VISTA DATA-FLOW SYSTEM SCIENCE ARCHIVE (VSA): REVISED PLAN

Introduction

1. The original project plan set out in the 2004 VEGA proposal described five versions (VDFSv1-v5) of the Science Archive, to be released annually at the ends of 2003-2007. Versions VDFSv1-3 were for WFCAM data, offering progressively more functionality, while VDFSv4 is the WFCAM archive scaled for VISTA data volumes, required to be ready for VISTA commissioning in December 2006. The final release, VDFSv5, is the Archive fully shaken down with VISTA data. Owing to the delay in delivery and commissioning of WFCAM, the time-scale between WFCAM coming into operation and the scheduled VISTA delivery has been greatly compressed, necessitating revision of the plan. Now that WFCAM science verification data are being retrieved from the Archive by the community, and the group in Edinburgh is at last fully up to VDFS-funded complement for VDFS work, it is timely to revisit the plan.

2. The Science Archive has now been serving WFCAM data from the Science Verification observations to members of the UKIDSS consortium for some weeks, and we received valuable feedback on operation of the archive at the first science verification meeting on August 18. We are using this feedback to prioritise the work in the coming months.

Revised Project Plan

3. The prime goal of the project remains the production of the Science Archive for VISTA data, including tools for data querying and analysis, and using VO-compatible methods and data formats. We are still working to a target date of December 2006 to have a system (VDFSv4) ready for the VISTA camera commissioning data. The project plan from now (September 2005) until the end of the funded VEGA project (September 2007) is still built around the original five versions, but schedules preparatory work for the VISTA-ready version, VDFSv4 (scalability of WSA design, OS, DBMS and hardware for VISTA data volumes) in parallel with further development of data access tools (v3) for the WFCAM Science Archive (WSA). Some of the design work for the VISTA Science Archive (VSA), primarily the database schemas for the public surveys, can only be completed when the surveys themselves have been defined and accepted by ESO. This information may not be available until mid-2006 and the plan must take account of this. The overall plan is given below, with details in the following sections:

v2: WSA shaken down with experience of real WFCAM data (finish 30/9/2005);
v3: WSA with advanced data access and maintenance tools (1/10/2005 - 31/3/2006);
v4: scalability of WSA design, OS, DBMS, hardware for VISTA (1/10/2005 - 31/3/2006);
v4: VSA detailed design and coding for VISTA commissioning (1/4/2006 - 1/12/2006);
v5: VSA shaken down with experience of real VISTA data (1/12/2006 - 30/9/2007).

Version 2

4. Version 2 was envisaged to be available one year after survey operations began but, because of the delay in WFCAM commissioning and the need to adhere to the VISTA time-table, we will close this version at the end of September 2005. The hardware for the Catalogue Servers and procedures for data transfer and ingest have required considerable effort to shake down (owing to problems with data format) and this experience provides valuable lessons for Version 4. The WSA has served science verification data, including source catalogues and image data, to members of the UKIDSS consortium. It offers users facilities for querying UKIDSS survey catalogues (LAS, GPS, GCS, etc) and other catalogues (2MASS, USNO-B, SDSS-DRn), as well as WFCAM open-time (non-survey) data. Facilities for access control and user registration have been provided. The remaining functions we plan to provide at the end of this phase include:

- "Query builder", a WWW-based tool to facilitate searching the LAS together with external catalogues (SDSS, 2MASS) (joint searches are already possible using SQL, but this will help users unfamiliar with SQL)
- Tool for providing mosaic (filled WFCAM tile, ~ 0°.8) images
- Tools for searching open-time observations
- Tool for photometric re-calibration
- Tool to allow cross-matching by positions in user-supplied lists and searches on UKIDSS parameters (brought forward from V3 in response to UKIDSS science verifiers' requests).

Version 3

5. From October 2005, part of the WFAU VDFS team will continue to work on the WSA, VDFSv3, operating the archive until the database curator (funded separately) is up to speed, re-factoring database schemas and code in the light of experience and feedback from users, and adding additional facilities for catalogue maintenance and data delivery:

- Operate WSA and train archive curator
- Re-factor v2 database design, data ingest and catalogue maintenance tools in the light of experience and feedback from the UKIDSS consortium and other users
- Procure and install further disk sets for data servers
- Build tool for list-driven measurements between WFCAM detections
- Build tools for re-calibrating astrometry and computing proper motions
- Build tool for spatial indexation of revised coordinates
- Re-factor v2 data delivery tools in light of experience and user feedback
- Extend interface/tools to allow cross-matching by positions in user-supplied catalogues and searches on UKIDSS parameters
- Extend WWW interface and tools to return mosaiced images larger than 0°.8, blocked down as specified by user
- Build tools to generate stacked images given a user-selected list of input images and to allow use of user-selected stacking algorithms

Version 4

6. The first phase of work from October 2005 will be studies of the DBMS/OS (MS SQL Server running under Windows) and hardware adopted for the WSA to see whether they are still appropriate for the VSA. If a hardware architecture different from that hosting the WSA is indicated, we will procure and install the first nodes of the servers required for the VSA, to allow plenty of time for any bugs to be ironed out:

- Investigate DBMS/OS scaling for VISTA data volumes: MS SQL Server *vs* IBM DB2 comparison
- Investigate hardware scaling to VISTA data volumes including speed tests
- Specify and agree data transfer interface control document (ICD) with CASU
- If hardware investigation points to different architecture from WSA, procure and install initial servers
- Requirements capture from URD 3.0
- Draw up Curation Use Cases analogous to those for the WSA

7. The second phase, from about April 2006, will include the whole WFAU VDFS team to build the VSA. The database schemas will depend on the forms of the public surveys expected to be carried out by VISTA, which we will need to know by April 2006. As far as is possible, we will adapt techniques and code developed for the WSA for the VSA. Priority will be given to the database and data ingest tools in readiness for commissioning of the VISTA camera in December.

- Design and implement the survey database schemas in the light of VISTA public survey decisions
- Adapt WSA data ingest tools (CU1-4) to fit the ICD and VSA database design
- Adapt WSA catalogue maintenance tools to fit VSA database design
- Procure and install further disk sets for data servers as required
- Design, implement and test survey progress tools for VISTA surveys
- Design and build VSA User interface and adapt WSA data delivery tools for VSA data

Version 5

8. We expect a period of shake-down of the data-flow system and VSA once real data flow from VISTA. In the light of operational experience, and user feedback, we may have to re-factor some of the tools for data ingest, catalogue maintenance and data delivery. Unless there is a delay in the commissioning of the camera, there should be time for feedback and re-factoring before funding ceases in September 2007. In any event, during this phase, we will add further data access tools as specified in the UK VISTA User Requirements (URD 3.0) and upgrade our VO-compatibility.

- Operate VSA and train archive curator
- Re-factor v4 database design, data ingest and catalogue maintenance tools in the light of experience
- Re-factor v4 data delivery tools in light of experience and user feedback
- Implement further data access tools in the URD 3
- Develop system to allow up-loadable user-specified algorithms
- Re-cast internet services as grid services

9. The table below sets out the requirements on the archive from the UK VISTA User Requirements (URD 3.0) and the stages at which these are planned to be met. Some of these are "rolling", such as compliance with developing Virtual Observatory standards, and we intend to meet this by continuing the close cooperation between VDFS and VO staff in Edinburgh and consultation with ESO staff. The URD 3.0 also sets out (5.16) goals for the archive operational uptime and downtime, and recovery from hardware failures. The achievement of these depends on robustness of design, infrastructure (including power) and the provision of resources for staffing, hardware maintenance and spares. The goals will be used in preparation of our Rolling Grant proposals, which fund the archive operation. The goals for archive response times (5.17) and handling of long and short queries are implemented slightly differently: the time estimates by SQL Server's query planner are not always accurate enough to be useful so long queries are put in the background after an interval and the user e-mailed when the results are ready.

Ref	Archive functions from the UK VISTA User Requirements 3.0	V2	V3	V4	V5
5.2	Take account of any proprietary periods	●	●	●	●
5.3	Compliance with the Virtual Observatory standards	●	●	●	●
5.4	Archive data at a peak rate of 1000 GB/day for 10 days, sustained mean rate 650 GB/day			●	●
5.9a	Provide "simple" and "advanced" search interfaces	●	●	●	●
5.9b	Queries shall permit any combination of Boolean conditions		●	●	●
5.9c	Users can define own parameters as combinations of catalogue parameters for searching	●	●	●	●
5.9d	Search by RA/Dec, Galactic or ecliptic coordinates, including circular, square or rectangular windows		●	●	●
5.9e	Facility to upload files of ~1000 positions for thumbnail images and catalogue objects		●	●	●
5.9f	Combine queries with 2MASS, SDSS, USNO-B, FIRST, IRASPSC, UKIDSS surveys &c	●	●	●	●
5.9g	Ability to return any of all, short-list or user-selected source parameters	●	●	●	●
5.9h	Return a random sample (e.g. 1%) of objects within user-selected sky area and magnitude limits	●	●	●	●
5.10	Selections by colour to allow specified cut on time-lag between different filter observations				●
5.11	The VDFS Archive shall have capability for user to request arithmetic image processing.				●
5.12	Archive shall comprise a regularly updated internal database and occasional data releases	●	●	●	●
5.13	Each data release shall remain indefinitely available	●	●	●	●
5.14	Upload a list of target positions and search radius for improved positions and reference stars		●	●	●
5.15a	The archive shall be capable of limiting data access to suitably authorised subsets of users	●	●	●	●
5.15b	Access shall allow for any proprietary periods as necessary.	●	●	●	●
5.15c	As far as reasonably practical, it shall be impossible for users to corrupt data or overload Archive	●	●	●	●
5.15d	Releases will require prior authorisation by the VISTA PI	●	●	●	●

The VDUC is invited to comment of the proposals and priorities.