusion of additional data

Including list-driven catalogues resulted in more data to be stored and processed. Additional code has been developed and procedures have been put in place to ingest these.

[1] http://osa.roe.ac.uk

Another difference to the data products from VISTA is the application of illumination corrections in the CASU pipeline. Whereas for VISTA simple monthly illumination correction tables are made available by CASU and applied during photometry calculations at WFAU, the procedure for VST data is more convoluted: the illumination corrections are applied to the fluxes at CASU and are not accessible to the users. Since the ATLAS consortium and users requested it, we developed code to reverse calculate the corrections applied per detection and made it available via the database tables. This involves the storage of uncorrected catalogues in addition to the corrected ones in the database.

Achievements: new user interface functionality

A new user interface has been developed for the OSA to enhance ease of use and, in the long term, interoperability of WFAU Science Archives with all external databases registered in the IVOA (International Virtual Observatory Alliance[2]). As the testbed for these developments the OSA was chosen.

To standardise queries on all databases in IVOA the Astronomy Data Query Language (ADQL) was chosen. In a first stage the previously used freeform SQL query interface has been rewritten to be based on ADQL. This provides additional features like plotting the query results or viewing the table metadata. The next stage is the development and implementation of FireThorn which allows users to combine data from local and remote archives into a virtual dataset (see Section 4.2.2).

2.4.2 Proposed programme of work Q2 2016 − Q1 2019

Inclusion of VPHAS+

Since we transfer all the data that CASU processes and now already have the infrastructure in place to process VST data, we are able to host another of the VST surveys, VPHAS+ (VST Photometric Hα Survey of the Southern Galactic Plane and Bulge[3]). Minor developments are still needed and additional time to process releases needs to be budgeted for as well as standard operations time to process all VST observed in the time frame in question.

Future Releases

ATLAS will observe until at least the end of P95 (2015-09-30), which means at least one more consortium release. Depending on the integration of observations done under Chilean Time, which might contribute to ATLAS and extend post P95, there is the possibility of further yearly releases combining these data sets with ATLAS. Since the calibration is not final and efforts are under way already to enhance it, a reprocessing of the complete data set is highly likely. This will result in the re-transfer, storage and ingest of the complete data set as well as in a final database release for ATLAS.

If discussions with the VPHAS+ PI are conclusive, we might be asked to produce releases on a yearly schedule. Since VPHAS+ is observed until 2017, this would mean at least 2 more releases have to be prepared (again plus a final one in case of re-calibration).

Each ATLAS release is about 1.5TB in size while each VPHAS+ release is about 3-4TB in size.

Integration of VST optical data with VISTA near-IR data

VST optical and VISTA near-IR surveys are very compatible, in terms of area surveyed (VST-ATLAS & VISTA-VHS; VST-VPHAS+ & VISTA-VVV; VST-KiDS & VISTA-VIKING) and many science goals rely on estimates of temperature, mass, star-formation history from combined optical near-IR spectral-energy-distributions. For stars, a neighbour table is usually sufficient and given significant proper motions is often ideal, but for extended galaxies it is necessary to be sure that the same part of the galaxy is being measured at each wavelength, so matched aperture photometry is best (see § refvsa:soft).

[2] http://www.ivoa.net/

[3] http://www.vphasplus.org/

2.5 WP1 DELIVERABLES and MILESTONES

Milestones:

QN 201N: Milestone 1

QN 201N: Milestone 2

Deliverables:

QN 201N: Deliverable 1

QN 201N: Deliverable 2

2.6 RESOURCES REQUESTED

(a) *Staff.*
(b) *Travel and subsistence.*
(c) *Consumables:*
(d) *Maintenance:*
(e) *Equipment:*

2.7 REFERENCES

WP2: Spectoscopic Surveys

Staff involved

R. S. Collins
A. C. Davenhall

3.1 INTRODUCTION

TBD: RGM. ESO SAF will be file repository, but need more flexible archive for spectroscopic surveys to facilitate survey science, just as for imaging surveys

3.2 WP2.1: GAIA-ES0 SPECTROSCOPIC SURVEY ARCHIVE

3.2.1 Report on recent work: 2013 – 2015

Achievements: first achievement

3.2.2 Proposed programme of work Q2 2016 – Q1 2019

3.3 WP2.2: MOONS SCIENCE ARCHIVE

The Multi-Object Optical and Near-infrared Spectrograph (MOONS)[1] is currently being designed and constructed for the ESO VLT. It will have a wavelength coverage of 0.6-1.8μm (corresponding to the red and infrared regions of the spectrum) and either medium ($R \sim 4000 - 6000$) or high ($R \sim 9000$ or $R \sim 20,000$) resolution modes. It will have ~ 1000 fibres deployable on target objects. MOONS is scheduled to start operations during 2019, when it will become the long-anticipated 'workhorse' near-infrared multi-object spectrograph for the VLT. It is likely to prove a popular and widely-used instrument.

Design and construction of MOONS is being overseen by a Consortium of interested groups and individuals. Though MOONS will be a common-user instrument the MOONS Consortium will be allocated a substantial amount of observing time, which they plan to use to conduct a series of coordinated surveys. WFAU seeks to work with the Consortium to facilitate these surveys. Specifically it wishes to contribute in two ways: (i) by providing assistance with target selection and (ii) by designing, building and operating a science data archive for the Consortium's surveys.

MOONS is a near-infrared spectrograph. Thus, the infrared VSA and WSA archives operated by WFAU are natural sources from which to select target objects for it. WFAU will work with the MOONS Consortium to ensure that the VSA and WSA are used effectively to find targets for MOONS in a timely fashion. WFAU also hopes to work with the Consortium to design, build and operate a science data archive for MOONS Consortium surveys, by adapting the GES science data archive (WP2.1, above).

MOONS will not start operating until 2019 and planning and preparation for the MOONS work are necessarily at an early stage.

[1] See URL: `http://www.roe.ac.uk/∼ciras/MOONS/VLT-MOONS.html`

3.3.1 Report on recent work: 2013 − 2015

Discussions between WFAU and the MOONS Consortium are in progress. Currently it seems highly likely that the Consortium will invite WFAU to provide the science data archive for their surveys, but it has not yet formally done so. MOONS will not start operations until 2019 and it is premature to start target selection. However, WFAU has written a note summarising which of its data-holdings are likely to be useful for this purpose.

Achievements: target selection note

The note *MOONS Target Selection from WFAU Surveys* (Davenhall *et al.*, 2015) was successfully completed and made available to the MOONS Consortium.

3.3.2 Proposed programme of work Q2 2016 − Q1 2019

As discussed above, the planned programme of work for the MOONS Consortium comprises two items (i) assistance with target selection and (ii) designing, building and operating a science data archive. These items are discussed separately below.

Target selection The work here mostly involves giving advice and assistance to members of the MOONS Consortium on using the WFAU data holdings, principally the VSA and WSA archives, to select target objects for subsequent observation by MOONS. The Gaia data archive is also likely to be important for planning MOONS observations and since WFAU is involved in the design and construction of this archive it is well-placed to include assistance on using it. This activity is not dissimilar to the advice that WFAU has traditionally offered to users of its archives, but it is more focused, and perhaps involves acquiring some expertise in the particularities of the MOONS instrument. Some limited software development, or more likely customisation, may be required to provide convenient, bespoke tools for finding MOONS targets, but no major effort is foreseen.

Although this work will be undertaken to facilitate target selection by the MOONS Consortium it will be equally useful to other users of MOONS. Since MOONS is expected to prove a popular instrument it will make a significant contribution to maximising the use of the VSA and WSA archives. Similarly, it may have at least some relevance to target selection for other spectrographs, though most of the other instruments that are operational or planned are optical and less closely matched to the VSA and WSA infrared surveys.

MOONS science data archive This work will be undertaken if, as seems highly likely, WFAU is invited to develop the science data archive for the MOONS Consortium. The work involves designing, building and operating this archive. MOONS will not become operational until later in 2019, after the period covered in this application. Consequently here we only request resources to design and build the archive. However, it should be noted that there may be some limited use of the archive before the instrument becomes operational, to hold, for example, lists of target objects and lists of atomic and molecular line data.

There are two aspects to the archive design: (i) the tables that it is to contain and the relations between them and (ii) the ingest procedures (in practice a collection of python scripts) to import the calibrated observations produced by the reduction pipeline into the archive. We envisage a design based on the one developed for GES (see WP2.1, above), which is itself based on those for the VSA and WSA. Hence we will be re-using and capitalising on the substantial investment and expertise in these archives.

As for GES, fully-reduced one-dimensional spectra in the archive will he held as files in the standard astronomical FITS format stored in a Unix directory structure. The metadata for these files, including the astrophysical parameters (such as the effective temperature and surface gravity) and elemental abundances derived from them, will be held as a series of linked tables in a relational database management system. Users will be able to interrogate these tables using queries expressed in SQL, the standard (and ubiquitous) query language for relational databases. This approach allows users to make powerful and flexible queries tailored to their precise requirements. This approach is the same as that successfully adopted for GES and, indeed, the VSA and WSA.

However, there are at least two notable differences between the GES and MOONS archives. Firstly, the GES Consortium performed multiple, independent analyses of each spectrum in order to generate a series of estimates of each astrophysical parameter and elemental abundance. The MOONS Consortium, by contrast, is currently planning only a single analysis of each spectrum. This difference represents a considerable simplification for MOONS. Secondly the GES survey solely observed stars in the Galaxy. The MOONS Consortium surveys, by contrast, will include both Galactic stars (and stars in satellite galaxies such as the Magellanic Clouds) and composite spectra of external galaxies. Analysis of stellar and composite extragalactic spectra yields very different parameters and separate tables will be required to store them. The composite extragalactic spectra have no analogue in GES, but there are some similarities to the 6DF archive[2] that WFAU created some years ago and still curates.

A further difference concerns the files produced by the MOONS reduction pipeline. The GES, VSA, WSA and OSA archives all ingest FITS files produced by pipelines operated by CASU. By contrast, the MOONS Consortium will use a pipeline designed and constructed by a group led by Frederic Royer at GEPI[3], Observatoire de Paris-Meudon. Like the CASU pipeline the MOONS pipeline will produce standard-conforming FITS files. However, FITS is a flexible and extensible standard and there will doubtless be significant differences between the MOONS FITS files and those WFAU has processed hitherto. Also note that astrophysical parameters and elemental abundances will not be computed in the pipeline but will be calculated elsewhere in the Consortium.

Design of the archive could start quite early in the period covered by the current application. However, implementation is better done towards the end of the period, when it is anticipated that many elements of the data reduction pipeline will be in place. Indeed, design of the final stages of the pipeline and the archive could usefully receive joint internal review within the MOONS Consortium to ensure that they inter-operate smoothly. In order to ensure that the archive meets the requirements of the Consortium we will use the same 'twenty questions' approach that has proved successful for other WFAU archives. A list of approximately twenty representative types of query of the archive, capturing its various likely types of use, will be assembled and the archive designed and tested to provide these queries.

The MOONS Consortium intends that on completion their surveys remain accessible as a continuing resource for the astronomical community. WFAU, with its expertise in maintaining, curating and making accessible completed, legacy surveys is well-placed to host this archive. Moreover, the Consortium wishes that the final archive will include both reduced, calibrated spectra and the astrophysical parameters and elemental abundances, as was done for the SDSS spectroscopic surveys[4] and WFAU is doing for GES.

Finally, though the archive is intended as a repository for the surveys conducted by the MOONS Consortium it could also hold other similar, but probably smaller, surveys conducted by other users of the instrument, again as happened for the VSA and WSA archives.

3.4 WP2 DELIVERABLES and MILESTONES

Milestones:

Q1 2016: Outline requirements for target selection.

Q2-2016: Agree scope and contents of the science data archive with the MOONS Consortium.

Q4 2016: Detailed requirements for target selection.

Q4-2016: Preliminary design of the archive complete; includes an outline set of entity-relation diagrams for the database tables.

Q4-2017: Customised tools in place for target selection.

Q4-2017: Detailed design of the archive complete; includes a complete set of entity-relation diagrams for the database tables.

Q1-2018: Target selection by the MOONS Consortium starts.

Q1-2019: Initial target lists finalised prior to operations starting later in the year.

[2] See URL: `http://www-wfau.roe.ac.uk/6dFGS/`

[3] See URL: `http://www.obspm.fr/gepi.html`

[4] See URL: `http://www.sdss.org/surveys/`

Q1-2019: Initial implementation of the archive ready to ingest data.

Deliverables:

Q2 2016: Document specifying the scope and content of the MOONS science data archive.

Q4-2016: A requirements document for selecting MOONS targets from the WFAU and Gaia data archives.

Q4 2017: Tools for conveniently selecting MOONS targets from the WFAU and Gaia data archives.

Q4-2017: Detailed design document for the MOONS science data archive.

Q1-2019: Operational science data archive.

3.5 WP2 RESOURCES REQUESTED

(a) *Staff.*
(b) *Travel and subsistence.*
(c) *Consumables:*
(d) *Maintenance:*
(e) *Equipment:*

3.6 REFERENCES

Davenhall, A.C., Cross, N.J. and Read, M.A., 2015, “MOONS Target Selection from WFAU Surveys” (WFAU).

WP3-WP5: Management & Infrastructure

Staff involved

A.C.Davenhall (WP3: 0.40 FTE)
R.P. Blake (WP4: 0.40 FTE)
M.S. Holliman (WP4: 0.33 FTE)
N.J. Cross (WP5: 0.30 FTE, Q2 2016-Q2 2017)
S.T. Voutsinas (WP5: 0.30 FTE, Q2 2017-Q1 2019)

4.1 INTRODUCTION

The work presented above in WPs 1 and 2 constitutes an ambitious programme of research, development and operations for the next three years, with a significant impact on the UK astronomy programme, as demonstrated by the letters of support in the Appendix. For that programme to be successful, the grant-funded RAs undertaking it must be well supported in a number of ways: they must receive both scientific direction and technical guidance, to ensure that they are doing what the WFAU user community requires and that they are doing it using sensible techniques and technologies; they must have project management support to ensure that they work effectively towards clearly defined goals in a timely fashion, and that they report their progress as required to satisfy funders and other key stakeholders, such as survey PIs; they require well designed and well maintained computing infrastructure of a specification that meets their needs; and they need basic administrative support so they can concentrate on the work they are funded to perform.

WPs 3–5 provide that support: WP3 covers project management, while WP4 and WP5 support, respectively, the hardware and software sides of the WFAU archive infrastructure.

4.2 WP3: PROJECT MANAGEMENT

The scientific direction for WFAU's work is provided by the grant PI, for whom we request continued funding at the 0.2 FTE level in order to fulfill that role. Since 2009 technical direction for the WFAU programme has been provided by Keith Noddle, who was appointed at that 0.75 FTE into a joint Technical Lead/Project Manager post. Noddle provided a pulse of technical innovation, which defined the new archive software infrastructure whose continued development and full deployment we outline in Section **??** below. As described there, that project is progressing very well towards its completion, and we forsee no need for a significant technical initiative over the coming three years, so we are reducing our management request to a 0.4 FTE Project Manager post, filled by Clive Davenhall.

Davenhall's post covers routine management and administration to ensure the efficient running of the grant. The WFAU archives comprise a complex operation which involve ssimultaneously preparing a number of different data archives, each with multiple external partners and which are increasingly international in scope. This operation needs to be managed effectively if releases for all archives are to be delivered in a timely fashion. Specific tasks which need to be addressed include the following:

- Monitoring the availablility of new data releases for ingest into the archives; superintending the preparation of new releases of the archives; monitoring progress and ensuring that deadlines are met.
- Superintending equipment procurement to ensure that adequate processing power and disk space are

available in a timely fashion as needed for new data releases.

- Ensuring that information about the WFAU archives is disseminated to the astronomical community via a variety of channels, including initiating and superintending the preparation of publications and presentations at meetings.
- Managing budgets, primarily for equipment purchase and travel.
- Managing WFAU's interaction with ESO and with public survey PIs.
- Ensuring that WFAU archives are accessible to VO services and that WFAU development of VO-aware software dovetails with other VO software development efforts. The latter aspect requires particular care because of the international and often consensual nature of VO standards and software development.

While most of these tasks are routine, in the sense that the form a regular procedure, they require technical knowledge. Since the significant administrative tasks are undetaken by the Project Manager, we are not requesting secretarial support (either DI or pooled DA) for this grant.

4.3 WP4: *ARCHIVEHARDWAREINFRASTRUCTURE*

4.3.1 Introduction

The WFAU data centre currently hosts over 1.1 petabytes of archived data, consisting of .75PB of image data and and .35PB of databases. All of these data are made available to users through online services and standardized interfaces that allow varying levels of access. Our operational policies utilize both hardware design and software tools to maximize service availability while minimizing the risk of data loss, all within the minimal possible budget. Over the period of the next grant our data archives will grow to 1.36PB of image data and .64PB of database data (a rise of %82 for both respectively) and reach a combined total of over 2PB. A number of our hardware systems will also reach obsolescence during this period, and the services and data they host will need to be consolidated onto new hardware. We have prepared a hardware purchasing plan that meets the requirements for both our continued operations and the consolidation of retired equipment. This plan takes advantage of the decreasing cost of equipment over time by making purchases only when required to meet our needs. Assuming our resource request is met in full, we will be able to continue to provide our world class archives and services to the scientific community through 2018 and beyond.

4.3.2 Existing Systems Summary

The hardware systems we operate can be roughly divided into five main categories: *WebServers*, *DatabaseCluster*, *Curati*

•**Web Servers**: *providepublicfacingservicestoendusers*(*bothscientistsandprojectdevelopers*).*Theseservicesincludethe*

•**Database Cluster**: *hostsallofourscienceready databasesandthecurationdatabasesusedtoingestdataandgeneraterelea*

•**Curation Servers**: *curatethepipelineprocessedimageandcataloguedataintoscienceready dataproducts*.*Theyalsoservea*

•**Storage Servers**: *storethelargevolumeofimagefilesandotherflatfiledataproductsweprovidetoourusers*.*Thesemachi*
60*diskswithina*4*Ufootprint*).*Thedisksareformattedinlarge RAID*6*arraystomaximizeavailablediskspacewhilemaintai*
EDSSthatprovidesthenecessarymountprotocolsonastableplatformatnegligablecost.*Wecurrentlyhave*9*storageserversi*

•**Backup Equipment**: *usedtoprovidebothliveandofflinebackupsofourprocesseddataproductsanddatabasesfordisaster*

4.3.3 Predicted Operational Requirements

The WSA/VSA/OSA projects all receive new data at predictable rates. The pixel data from each of these projects is also expected to receive significant reprocessing and republishing during the next grant period. The earlier versioned data will be stored alongside the reprocessed data in order to preserve the scientific

Table 4.1: WSA Operational Requirements

Year	Pixel Total TB	DB Total TB	Pixel Reprocessed TB
2016	0	5.2	100
2017	0	5	
2018	0	0	

Table 4.2: VSA Operational Requirements

Year	Pixel Total TB	DB Total TB	Pixel Reprocessed TB
2016	73	73.75	
2017	73	92.95	120
2018	73	107.15	

validation of existing published papers. We have calculated the new data flows and estimated the reprocessing volumes over the next three years to determine our storage and processing requirements. The breakdown per project is as follows:

•**WSA**:$TheWSAwillnolongercollectpixeldataduringthegrantperiod, butitwillhave2databasereleasesandalargevolume$

•**VSA**:$TheVSAcollectsnewpixeldataatarateof.2TB/night.Ithasa3:1ratioofpixeldatatodatabasedatabasedontheexisti$

•**OSA**:$TheOSAcollectsnewpixeldataatarateof.05TB/night.Ithasa10:1ratioofpixeldatatodatabasedatabasedontheexi$

4.3.4 Consolidation Requirements

We utilize all our systems for as long as possible before replacing them in order to maximize the return on investment. The average lifetime of all our systems is currently 5.4 years, with 23 out of 44 machines having been in service for 6 years or longer. As noted in the summary above, there are a number of systems that are reaching obsolescence and need to be replaced during the next grant period. All of the services and data on those systems will need to be migrated to new equipment in order to continue their operation. There is a significant risk of data loss and serious service disruption if the necessary migrations do not occur. The most critical systems that need to be replaced consist of the 10 database nodes (nominally totaling .15PB) and 6 of the storage servers (nominally totaling .464PB), as highlighted below:

Fortunately these machines do not need to be replaced on a one for one basis due to the advances in disk size and CPU counts/speeds. Therefore the data stored on these systems can be consolidated onto a smaller number of new servers, purchased according to Table **??**.

We examined two alternative options for reducing the costs associated with the hardware consolidation, but neither appears to offer sufficient cost savings when weighed against their operational impact. They are as follows:

•**Tape Storage of old data**$-WeexaminedthepossibilityofmigratingolderUKIDSSandVISTAimagedataontoLTO$

•**Old database migration**$-WeexaminedthepossibilityofmigratingolderUKIDSSandVISTAdatabases(approximat$
500

4.3.5 Resources Requested

We assembled our purchasing plan from the operational and consolidation requirements higlighted above. The prices for each hardware type are taken from recent purchase prices for similar equipment. We have

Table 4.3: OSA Operational Requirements

Year	Pixel Total TB	DB Total TB	Pixel Reprocessed TB
2016	18.25	1.5	
2017	18.25	2	
2018	0	2.5	140

Table 4.4: Consolidation Requirements

Year	Retired Storage TB	Retired Database TB
2016	226	120
2017	0	15
2018	238	15

Table 4.5: Hardware Price by Year

Year	Storage Server Price	Database Node Price	Web Server Price
2016	£22,097.34	£17,588.33	£9360
2017	£19,927.05	£15,786.35	£8440
2018	£17,969.92	£14,168.99	£7611

applied a cost depreciation factor to the prices over time which was calculated by examining our historical prices for each system type (9.82% and 10.25% price decrease per year for storage servers and database nodes respectively). This yearly price information is highlighted in Table **??** below.

The purchase plan for our operational needs is detailed in Table **??** below. The yearly needs for each survey project (Table **??**, Table **??**, and Table **??**) have been combined to provide a view of the total storage requirements for both categories of systems per year. The plan also includes the expected costs of renewing maintenance contracts for critical servers, replacement webservers for retired machines, and the incidental costs incurred each year (LTO tapes, spare disks, etc).

The purchase plan for our consolidation needs is detailed in Table **??** below.

By combining the yearly costs of operations and consolidation we get the total yearly hardware budget for requested resources as seen below:

4.4 WP5: ARCHIVE SOFTWARE INFRASTRUCTURE

4.4.1 Introduction

The great increase in the volume of WFAU data holdings over the past decade has been more than matched by the increasing complexity of their inter-relationships and of the manner in which the community wishes to use the data. The SDSS *SkyServer* was the model for the development of WFAU's science archive, but it was conceived as something very different from today's WSA and VSA, namely a standalone archive serving a user community whose research was focussed on the monolithic SDSS optical survey. Users were generally expected to download datasets for analysis and, while the SDSS archive introduced the innovative *MyDB* system for providing user space in a database, results held there could only be used as part of further queries with SDSS data, and were not externally queryable. By contrast, the WSA and VSA both hold data from a number of scientifically distinct surveys, each of which was conceived as part of wider, multi-wavelength research programme, which could be supported by the developing global Virtual Observatory (VO) system.

Several years ago it became apparent that these changing requirements and technical opportunities necessitated the development of a new software infrastructure for WFAU's science archives, that built on the lessons of the SDSS-inspired initial deployments of the WSA and VSA, but could set those archives in a wider, distributed environment through the use of VO technologies, to enable them to better support the multiwavelength analyses that would come to dominate the scientific use of WFAU-curated sky survey data.

Table 4.6: Operational Purchase Plan

Year	Pixel Total TB	Storage Servers	Database Total TB	Database Nodes	Web Servers	Maintenance	Inci
2016	191.25	2	80	3	0	£2995	£
2017	211.25	1	100	1	1	£3820	£
2018	213	1	110	2	1	£3050	£
							Tota

Table 4.7: Consolidation Purchase Plan

Year	Pixel Total TB	Storage Servers	Database Total TB	Database Nodes	Cost
2016	226	1	120	1	£39,685
2017	0	0	15	1	£15,786
2018	238	1	30	0	£17,970
				Total	£73,442

Table 4.8: Total Hardware Costs

Year	Operations Cost	Consolidation Cost	Total Cost
2016	£103,720	£39,685	£143,405
2017	£51,740	£15,786	£67,526
2018	£60,735	£17,970	£78,705
		Total	£289,638

This new software infrastructure, known as *Firethorn* internally, was largely designed by Noddle and Dave Morris, a former *AstroGrid* software engineer, who was recruited to WFAU in 2012. As the *Firethorn* design took shape, it quickly became clear that the same infrastructure could be of great benefit to multiwavelength research consortia, both in supporting their collaborative operations while they are taking data, and, potentially, in ensuring that their data is preserved in a re-usable format long after the consortium has dissolved.

Over the past three years, *Firethorn* has been developed by Morris and Stelios Voutsinas, partly through predecessors to this grant, using the OSA as a testbed, and, most recently the EU FP7 project *GENIUS* which is prototyping data access services to enhance the capabilities of the *Gaia* archive to be hosted at ESA's European Space Astronomy Centre (ESAC) near Madrid. As described in Section **??** below, most of the components of *Firethorn* have been prototyped, but their implementation has been driven by the particular requirements of *Gaia*. WP5 seeks support for the deployment of the *Firethorn* platform tailored to the WFAU science archives being developed and deployed through WP1 and WP2, and its extension to meet additonal data analysis requirements from the research communities that use them.

4.4.2 Report on recent work: current status of *Firethorn*

Figure A above illustrated our vision for how WFAU archives should participate in the distributed system of astronomical databases published through the IVOA Table Access Protocol (TAP), and Figure B shows how far we have got with implementing that through the *Firethorn* project. All the key components required to support distributed querying between local and remote TAP services are in place. OSA users may pose queries through the OSA web interface using either the SQL dialect native to our RDBMS (Microsoft SQLServer) or the IVOA ADQL standard and the results of their queries are written into their user space in a database before they are displayed on the results webpage. From there, they may be exported to TOPCAT or Aladin for further manipulation using the IVOA Simple Application Messaging Protocol (SAMP).

Most of the recent work on *Firethorn* has been funded through the *GENIUS* project and has, therefore, focussed on the particular requirements of *Gaia*. The ESA Science Archive Team at ESAC are already developing a *MyDB*–like facility for use in the *Gaia* archive, so *GENIUS* is more interested in the distributed query processing capabilities of *Firethorn* than in the user data side (even though that offers significantly greater functionality, as discussed in Section **X** below).

When *Firethorn* metadata service receives an ADQL query (e.g. through our *Gaia* test portal[1]) it first parses it to determine which TAP service(s) it is targetting. Since it is configured to know which services are co-located with it, it can make a direct database connection for faster execution of a query against a single local database; this is what currently happens for OSA queries. If multiple services are involved in the query, *Firethorn* invokes an customised version[2] of the *OGSA-DAI* Distributed Query Processor (DQP) to decompose the original query into sub-queries which are executed by the multiple TAP services and the

[1] `http://genius.roe.ac.uk`

[2] *OGSA-DAI* (see `http://www.ogsadai.org.uk/`) is a generic distributed data access and management system developed as part of the UK e-Science Programme. We have worked with OGSA-DAI developers at EPCC to customise their software for our purposes and WFAU now maintains a branch of the relevant code, which is now open source.

result sets from each are then aggregated before being written into the user's space in a WFAU database and then passed back to the user through the web interface.

Currently, the TAP service for WFAU databases is provided by an extended version of the old *AstroGrid* DataSet Access webservice, which is not fully compliant with the current version of the TAP standard, so we are developing a new TAP service, which we plan to have in operation before the start of the new grant. It is well known that one of the major problems with distributed systems is the difficulty of feeding back to the client useful error information: a user may launch a task that invokes a chain of distributed services and it can be frustrating if the difficulty of passing error information back up through that chain means that the user is provided with nothing more than a generic error message that says something has gone wrong somewhere. To address that, we are currently developing an asynchronous callback mechanism that allows the DQP service to feed information from the TAP services back to the user through *Firethorn*. Not only will this provide the user with more specific error information, from the specific TAP service that has raised an exception, but it will also allow features like paging through partial result sets as they become available, instead of having to wait for the execution of the full query to complete; this is very valuable, as it will allow users to cancel queries as soon as they discover they are erroneous, thereby saving resources that would otherwise be wasted due to user error.

Testing distributed systems is also difficult, so we have set up a Python-based test infrastructure that can execute queries against combinations of WFAU archives either directly or via TAP as a simulated distributed environment. For this activity we have used queries from the logs of WFAU archives, to ensure that our distributed system can support the types of queries that users want to pose; this has identified a number of instances where ADQL does not support particular features of SQLServer's SQL dialect that are employed by our user community, and we are taking these through the IVOA standardisation process as enhancements to ADQL.

4.4.3 Proposed programme of work Q2 2016 − Q1 2019

We propose to extend the *Firethorn* system in two ways during the grant period: (i) completing implementation of the user data functionality; and (ii) allowing users to perform more analysis in the data centre. These are both targetted at supporting multi-wavelength research consortia and we intend to develop the new functionality in collaboration with one such team – Herschel ATLAS (H-ATLAS) – while neither is a priority for the *GENIUS* project, with its tight focus on *Gaia*, which is why we seek support from this grant.

H-ATLAS as a testbed for *Firethorn* development

The H-ATLAS survey (Eales et al 2010) is an excellent example of the kind of multi-wavelength survey that can benefit from the *Firethorn* architecture. The largest open-time survey undertaken with the *Herschel Space Observatory*, it is being undertaken by a multi-institutional, multi-national team, who have amassed a rich multi-wavelength dataset in its three survey areas, at the North and South Galactic Poles, and in an equatorial region also surveyed by the GAMA (Driver et al 2009) survey. Some of this data comes from public surveys - e.g. GALEX, SDSS, UKIDSS and VISTA/VIKING - for some of which the default data products are used, while customised data products are generated from others. The H-ATLAS team are also undertaking their own follow-up observations of their far-infrared sources, and so have produced a very valuable dataset, that is heterogeneous in nature, but whose whole is much greater than the sum of its parts in scientific terms.

Like most such consortia H-ATLAS has struggled to deploy enough effort managing its data – postdocs and PhD students are recruited, and want, to do science, not data management. This has a detrimental effect both to the scientific effectiveness of the team (as data is not always as well documented or stored in as logical a structure as might be hoped) and on the long-term preservation on the scientific value that resides in the multi-wavelength corpus. The constituent parts of the dataset will continue to be available in separate telescope archives long after the H-ATLAS team dissolves, but the team's published papers will not have captured all the scientific value incorporated in the combined dataset, which is unlikely to be actively managed once the H-ATLAS project is complete.

The H-ATLAS data are most complete in the equatorial fields in which there are images and catalogues from a total of 25 passbands, from the GALEX FUV band through to the Herschel 500μm band. The images comprise about 3TB, and, adding in the data from the NGP and SGP fields, and all the source catalogues, the total H-ATLAS data volume is about 10 TB. The dataset is, therefore, modest enough in volume, but complex enough in structure, to serve as an ideal testbed for *Firethorn* development. Most importantly, two key members of the H-ATLAS team – Loretta Dunne (Co-PI) and Steve Maddox – are currently spending half of their time in Edinburgh, which will provide the ideal opportunity for close collaboration in the early stages of the project, to ensure that our development captures the detailed scientific requrements of such multi-wavelength research consortia. With that in mind, we intend to deploy effort from an astronomer/developer, Nick Cross, on this project initially, with a software engineer, Stelios Voutsinas, taking over once the project reaches the stage of deploying generic services that have been prototyped successfully with H-ATLAS.

User data space

As noted above, results from OSA queries are currently written into tables within a database that are owned by the user who has executed the query although those tables are not currently exposed to the user. The required extension to support multi-wavelength research consortia is to make the data in those user spaces shareable within groups and usable within the VO context. This would then make much simpler many of the scientific workflows that such consortia currently undertake in a much more cumbersome manner at the moment.

To take an example from H-ATLAS science, an astronomer could identify a sample of candidate obscured quasars from a distributed query that made use of a number of the H-ATLAS datasets plus an *XMM-Newton* or *Chandra* X-ray catalogue published through a remote TAP service, store the sample in their user space and share it with their H-ATLAS collaborators via a TAP service in front of the user database that only accepted queries from authorized users. Once the analysis of the sample was complete, the catalogue could be published to the VO in parallel with publication of the paper describing it, so that the wider community would have ready access to the sample for further analysis.

Supporting that type of use case involves a number of enhancements to the existing *Firethorn* software, namely:

- **Access controls**. The current *Firethorn* code supports the identification of a user as the owner of a database table, but needs to be extended to support access to the table by a range of users who may have different rights: e.g. some users may only be able to read the table, while others may also be able to write to it or delete it. This is standard functionality, but it needs to be included in *Firethorn*.
- **User data interface**. A user-friendly interface will be needed so that a user can manage the data in their user space (which will ultimately include flat files as well as database tables) and also manage the access controls they wish to apply to them (e.g. setting up and managing groups). We have prototyped this functionality, but its design has yet to be tested through use in realistic scientific workflows. This will best be done through use within a small subset of the H-ATLAS team who are willing to use it for a real scientific project while the software is still in development and to provide rapid feedback on their experiences of using it.
- **Admin user interface**. Rolling out the user data functionality is a producton service will generate sysadmin requirements that are not met by current WFAU archive software. The SDSS archive team experienced significant practical problems administering *MyDB* databases for hundreds of users with standard tools: for example, a drop-down menu for database selection works when there are ten databases on the server, but not when there hundreds.

Data analysis in the data centre

The science archive paradigm pioneered by SDSS envisages users selecting data in a database (and, perhaps, generating some aggregate statistics for it there), but needing to download the selected dataset to their workstation for more involved analysis. This approach is already cumbersome in the larger WFAU archives, and will become impossible for next generation of sky surveys – e.g. Euclid and LSST – for which the scientifically-valuable subsets needed for many analyses will be too large for users to download and those analyses must be undertaken on compute resources co-located with the archive in the data centre.

This is well understood and, for example, the LSST data management system envisages Data Access Centres having both storage and compute resources to do that, but there is limited practical experience in making such a system work. The one notable exception is the *CANFAR* system operated by the Canadian Astronomy Data Centre (CADC), which supported multiple teams analysing some of its archive data by having each deploy virtual machines (VMs) in a cloud computing environment close (in network terms) to the data. The CANFAR experience (**ref**) was a very positive one for the scientists, but not a complete success for the data centre managers, who struggled with some of the practicalities of operating large numbers of VMs running complicated stacks of analysis software in such a system, suggesting that a better technology is needed.

A promising contender for such a technology is *Docker*[3], which is gaining considerable interest in the software development community. In contrast to the conventional VMs used in cloud computing, which emulate all aspects of a real computer, *Docker* is based on the concept of namespacing in the linux kernel, which essentially allows multiple processes, called containers, to run independently within the same linux kernel on the same physical machine, without the requirement to emulate a full computer. This has obvious performance benefits, while other advantages of *Docker* are **...Dave to add something here.**

We have been conducted initial experiments deploying *Firethorn* components in *Docker* containers and this approach appears very promising as a flexible, scalable way of providing users – and groups of users, in a research consortium – with data analysis services within the data centre. The next step is to test it out on real scientific use cases, and here, again, collaboration with the H-ATLAS team can provide the necessary scientific input.

For a multi-wavelength consortium like H-ATLAS, a number of standard scientific operations can be simplified by performing them in the data centre, in or close to the database holding the data. Firstly, associations must be made between entries in catalogues derived from different passbands. The standard WFAU archive software generates ”cross-neighbour” tables between catalogues, which store candidate counterparts on the basis of spatial proximity, but in many cases it is necessary to supplement the spatial matching stage with a statistical assessment of those potential candidates, using the Likelihood Ratio method (e.g. Sutherland & Saunders 1992, Smith et al 2011) or alternative methods (e.g. Budavari & Szalay 2008). Other standard operations that follow from that are the estimation of photometric redshifts and running an SED-fitting code like MAGPHYS (da Cunha, Charlot and Elbaz 2008), while all three of these steps may have to be prefaced by running a matched-aperture photometry code like that described in Section **X** and, e.g. by Clark et al (2015).

To test the usability of a container-based approach to running data analysis jobs in the data centre, we will work with the H-ATLAS team to configure *Docker* containers to hold standard codes for each of these tasks and get them working with *Firethorn*. We can then offer these (and similar applications) as standard *Docker* images that can be deployed by users within the WFAU data centre.

4.5 WP3-WP5 DELIVERABLES and MILESTONES

Milestones:

QN 201N: Milestone 1

QN 201N: Milestone 2

Deliverables:

QN 201N: Deliverable 1

QN 201N: Deliverable 2

[3] `www.docker.com`

4.6 WP3-WP5 RESOURCES REQUESTED

(a) *Staff.*
(b) *Travel and subsistence.*
(c) *Consumables:*
(d) *Maintenance:*
(e) *Equipment:*

4.7 REFERENCES

Risks

5.1 INTRODUCTION

Impact Plan

6.1 INTRODUCTION

Appendix : Letters of Support

7.1 INTRODUCTION

Bibliography

[1] B. Q. Chen, M. Schultheis, B. W. Jiang, O. A. Gonzalez, A. C. Robin, M. Rejkuba, and D. Minniti. Three-dimensional interstellar extinction map toward the Galactic bulge. , 550:A42, February 2013.

[2] N. Cross, R. Collins, M. Read, R. Blake, E. Sutorius, N. Hambly, M. Holliman, and R. Mann. A Matched Aperture Photometry Pipeline Incorporated into the WFAU Archives. In N. Manset and P. Forshay, editors, *Astronomical Data Analysis Software and Systems XXIII*, volume 485 of *Astronomical Society of the Pacific Conference Series*, page 371, May 2014.

[3] N. Cross and M. Read. Extinction Maps in the WFAU Archives. *ArXiv e-prints*, May 2015.

[4] P. Fernique, T. Boch, T. Donaldson, D. Durand, W. O'Mullane, M. Reinecke, and M. Taylor. MOC - HEALPix Multi-Order Coverage map Version 1.0. *ArXiv e-prints*, May 2015.

[5] T. Muraveva, G. Clementini, C. Maceroni, C. J. Evans, M. I. Moretti, M.-R. L. Cioni, J. B. Marquette, V. Ripepi, R. de Grijs, M. A. T. Groenewegen, A. E. Piatti, and J. T. van Loon. Eclipsing binary stars in the Large Magellanic Cloud: results from the EROS-2, OGLE and VMC surveys. , 443:432–445, September 2014.

[6] V. Ripepi, M. I. Moretti, M. Marconi, G. Clementini, M.-R. L. Cioni, J. B. Marquette, L. Girardi, S. Rubele, M. A. T. Groenewegen, R. de Grijs, B. K. Gibson, J. M. Oliveira, J. T. van Loon, and J. P. Emerson. The VMC survey - V. First results for classical Cepheids. , 424:1807–1816, August 2012.