VISTA Data Flow System survey access and curation: The WFCAM Science Archive

Nigel C. Hambly[a], Robert G. Mann[a,b], Ian Bond[a], Eckhard Sutorius[a], Michael Read[a], Peredur Williams[a], Andrew Lawrence[a] and James P. Emerson[c]

[a]Wide Field Astronomy Unit, Institute for Astronomy, School of Physics, University of Edinburgh, Royal Observatory, Blackford Hill, Edinburgh, EH9 3HJ, UK;
[b]National e-Science Centre, 15 South College Street, Edinburgh, EH8 9AA, UK;
[c]Astronomy Unit, Queen Mary University of London, Mile End Road, London E1 4NS, UK

ABSTRACT

VISTA Data Flow System (VDFS) survey data products are expected to reach of order one petabyte (10^{15} bytes) in volume. Fast and flexible user access to these data is pivotal for efficient science exploitation. In this paper, we describe the provision for survey products archive access and curation which is the final link in the data flow system from telescope to user. Science archive development at the Wide Field Astronomy Unit of the Institute for Astronomy within the University of Edinburgh is taking a phased approach. The first phase VDFS science archive is being implemented for WFCAM, a wide–field infrared imager that has similar output to, but at a lower data rate than the VISTA camera. We describe the WFCAM Science Archive, emphasising the design approach that is intended to lead to a scalable archive system that can handle the huge volume of VISTA data.

Keywords: VISTA, VDFS, science archives, WFCAM

1. INTRODUCTION

The use of the phrase 'science archive' implies an entity that is more than a simple repository of data. The concept[1] of a science archive includes, amongst other things, provision for highly flexible querying via a variety of interfaces, and science–driven and database–driven data products that are significantly enhanced with respect to the ingest data. The need for astronomical science archives is becoming increasingly clear. For example, current data volumes in astronomy are undergoing expansion at a rate that outpaces Moore's Law[2] while the limiting hardware factor (disk I/O performance) is increasing more slowly than Moore's Law. The large legacy survey imaging datasets originating from digitised scans of photographic plates[3,4] comprise tens of terabytes (TB) of imaging data (and approximately 10% of that volume, or $\sim$ 1 TB, of corresponding catalogue and descriptive data). The current generation of infrared survey instruments (e.g. WFCAM[5]) will yield datasets an order of magnitude larger, i.e. $\sim$ 100s of TB of data; within two years VISTA[6] will be producing more than eight times as much data again, i.e. petabytes. Issues concerning centralised, controlled (re)calibration and release of prepared datasets also imply the need for well organised and curated data, as does the concept of data mining in the context of the Virtual Observatory.

The VISTA Data Flow System (VDFS) is a project that aims to handle the torrent of data from the new generation of large format, wide–angle infrared survey instruments by building on UK expertise in curating and disseminating large scale optical datasets. Pipeline processing and science archiving is being implemented first for WFCAM data (where the challenges that this presents in itself have been summarised previously[7]; separate papers[8,9] in these proceedings give an overview of the VDFS and describe pipeline processing). Subsequently, it is intended to deploy, operate and maintain a similar but suitably scaled system to handle VISTA data.

This paper describes the design, prototyping and deployment of the first phase VDFS science archive known as the WFCAM Science Archive[10] (WSA).

Further author information: Send correspondence to N.C.H.: E-mail nch@roe.ac.uk; Telephone +44 (0)131 668 8234

2. DESIGN OF THE WSA

2.1. Requirements capture

Comprehensive requirements capture is fundamental to the success of any large, distributed technological project. A set of top–level requirements[11] were assembled and agreed between the relevant observatory and representatives of the VDFS and the UK user community in the form of the UK Infrared Deep Sky Survey[12] (UKIDSS) consortium. For the science archive, these requirements were then supplemented by a representative set of archive usage modes,[13] following the '20 queries' approach advocated by Gray, Szalay and colleagues[1] in development of the science archive for the Sloan Digital Sky Survey. Finally, these were analysed and incorporated into a science requirements analysis document[14] which was iterated and agreed between the archive designers and the UKIDSS consortium. The science archive requirements specified therein clearly illustrate the need for a flexible, scalable archive system with specific access modes tailored to both novice and experienced users. The specifications on hardware (e.g. storage) and software (e.g. user interface functionality and provision of enhanced database–driven products) followed on naturally from this phase of requirements capture.

2.2. Hardware architecture

The science archive hardware for the VDFS is based around 32-bit rack–mounted PC technology, which provides relatively high performance at a relatively modest cost. The design is modular and can be expanded considerably as storage requirements dictate. Figure 1 shows a schematic diagram of the WSA 'private area network'. Features of the design include:

- the private area network (which presently comprises a simple gigabit switch) is isolated from the general site ethernet and network links in order to ensure maximum archive bandwidth for the purposes of ingest/curation/querying and also that such large data volume operations have minimal impact generally;
- the archive hardware is interconnected at 1 Gbit s^{-1} and has a 1 Gbit s^{-1} internet connection directly to a JANET backbone access router straight onto the 10 Gbit s^{-1} JANET backbone;
- internally firewalled PCs protect the WSA and its associated catalogue servers from unauthorised access;
- mass pixel storage (requiring expansion at a rate of $\sim$ 10 TB $year^{-1}$) is achieved using RAID5 arrays of low–cost SATA disks which provide good speed of access;
- the catalogue servers employ Ultra320 SCSI RAID arrays which provide very high aggregate disk IO rates (see below);
- the catalogue servers are mirrored to provide redundant storage and to isolate day–to–day intensive curation activities (which take place on the load server) from users of periodically released and published database products (on the public catalogue server);
- Ultrium–II LTO tape (200 GB per tape native capacity and read/write rates of 100 GB per hour) is used as a catalogue backing store.

The Ultra320 SCSI systems are worth further discussion at this point. We have experimented with several different disk configurations and interfaces to achieve high throughput at reasonable cost. Arrays using different disk interfaces (SCSI, IDE and fibre channel), different interconnection (Ultra160, Ultra320, IDE–ATA and fibre) and both hardware and software RAID were analysed, and our findings were in agreement with others (e.g. Ref. 15) in that software striping provides the best performance (a single Ultra320 channel being capable[16] of sustained read/write to the tune of $\sim$ 200 MB s^{-1}). However, we found that our Tyan 'Thunder Pro' motherboards (which incorporate two independent PCI–X and one PCI bus slots), when fully loaded with dual–channel interface cards, were not stable with 32 Ultra320 disks running at full speed (although they are stable when clocked back to Ultra160, as has been done for the deployed prototype – see later). The source of this instability is not fully understood, but we are presently investigating a hardware RAID solution that employs multi–channel PCI–X interface adapters capable of supporting 'spanned arrays' (also known as RAID50 configurations) where the sets of disks on individual SCSI channels are configured as RAID5 logical volumes,

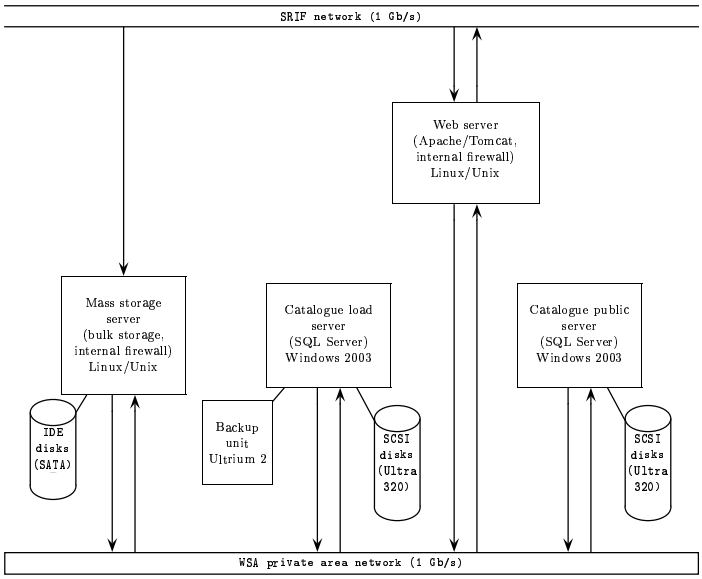

Figure 1. Schematic diagram of the WSA private area network. Arrows indicate the dataflow between the various archive servers.

those volumes then being incorporated into stripe sets within the adapter. Initial results indicate that read performance is excellent, while there is of course a penalty to be paid in RAID5 write performance. The WSA requirements are most stringent on read performance; hence we feel able to accept the RAID5 write penalty.

2.3. Database management and operating systems

The VDFS science archive team have chosen to use a relational database management system (DBMS) for the project. The choice of relational over object–oriented is based on preference: simplicity of application interfaces (e.g. standard structured query language interfaces) for both developers, operators and users is a major requirement. Generally, the use of an off–the–shelf DBMS brings many benefits, including industry–standard interfaces (e.g. ODBC/JDBC and SQL). There are several freeware and proprietary DBMS products to choose from. The WSA requirements indicate the need for an enterprise–class product that:

1. is capable of scaling to large database volumes;
2. has comprehensive documentation and technical support;
3. is available at reasonable cost; and
4. is capable of non–expert administration.

Items 1 and 2 essentially excluded all freeware products and left us with the choice of DB2 (IBM), Oracle and Microsoft SQL Server. We note that Oracle is being used as the back–end store in the ASTRO–WISE project[17] while the SkyServer[18] science archive for the SDSS employs SQL Server. With due regard to items 3 and 4 above, the WSA have chosen to follow the SDSS lead and use SQL Server, enabling us to make use of astronomy–specific developments for that system (e.g. DBMS implementation of the spherical spatial indexing

scheme known as Hierarchical Triangular Mesh[19]) via invaluable collaboration with the SkyServer development team within Johns Hopkins University and Microsoft Research. Experiments are continuing with the prototype science archive dataset (see below) before finalising the scale–out of the science archive design.

With reference to the hardware description in Figure 1, the catalogue servers run Windows Server 2003 (enterprise edition) along with separate instances of SQL Server. The front–end, firewalled mass storage and web servers run Debian Linux.

2.4. Software architecture

The top–level WSA software architecture is designed to provide a scalable, easily maintained system that is flexible to refactoring in the light of iterative software development and 'schema evolution' due to changes to the detailed requirements of the data flow and users. The following sections describe in some detail the software architecture of the WSA.

2.4.1. Data modelling

We have used generalised entity–relationship (ER) modelling to provide a schematic relational model of the WSA for ease of review with the end users and also to assist in the coding and implementation phases. Figure 2 shows the entity–relationship model for *all* image data in the WSA. This model describes the smallest, basic unit of image data (a single device detector image) through calibration, multi–device camera 'paw prints', stacks and mosaics to the most heavily processed database driven products like tiled and stacked contiguous images. The basic concept has two entities: 'multiframe' and 'detector frame' that contain, amongst other information, the FITS keywords (primary and extension headers of the multi–extension FITS, or MEF, source files respectively) describing the image data. A simple one–to–many relationship specifies that each multiframe consists of one or more detector frames. Note that the actual number of detectors is not specified; hence this model will work for VISTA data (16 detectors) as well as WFCAM (4 detectors). Furthermore, this model can cope with the possibility of loss of one or more detectors during the operating lifetime of the instrument; it is in fact applicable to any survey imaging system.

A comprehensive set of similar ER models was developed[20] early on in the design phase of the WSA, covering object catalogue data, calibration metadata and database–driven functions (ingest, curation and generation of database–driven products). Once the ER models were developed, SQL schema script files were created to implement the tables, their attributes and relationships.

2.4.2. Schema–driven features

To cope with schema evolution during development, the decision was taken to use the SQL schema scripts to drive data ingest. An *ad hoc* system of comment tags was developed to annotate the SQL scripts to drive certain ingest functions, e.g. mapping of FITS keywords (which are limited to 8 ASCII characters in standard FITS) to table attribute names, transformation of units, and generation of new attributes from ingest records (for example the 20–level HTM index given a spherical co-ordinate pair).

2.4.3. Compartmentalisation

A key requirement arising during the design review process[21] was for an architecture with 'clean' programming interfaces and a modularised approach to yield an easily maintained and scalable archive system. We have followed established best practice and the WSA software architecture incorporates the following features:

- choice of syntactically clear, platform–independent object–oriented (OO) languages (Python and Java) for high level application code;
- OO approach to persistent database driving metadata whereby curation information (e.g. source merging specifications, descriptions of required enhanced images products) is stored in the DBMS in tables/rows/columns that map directly onto a class/instance/attribute OO hierarchy in the curation application software (this is similar to the ASTRO–WISE[17] approach);

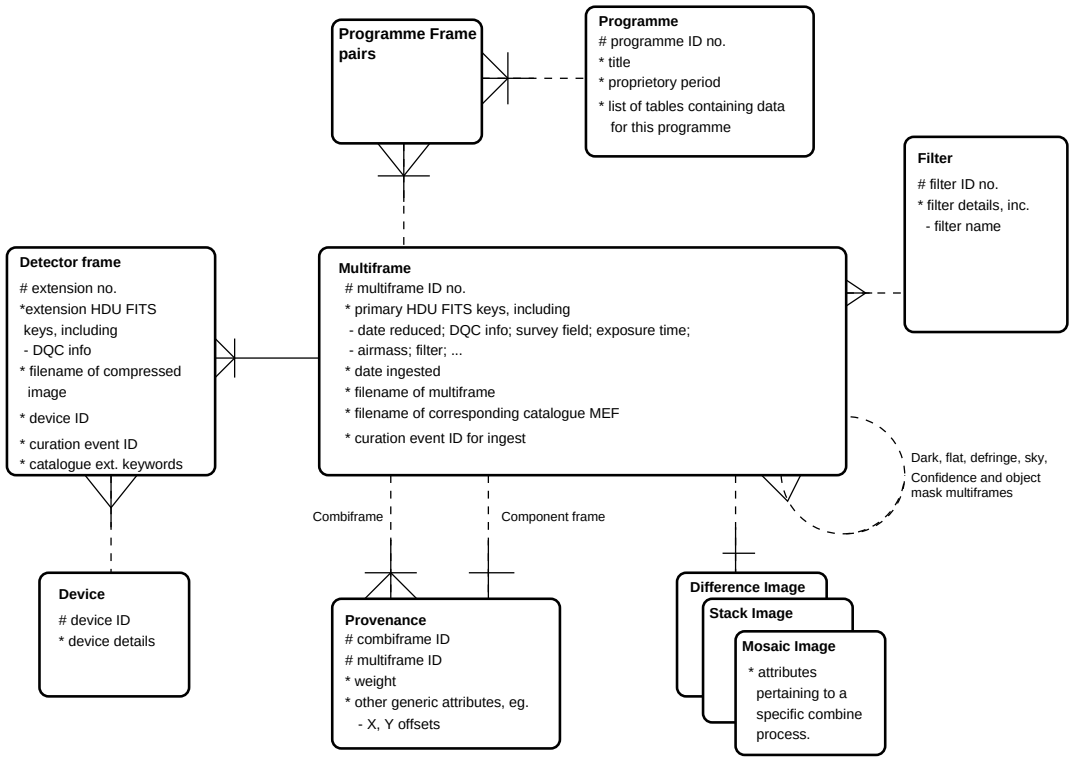

Figure 2. Data model for all image data in the WSA. This generalised entity–relationship model shows how image metadata are stored and related in the archive via a set of optional one–to–many relationships between entities indicated by lines joining the boxes. Note how a multiframe may have provenance (i.e. may have some dependency on other images in the archive – e.g. a combined frame product) and also how calibration frames are stored along with science images. In the mapping of this model to an RDBMS schema, the boxes translate into tables, hashed attributes to primary keys and relationships to foreign keys.

- use of off–the–shelf middleware and an abstracted, layered approach that compartmentalises the curation software;
- client/server architecture between the curation driving server and web server, and Microsoft SQL Servers employing standard, layered interfaces (e.g. ODBC/JDBC, XML–RPC).

With due regard to scalability, some curation applications (e.g. bulk ingest, source merging, billion–row source list associations) need careful implementation. For example, data structure overheads in a multi–layered environment and/or client/server connection overheads, or high level interface overheads such as use of server–side SQL cursors, can easily make even modestly–sized multi–row data modification extremely time consuming. We have developed some applications outside the DBMS via efficient C/C++ code, operating on flat files, that is easily bound in to higher level Python wrapper scripts, and employing the native binary file ingest/output and minimally transaction–logged facilities that are provided in SQL Server.

Needless to say, all WSA software development is being undertaken in a disciplined environment employing CVS as a centralised, version–controlled repository and software documentation tools (JavaDoc, Doxygen and Epydoc) to provide software description/maintenance documentation.

3. WSA PROTOTYPE: THE SUPERCOSMOS SCIENCE ARCHIVE

As a real–world test of the development and deployment of a terabyte-scale science archive, in the absence of such large amounts of WFCAM or VISTA data we have ingested the legacy SuperCOSMOS Sky Survey[3] catalogue data into a WSA prototype system as described above. This prototype is known as the SuperCOSMOS Science Archive (SSA); the homepage is at `http://surveys.roe.ac.uk/ssa`, and further information is given in Ref. 16. The SSA illustrates many features of the WSA, including a similar (albeit much simpler) data model, incorporation of several large external datasets (as specified in the WSA requirements) that are prejoined to the SuperCOSMOS data for the purposes of joint querying, and provision of a variety of user interfaces that give immediate access to novice users as well as an SQL interface that exposes the full power of structured query language to expert data miners.

The SSA includes 3.7 billion individual source detections arising from over 3000 photographic plates; these individual passband detections are merged into over 1 billion multi–colour, multi–epoch source records. Cross–identifications between that merged source table and the USNO–B[4] catalogue ($\sim$ 1 billion rows), 2MASS[22] catalogues ($\sim$ 0.5 billion rows) and the SDSS DR1 and EDR catalogue products are provided, along with a cross–match of the source list with itself. These result in a combined dataset containing nearly 10 billion rows. The disk I/O subsystem of the SSA is currently based on the software striping configuration described above, but running at Ultra160 speed. Even so, the aggregate IO rate (limited by software destriping in memory) is 300 MB s^{-1}. This in turn enables wholesale trawling of the 1 billion row merged source table ($\sim$ 300 GB) in just 15 minutes, which opens up all sorts of data mining opportunities as well as exploratory science applications. Of course, most commonly executed queries are predicated on a small subset of attributes, i.e. position, brightness and source class. DBMS indexes on those attributes in the SSA result in query response times of seconds for most archive usage modes (such response times being largely limited by network connectivity).

Figure 3 shows the homepage of the SSA. This web interface gives an good indication of how the WSA will function. Features that are relevant to the WSA include:

- variety of output formats, including plain text, FITS and VOTable[23] (see below);
- variety of access modes, including simple radial query (cone search) via a web form, through menu–driven SQL construction to full–blown free–form SQL (there is also cross–identification via provision of upload of a user–supplied source list);
- for FITS and VOTable output formats, availability of push–button start–up of the interactive catalogue browsing utility TOPCAT[24] (this uses Java 'WebStart' software);
- comprehensive online documentation including a database browser and an SQL cookbook aimed at the astronomer end users;
- result–set tie–in with source pixel data, where image thumbnails for selected sources in selected wavebands can be viewed and/or downloaded.

The main difference between the SSA and the WSA is in day–to–day curation. The SuperCOSMOS Sky Survey[3] southern hemisphere dataset is a finished data product like the USNO–B, 2MASS or SDSS data releases, and was ingested as a single bulk load operation. WFCAM and VISTA data, on the other hand, will be ingested periodically on a weekly timescale and curated (e.g. source merged, recalibrated) within the DBMS. Major coding effort at the time of writing is being employed in the deployment and testing of the curation software for the WSA, using simulated and real test data from existing large–format imaging instruments.

4. WSA, VISTA AND THE VIRTUAL OBSERVATORY

The WSA database browser corresponding to the same for the SSA is illustrated in Figure 4 and can be viewed at the URL specified in Ref. 25. At the time of writing (May 2004), the UKIDSS user community are reviewing the organisation and testing the functionality of the WSA browser and the SSA; we anticipate first data ingest into the WSA later this year. At the same time, a phase of requirements capture and review is being undertaken

Figure 3. Screen shot of the home page of the SuperCOSMOS Science Archive – see http://surveys.roe.ac.uk/ssa

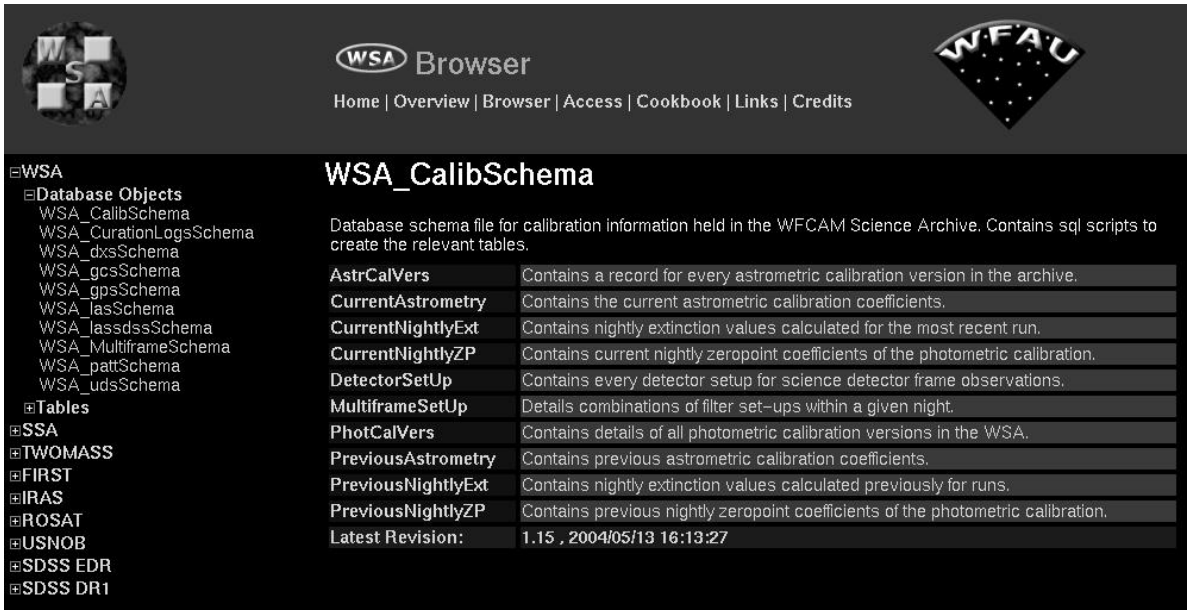

Figure 4. Screen shot of the prototype web browser pages of the WFCAM Science Archive – for more details, see http://www.roe.ac.uk/∼nch/wfcam/SchemaDocs/www/wsa_browser.html.

for VISTA, which will feed into enhancements in the archive design over the next year or so culminating in deployment of a system analogous to the WSA for VISTA data.

In the context of Virtual Observatory (VO) initiatives, we are integrating a number of prototype features. As has already been indicated above, the SSA output format has options to specify VOTable,[23] a prototype XML format for astronomical data. The WSA and SSA schemas also incorporate unified content descriptors[26] (UCDs) which are being developed to facilitate VO functionality transparently to the user. As for integrating the SSA, and subsequently the WSA, into the VO, over the next few months we will be deploying AstroGrid server–side infrastructure. This deployment will consist of:

- installation of a web interface on the existing web server that incorporates VO standard interfaces including cone–search, Simple Image Access Protocol and SkyNode[27] utilities (all contained within the Publisher's AstroGrid Library);
- installation of resource description metadata specifically in the AstroGrid Registry (those data also being generally applicable to any VO Registry conforming to protocols agreed within the International Virtual Observatory Alliance);
- installation of 'MySpace' and other web/grid–oriented facilities (e.g. specific tools such as SExtractor[28] cast as web services) for general use as part of the prototype distributed system.

ACKNOWLEDGMENTS

It is a pleasure once more to acknowledge the kind assistance of Jim Gray and Alex Szalay, Ani Thakar and colleagues within The Johns Hopkins University for their help and advice in developing and deploying the SSA. Thanks are also due to Microsoft for provision of large amounts of software via the Academic Alliance. We thank IT Support within the UK ATC at Edinburgh, particularly Jonathan Dearden and Horst Meyerdierks, for setting up the archive servers, and the UK AstroGrid Project for their software and advice; N.C.H. would also like to thank Andy Knox for much advice concerning large disk arrays. The WSA's direct network connection to the JANET backbone was funded through the Strategic Research Infrastructure Fund (SRIF); the VISTA Data Flow System is funded by grants from the UK Particle Physics and Astronomy Research Council.

REFERENCES

1. J. Gray, A. S. Szalay, A. R. Thakar, C. Stoughton, and J. Vandenberg, "Online scientific data curation, publication and archiving," in *Virtual Observatories*, A. S. Szalay, ed., *Proc. SPIE* **4846**, pp. 103–107, 2002.
2. I. Tuomi, "The lives and death of Moore's Law." First Monday (peer–reviewed journal on the internet), Volume 7 No. 11: http://www.firstmonday.dk/issues/issue7_11/tuomi/
3. N. C. Hambly, H. T. MacGillivray, M. A. Read, S. B. Tritton, E. B. Thomson, B. D. Kelly, D. H. Morgan, R. E. Smith, S. Driver, J. Williamson, Q. A. Parker, M. R. S. Hawkins, P. M. Williams, and A. Lawrence, "The SuperCOSMOS Sky Survey - I. Introduction and description," *Monthly Notices of the Royal Astronomical Society* **326**, pp. 1279–1294, 2001.
4. D. G. Monet et al., "The USNO–B Catalog," *Astronomical Journal* **125**, pp. 984–993, 2003.
5. D. M. Henry et al., "Design status of WFCAM: a wide field camera for the UK infrared telescope," in *Instrument Design and Performance for Optical/Infrared Ground–based Telescopes*, M. Iye and A. F. M. Moorwood, eds., *Proc. SPIE* **4841**, pp. 63–71, 2003.
6. A. M. McPherson, S. C. Craig, and W. Sutherland, "Project VISTA: a review of its progress and overview of the current program," in *Large ground–based telescopes*, J. M. Oschmann and L. M. Stepp, eds., *Proc. SPIE* **4837**, pp. 82–93, 2003.
7. A. Lawrence, N. Hambly, B. Mann, M. Irwin, R. G. McMahon, J. R. Lewis, and A. J. Adamson, "The WFCAM/UKIDSS data archive: problems and opportunities," in *Survey and Other Telescope Technologies and Discoveries*, J. A. Tyson and S. Wolff, eds., *Proc. SPIE* **4836**, pp. 418–425, 2002.
8. J. Emerson et al., "VISTA data flow system: overview," in *Optimizing scientific return for astronomy through information technologies*, P. J. Quinn and A. Bridger, eds., *Proc. SPIE* **5493**, in print, 2004.

9. M. J. Irwin et al., "VISTA data flow system: pipeline processing for WFCAM and VISTA," in *Optimizing scientific return for astronomy through information technologies*, P. J. Quinn and A. Bridger, eds., *Proc. SPIE* **5493**, in print, 2004.
10. http://www.roe.ac.uk/∼nch/wfcam
11. http://www.jach.hawaii.edu/∼adamson/wfarcrq.html
12. http://www.ukidss.org/
13. http://www.roe.ac.uk/∼nch/wfcam/misc/wsausage.html
14. http://www.roe.ac.uk/∼nch/wfcam/srd/wsasrd/wsasrd.html
15. J. Gray et al., "Data Mining the SDSS SkyServer Database." Microsoft Technical Report MSR–TR–2002–01.
16. N. C. Hambly, M. Read, R. G. Mann, E. T. W. Sutorius, I. Bond, H. T. MacGillivray, P. M. Williams, and A. Lawrence, "The SuperCOSMOS Science Archive," in *Proceedings of the 13th ADASS meeting*, F. Ochsenbein and M. Allen, eds., *ASP. Conf. Ser.*, in print, 2004.
17. http://www.astro-wise.org/
18. http://skyserver.pha.jhu.edu/dr1/en/
19. P. Z. Kunstz, A. S. Szalay, I. Csabai, and A. R. Thakar, "The Indexing of the SDSS Science Archive," in *Proceedings of the 9th ADASS meeting*, N. Manset, C. Veillet, and D. Crabtree, eds., *ASP. Conf. Ser.* **216**, pp. 141–144, 2000.
20. http://www.roe.ac.uk/∼nch/wfcam/VDF-WFA-WSA-007-I1/VDF-WFA-WSA-007-I1.html
21. http://www.roe.ac.uk/∼nch/wfcam/WSA_CDR_report.pdf
22. http://pegasus.phast.umass.edu/
23. http://vizier.u-strasbg.fr/doc/VOTable/
24. http://www.star.bris.ac.uk/∼mbt/topcat/
25. http://www.roe.ac.uk/∼nch/wfcam/SchemaDocs/www/wsa_browser.html
26. P. F. Ortiz, F. Ochsenbein, A. Wicenec, and M. Albrecht, "ESO/CDS Data–mining Tool Development Project," in *Proceedings of the 8th ADASS meeting*, D. M. Mehringer, R. L. Plante, and D. A. Roberts, eds., *ASP. Conf. Ser.* **172**, pp. 379–382, 1999.
27. http://www.skyserver.org/skynode/
28. E. Bertin and S. Arnouts, "SExtractor: Software for source extraction," *Astronomy and Astrophysics Supplement Series* **117**, pp. 393–404, 1996.