VISTA DATA FLOW SYSTEM (VDFS)

for VISTA & WFCAM data

Science Archive integration into the VO

author
R.G. Mann (IfA Edinburgh)
WFAU VO Astronomer
number
VDF-WFA-VSA-010
issue
1.0
date
September 2006
co-authors
N.C. Hambly

Contents

1 SCOPE 3

2 OVERVIEW 3

3 THE VIRTUAL OBSERVATORY 4

3.1 Introduction . . . 4

3.2 VO development in Edinburgh . . . 4

3.3 The current status of the VO . . . 5

3.3.1 IVOA standards . . . 5

3.3.2 AstroGrid . . . 7

4 VO-RELATED REQUIREMENTS FROM THE SRAD 9

5 MEETING THE VO-RELATED REQUIREMENTS 10

5.1 Science Archive Functionality . . . 10

5.2 Science Archive Security . . . 11

5.3 Future plans . . . 12

6 Summary and Conclusions 12

7 ACRONYMS & ABBREVIATIONS 14

8 APPLICABLE DOCUMENTS 14

9 CHANGE RECORD 14

10 NOTIFICATION LIST 15

1 SCOPE

This document discusses the current plans for the integration of the WFCAM Science Archive (WSA) and VISTA Science Archive (VSA) into the developing international Virtual Observatory (VO). It does not cover any possible VO-enabling of the data processing side of the VISTA Data Flow System (VDFS).

The VO has yet to reach a state of maturity, and its current status is still characterized more by competing prototypes than by a stable set of internationally agreed standards. In light of that, this document can present no more than current thinking on how to VO-enable the WSA and VSA, since any detailed planning would inevitably become outdated before it came to be implemented.

This document does, however, show how the structures are in place - in terms of the close collaboration between the VDFS-funded science archive development team and VO development and operations staff funded from other sources - to ensure that the WSA and VSA will be integrated into the VO in the manner described in the VDFS Science Archive Science Requirements Analysis Document (SRAD; see AD01). Much of the work required to achieve that will be undertaken by staff funded from outwith the VDFS, but, for completeness, their future work is discussed here, since focussing solely on the VDFS-funded portion of this endeavour would result in a misleading view of a necessarily collaborative activity.

2 OVERVIEW

Large digital sky surveys, like those underway with WFCAM and planned for VISTA, provide some of the strongest science drivers for the development of the VO. Such surveys will yield, by design, data which can form the basis for a wide range of scientific analyses, and which, therefore, will be used in conjunction with a large number of other datasets. While the WSA and VSA database design (AD02) features the integration of certain key external datasets into the VDFS science archives, these cannot comprise an exhaustive list of the data which astronomers will want to use in conjunction with the WSA and VSA. It follows that the WSA and VSA must be made interoperable with external datasets geographically distributed around the world, and the VO aims to provide the infrastructure by which that becomes possible.

Conversely, the WSA and VSA have been designed and developed with the VO in mind from the outset. While the working VO that exists when the final VISTA data enter the VSA will bear little resemblance, at a detailed level, to that envisaged by the authors writing the first WSA design documents, the basic concepts behind the VO are likely to remain fairly constant, and science archive development within the VDFS has been planned and undertaken in a manner that ensures consistency with the relevant aspects of the VO, as it changes from being a set of concepts and aspirations to working software deployed throughout the astronomical community.

This document outlines the current thinking by the VDFS project on the shape that the VO will take on the timescale of WSA and VSA development and on how these science archives will be integrated into the VO. The contents of its remaining sections are as follows:

- Section 3 provides an introduction to those aspects of the Virtual Observatory relevant to the VO-enabling of the WSA and VSA. It briefly discusses the concepts which form the core of the VO, the current set of international standards which define how those ideas should be realised in detail, and the organisational structures through which those standards will be implemented for the WSA and VSA.
- Section 4 lists the requirements from the SRAD which are to be met through the VO

- Section 5 describes, for each of the requirements listed in Section 4, the current thinking on how it will be met, including estimated timescales, where such are available.
- Finally, Section 6 presents a summary and conclusions.

3 THE VIRTUAL OBSERVATORY

3.1 Introduction

The Virtual Observatory may be defined [7] as *"a system that allows users to interrogate multiple data centres in a seamless and transparent way, which provides new powerful analysis and visualisation tools within that system, and which gives data centres a standard framework for publishing and delivering services using their data"*. The achievement of this goal requires the definition of, and adherence to, a number of interoperability standards, which will be supported world-wide, reflecting the international character of the astronomical community and the global distribution of the resources which must be federated into the VO.

The standards agency for the VO is the International Virtual Observatory Alliance (IVOA [8]). It coordinates VO developments worldwide and its members are national and regional VO projects: in mid-September 2006 the IVOA has 16 members, representing Armenia, Australia, Canada, China, France, Germany, Hungary, India, Italy, Japan, Korea, Russia, Spain, the UK, the US, and Europe. The UK is (through PPARC) a member of Euro-VO [11], which coordinates the development of a Virtual Observatory for Europe, and involves the ESO, ESA and funding agencies for astronomy in France, Germany, Italy, the Netherlands and Spain, as well as the UK.

The IVOA members differ greatly in size and focus, but all have signed up to a common "road map" for VO development and all contribute, to varying degrees, to the definition of standards, through an open process reminiscent of that by which the World Wide Web Consortium [9] oversees the development of the standards which underpin the WWW. In the IVOA scheme, a standard starts off as a *Working Draft*, before becoming a *Proposed Recommendation*, if endorsed by the relevant IVOA Working Group, and, finally, a *Recommendation*, once it has been approved by the IVOA Executive. The IVOA Executive, in turn, may propose that IVOA Recommendations are considered for adoption as IAU Standards, through IAU Commission 5, reflecting the paramount role that the IAU [12] plays in professional astronomy.

While all the IVOA members adhere to a set of common standards, there can be differences in how they implement them. Of most direct relevance to the WSA and VSA is the implementation by AstroGrid [10], which is the UK's VO development programme. The first two phases of AstroGrid development were funded by PPARC through its e-Science programme, and the AstroGrid consortium is currently seeking a third phase of funding from PPARC to support operation of a working UK Virtual Observatory running software development during the first two phases of the project. The core of the AstroGrid consortium is a group of major UK astronomical data centres – in Cambridge, Edinburgh, Leicester and at Jodrell Bank, MSSL and RAL – and this core has been supplemented by institutions (QUB, Exeter, Leeds, Portsmouth) providing particular expertise to different phases of the AstroGrid project. AstroGrid is also a member (formally through the Universities of Cambridge, Edinburgh and Leicester) of the VOTECH [13] project, which was the first Euro-VO activity to be funded and which is conducting a series design studies for a European VO.

3.2 VO development in Edinburgh

WFAU is playing a leading role in the development of the VO. Andy Lawrence is the Project Leader of AstroGrid and the Lead Investigator for VOTECH, while Bob Mann leads VOTECH's Data Exploration Design Study (DS6), and, by late 2006, there will be five research staff working within WFAU on

VO-related topics. These five research staff are funded through three different sources: Kona Andrews is supported by AstroGrid and is a core member of its development team; John Taylor is a researcher for the VOTECH DS6 team; Mark Holliman is jointly supported by VOTECH and WFAU's PPARC Rolling Grant, as a web developer and systems adminstrator; Brian Walshe will join the VOTECH DS6 team on 1 October 2006, to research data mining in the VO; and Lorenzo Rimoldini will join WFAU on 1 November 2006 as a data mining applications scientist on the Rolling Grant. While none of these staff are directly funded by the VDFS, all are – to varying degrees – contributing to the integration of the WSA and VSA into the VO. The most direct contributions are from the two staff who receive funding from the WFAU Rolling Grant: Mark Holliman leads the deployment of VO services running against WFAU's science archives, while Lorenzo Rimoldini will be developing data analysis services to help users exploit to the full the data in those science archives.

This VO development team interacts closely with the other members of WFAU staff who are responsible for the development and operation of the Unit's science archives. WFAU's science archive and VO teams each hold a weekly meeting, but many members of the Unit usually attend both meetings, which helps to ensure that the two sides of WFAU's work remain integrated. This close interaction between the science archive and VO teams within WFAU has already ensured that the requirements of the WSA and VSA influence the design priorities within AstroGrid and VOTECH, and that the WSA and VSA receive early deployments of stable AstroGrid software: this is discussed in greater detail in Section 5 below.

3.3 The current status of the VO

As outlined above, the development of the VO proceeds through parallel tracks of standards definition and implementation, and here we briefly summarise the status of each activity at the time of writing (in mid-September 2006).

3.3.1 IVOA standards

There are currently fourteen technical specifications within the IVOA standards process (see [14]): six *Recommendations*; three *Proposed Recommendations*; and five *Working Drafts*.

The first *Recommendation* defines the **IVOA document process** itself. Next comes the specification for **VOTable**[15], which is an XML format for tabular datasets. The VOTable format has been designed with the structures of FITS[16] tables in mind: each VOTable contains a header section comprising metadata records, followed by one or more tables containing data values. The specification provides three ways in which the data can be included within the VOTable document. The simplest method sees the data stored in a pure XML representation in which the table of data values is constructed in a manner very similar to an HTML table, with <TR> and <TD> elements delimiting the rows and columns. The repetition of these tags makes this pure XML representation very verbose for large tables, so two more compact options are also defined: the data may be included in the XML document in binary form (e.g. gzipped, or using base64 encoding) or they may be stored in an external FITS binary table whose URL is included in the VOTable. Since the publication of the first version of the VOTable specification in 2002, the VOTable standard has been widely adopted within the VO community: many archives now offer VOTable as an ouput data format for query result sets, and many software tools (e.g. TOPCAT [17]) now accept inout data files in VOTable format.

The third *Recommendation* covers **Resource Metadata**[18] in the VO. A *resource* in this context is anything within the VO that can be uniquely identified, and, most importantly, this includes datasets and application software. The resource metadata standard lies at the heart of one of the key components of the VO, which is the *registry*: if the VO is to provide users with access to resources distributed across the Internet, they need a means of finding those which are of interest to them and that is done by querying a registry, which stores resource metadata entries for every resource accessible through the VO. The resource metadata standard illustrates one of the key principles of VO development, which is the re-use of well-established existing standards from outwith astronomy where possible. Many of the basic metadata entries – e.g. those recording the Title, Publisher and Creator of a resource – are drawn from the Dublin Core [19] metadata element set, which is a controlled vocabulary of terms, originating within the digital library community, but which has become a *de facto* standard across a much wider range of disciplines. Where Dublin Core elements do not exist, the resource metadata specification defines additional elements – e.g. for spatial and spectral coverage – so that the full metadata record for a resource stored in the registry should provide all the information that a user needs to decide whether it is of interest.

The final three *Recommendations* all relate to **Unified Content Descriptors** (UCDs), which are a controlled vocabulary of semantic types designed to aid the interoperability of data sources in the VO. The need for UCDs arises from the way that the VO is being built as a federation of existing data resources most of which were constructed originally as single, standalone archives, with no expectation that they would be queried in tandem with other archives. As a result, there is no standard set of column names used in these archives. So, the Right Ascension column may be called "RA" in one database, "RA_J2000" in a second and "Alpha" in a third. This level of variation may not confuse an astronomer, but it is a problem for the sort of computer-mediated interoperability envisaged for the VO. UCDs solve this problem by enabling archive curators to assign to a column in their database a label which essentially says, in the above example, *"This column is Right Ascension"*, so that users and their software agents understand what kind of quantity is stored in a particular database column, irrespective of the name which has been assigned to that column. Many data centres now include UCDs as part of the metadata describing columns in tabular datasets returned in response to query. This is readily implemented for VOTable output, since *ucd* is one of the attributes allowed to be present in the <FIELD> element which records the metadata for a column in a table stored in a VOTable document.

Two of the three *Proposed Recommendations* define quantities which form part of the VO's conceptual infrastructure: the **IVOA Identifiers** standards specifies a way for assigning globally unique identifiers to resources in the VO, while the **Space-Time Coordinate Metadata for the Virtual Observatory** specification provides a standard representation for metadata relating to space and time. The third *Proposed Recommendation*, **VOEvent**, defines a standard information packet corresponding to transient celestial events. It is primarily intended to facilitate real-time follow-up of events like gamma ray bursts and novæbursts, and so is of little relevance to the WSA and VSA, since both these science archives ingest data too long after they are taken for the triggering of follow-up observations of transient phenomena to be possible.

The five current *Working Drafts* cover a number of the most important aspects of the VO, and, in most cases, the fact that they have not progressed beyond the *Working Draft* stage indicates the difficulty that there has been in achieving consensus across the VO community as to their contents. The two prime examples of this are the *Working Drafts* on **Astronomical Data Query Language** (ADQL[20]) and the **SkyNode Interface** [21]. As noted above, most of the archives which are being federated through the VO were developed separately, so an important part of the VO is the creation of a standard method of querying them (both singly and jointly) and returning results from those queries. An early prototype implementing such functionality was the SkyQuery.Net [22] system developed at Johns Hopkins University. This provided a portal within which users could run queries (written in SkyQL, an extension to SQL) against a number of distributed datasets. SkyQuery.Net was an important demonstrator of some of the central ideas behind the VO, as well as a very early example of the use of web service technologies in astronomy, but it made heavy use of particular proprietary Microsoft technologies – notably having all data loaded into SQLServer databases and using the .Net framework for the web service layer – which made it inappropriate as a final solution for the VO community, where open source technologies are favoured, to aid interoperability. It was necessary, therefore, to seek to generalise the SkyQuery.Net system into something that could be used more widely in the VO, and this was started with the initiation of the SkyNode interface, which defines a standard set of services which a data centre must provide for each of its data resources, and ADQL, which was a modification of SkyQL and was intended as a first attempt at a standard query language for the VO. Both of these specifications are about to undergo a major overhaul at the time of writing and neither has been widely adopted within the VO community to date.

One of the *Working Drafts* which has been more widely adopted is the **Simple Image Access** protocol (SIAP[23]). As its name suggests, this provides a simple mechanism by which data centres can publish image archives. At the heart of SIAP is a two-stage process: a user submits to the image archive a query in the form of a rectangular area of sky, and receives in return a VOTable which contains URLs for a set of images contained in the archive in that area of sky; and then the user can then use URLs from that list to retrieve the images of interest. Several variants on SIAP exist, corresponding to different ways of defining the set of images whose URLs are returned to the user: the image archive may offer a cut-out service, which will return an image (either pre-existing or created on-the-fly) covering the full rectangular region specified in the query; or the archive may return a list of any image(s) it stores which cover the query region, and this could cover either a sky survey archive with a single or a few images in each region of sky or a telescope archive, which might have many hundreds of pointings in some areas and none in others.

The *Working Draft* on the **IVOA Single-Sign-On Profile** is the first step towards developing a security infrastructure for the international VO, through specifying an approved mechanism for authenticating users issuing calls to VO services: the issue of security in the VO will be discussed in greater detail in Section 5. The final *Working Draft* covers **VOResource**, which is a specific XML encoding for the metadata elements specified in the Resource Metadata *Recommendation*.

3.3.2 AstroGrid

The parallel tracks of standards definition and prototype implementation do not run in perfect synchronisation within the VO, nor can they, since, for example, AstroGrid is committed to delivering a working VO system for the UK astronomical community on a timescale which is inconsistent with seeing specifications describing all aspects of the VO pass through the IVOA standards procedure. It is inevitable, therefore, that in many places what AstroGrid implements runs ahead of the IVOA standards, and, in some cases, it lags behind, since AstroGrid has not implemented some standards currently in *Working Draft* or *Proposed Recommendation* form, since it wants to see them changed significantly before being accepted as *Recommendations*.

What follows is a brief description of AstroGrid at the time of writing. The first thing to note – and to be borne in mind when reading the extracts from the SRAD in Section 4 – is that AstroGrid is built using web services. Since it was funded through the e-Science programme, there was considerable pressure on AstroGrid to adopt Grid technologies (e.g. Globus Toolkit 2 and the like) during its early stages of development. However, after some testing, it was clear that the majority of Grid software was not sufficiently robust at that stage to be usable for a production system, so AstroGrid decided to use only web services, which are a much more well-established technology, with a wide user base in the commercial IT sector, as well as in academia. This proved to be a very prescient decision, since, with the advent of the Web Services Resource Framework (WSRF[24]), the Grid community has moved into much closer alignment with the commercial use of web services.

As discussed in Section 3.3.1, the Registry is one of the key components of the VO, and AstroGrid deploys two classes of IVOA-compliant registry, namely **Publishing Registries** and **Harvesting Registries**. Data centres (or even individual VO users, if in possession of an appropriate AuthorityID, as described in the *Proposed Recommendation* on IVOA Identifiers) can use a publishing registry to register resources (e.g. datasets or application software) under their control. If appropriately configured, the contents of such registries can be copied into a harvesting registry, so that a network of such harvesting registries comprises a repository of all registered resources, with the possibility for fault-tolerance through redundancy. Most of the data resources currently published within AstroGrid are databases containing catalogues of source attributes, like the WSA and VSA. These may be queried using the **DataSet Access** (DSA) component, which accepts queries in a particular version of ADQL.

At an early stage in AstroGrid's development it became clear that the project would have to wrap – and offer users a simple means of wrapping – many existing command line tools to expose them as web services. In practice what this means is providing for each service the WSDL[25] contract which describes what operations it performs in terms of what messages it will exchange with other services and how to call each operation. As AstroGrid staff started to wrap legacy command line codes as web services, it became clear that this process of WSDL-generation was very repetitive, and, also, that the WSDL contract itself did not specify all the information about the service that might be required later. So, AstroGrid developed the **Common Execution Architecture** (CEA), which is both a restriction on what can be expressed in WSDL (so that *all* services in AstroGrid are described by one of a small set of CEA WSDL documents) and an extension to it, by adding functionality which makes it possible to run the CEA-wrapped application code asynchronously (in contrast to standard web service systems, which are built on synchronous calls). One of the motivations for adding in the capability for asynchronous operations is that many tasks which users will want to run in the VO may be very time-consuming, so, in the standard, synchronous web service model, they could be timed-out and fail. The likely prevalence of these long-running jobs within the VO implies that users will want to be able to use the VO non-interactively, by launching a workflow which links together a number of operations and automatically executes them in turn, rather than requiring the user to set off each manually. To that end, AstroGrid has developed a **Job Execution System** (JES), which takes as input a workflow file, specifying a number of tasks in terms of calls to CEA services with particular inputs and outputs, and which turns the input workflow into a transcript which logs information as each constituent step in the workflow is executed. The information logged includes the names and locations of intermediate data files, which means that complex workflows which crash part-way through do not need to be re-run from the start. Many of the steps in a workflow will require the reading from, or writing to, temporary data files, and AstroGrid has implemented a distributed file storage system, called **MySpace**, to aid that process. MySpace is composed of two parts: the FileStore component corresponds to a physical unit of disk space; and the FileManager component looks after the movement of files between FileStores. The user views her holdings in MySpace as a single logical collection, even though they may be distributed across a number of FileStores in different parts of the country, thereby simplifying her task in managing her files within AstroGrid.

Users interact with the AstroGrid web services in two ways. Firstly, via a browser-based **Portal**, which is now deprecated, or, secondly, using the **WorkBench**, which is a Java client-side tool. Underpinning the Workbench is the **Astro Runtime** [26], which calls AstroGrid web services on the user's behalf, as well as providing a range of user-support tools, such as a **query builder** (which uses registry information to help users create ADQL queries to run against particular data resources exposed via the DSA), a job monitor (for tracking the status of jobs submitted through JES) and a MySpace browser. A particular bonus is that the user only has to log into the Workbench once to be authenticated with any service requiring credentials. These user credentials are managed by the **Community** component. Each user is a member of a community and when a user logs into AstroGrid, supplying a username, password and the name of the community in which they are registered, that community's server is queried to check that the user with that username and that password really is a member of that community. This authentication is, in itself, provides a very basic level of security (in that ensures that only registered users are using the system) but it forms the basis of more thorough system, which supplements authentication with authorization, to respect proprietary data access rights.

4 VO-RELATED REQUIREMENTS FROM THE SRAD

In this Section we gather together in one place all the requirements from the SRAD which have a bearing on the VO-enabling of the VDFS science archives. They are taken from Sections 4, 5 and 6 of AD01 and read as follows:

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

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

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

A2:
Archived data must be uncorruptable by Science Archive users.
Rationale: Scientific exploitation will be compromised if data are corrupted.

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

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

5 MEETING THE VO-RELATED REQUIREMENTS

In this Section we discuss how the VO-related requirements of Section 4 are being met, but before we do that, we note a few principles guiding the VO-enabling of the VDFS Science Archives.

Firstly, it is not within the remit – or the funding capabilities – of the VDFS to develop VO software: the VDFS and AstroGrid projects are closely related, but the distinction between them must be respected and neither can be allowed to stray into the realm of the other. This point has been highlighted by oversight committees and must be borne in mind by readers of this document.

Secondly, the VDFS team will not risk a negative impact on operation of its science archives caused by immature VO software. This means that VDFS staff must err on the side of caution when deploying AstroGrid software and will only deploy those components which they believe to be verified to be stable and functioning properly. This cautious approach may mean that the VDFS science archives do not, at any given time, have deployed against them all available VO software, but the VDFS archiving team will be pro-active in engagement with AstroGrid and will seek to be early adopters of mature AstroGrid software.

The six requirements listed in Section 4 can be considered in two groups, covering science archive functionality (T4, T5) and security (A1, A2, A3). We discuss these two groups in turn:

5.1 Science Archive Functionality

T4 requires that the VDFS-v1 (WFCAM) conforms to *"existing VO standards"* and, more specifically that VDFS-v2 (WFCAM) conforms to *"internationally agreed VO standards in access tools, data product formats and transfer protocols"*. The current status of conformance with the extant IVOA standards listed in Section 3.3.1 is as follows:

- **VOTable**. VOTables have been offered as one of the output data formats for the WSA since its inception, and will similarly be offered for VSA data in due course. The VOTable files produced by the WSA have been validated against the VOTable Document Type Definition to ensure that they do adhere to the VOTable specification exactly.

- **Resource Metadata**. A full resource metadata description for the WSA has been generated, and scripts exist to produce updated resource metadata if/when the WSA schema is changed.

- **UCDs**. UCDs were generated for all columns in the WSA as part of its schema definition phase. These have not yet been updated to "UCD1+" format, because this new standard is not yet widely adopted, and it was considered a low priority to implement this change before a clear requirement from users was perceived. UCDs are not included in the VOTables produced by the WSA web interface, which is acceptable, since the "ucd" attribute of the <FIELD> element is optional.

- **ADQL/SkyNode**. The small world-readable database from the WSA Early Data Release is accessible via a preliminary release of a version of the AstroGrid DataSet Access (DSA) component which features authentication. The lack of a security infrastructure within AstroGrid (see Section 5.2 below) has precluded the publication of the remaining data in the WSA as there is currently no way of ensuring that their proprietary data rights are respected when queried through AstroGrid.

- **SIA**. The absence of an AstroGrid component implementing the IVOA Simple Image Access protocol has prevented the publication of WSA image data through the VO. AstroGrid is currently developing an image access component, which will include an SIA service as well as a service which extended SIA to require authentication and authorization, and once this is available (in early 2007) and has been thoroughly tested, it will be deployed against the WSA to meet requirement T4.

T5 is essentially a specialisation of T4 concerned with query interfaces. As noted above, the AstroGrid DSA component – which will allow *complex queries* in the sense of T5, through use of ADQL – will be deployed against the WSA as soon as it is verified that it can support the authentication and authorization needed to implement the proprietary access restrictions on the WSA. It is anticipated that this will occur within the next few months, in which case the VSA should feature access via DSA from the outset.

5.2 Science Archive Security

A1 requires that archived data must be accessible only by validated users, while A3 commits the VDFS to enforcing proprietary rights on a per-frame basis. As noted above, the delay in the deployment of a security infrastructure for AstroGrid has caused a corresponding delay in the meeting of this requirement as it relates to VO access. Taken together, A1 and A3 require the presence of authentication and authorization functionality in AstroGrid. As described in Section 3.3.2, authentication is performed through the Community component in AstroGrid, while authorization is not yet implemented in any AstroGrid component. Since access to proprietary data had to be possible from the launch of the WSA, it was necessary to develop a temporary security infrastructure to make up for the lack of such a component in AstroGrid. This was delivered in the form of a username/password combination which users must supply before being granted access to the query webpages for proprietary WSA data. The list of valid users to be assigned a username/password pair was generated from data supplied by representatives within each UKIDSS institution and was implemented as a database table. Once authentication is provided by AstroGrid, and UKIDSS members are signed up to their local Community, this existing database table can be converted into an access control list providing authorization.

A2 demands that archive data be uncorruptable by users. This requires the DSA components deployed against the VDFS science archives to allow only read-only access to these databases. This is currently the normal behaviour in the DSA, but a version of that component which allows users to write to temporary tables in a database is under develoment, and it is important that this version should not be deployed against the VDFS science archives themselves, even if it is used to provide users of the archive will temporary databases which can be used in conjunction with those science archives.

5.3 Future plans

VDFS-v3 (WFCAM) envisages a significant enhancement in the data analysis functionality offered as part of the WSA, and much of this will be offered within the context of the VO. As mentioned in Section 3.2, WFAU are about to make two new appointments, to a position within the VOTECH dedicated to researching VO data mining, and to a ”data mining applications scientist” post, which will cover the development of data analysis algorithms to be run on the WSA and VSA. It is through these two new positions that the second piece of new funtionality listed in Section 7 of the SRAD will be delivered, namely *”server-based data analysis tools, e.g. cluster analysis, PCA, etc”*. These tools will be implemented as CEA web services, so that they can be called as part of an AstroGrid workflow. The third new piece of functionality – *”a system to allow uploadable user-specified analysis algorithms”* – has been the subject of a study performed by WFAU in cojunction with staff from the Digital Curation Centre (DCC[27]), which concluded that there is a series of measures which can be taken to ensure the integrity of a database which allows the upload of user-specified analysis algorithms. Some of these measures are specific to particular languages in which the algorithms may be coded – e.g. making use of the sandboxing provided by the Java Virtual Machine – while others (e.g. assigning restricted rights to particular users and groups) can be implemented at operating systen level or within the database management system itself. The most sophisticated option is to use *”proof -carrying code”* which brings with it a verifiable statement of its resource usage, but this may prove difficult to apply to the WSA and VSA as it would currently require users to annotate with special tags source code written in a restricted subset of Java, and users are unlikely to take the trouble to learn the restrictions to the language or the annotations required unless they really have to do that to make use of the WSA and VSA. The final piece of new functionality envisaged for VDFS-v3 (WFCAM) is the recasting of *”existing web services as grid services, in order to allow queries and analysis using shared managed resources with other data centres”*. As mentioned in Section 3.3.2, the distinction between web and grid services has largely disappeared with the adoption of WSRF by the Grid community. The VDFS science archives will continue to use AstroGrid services, and to the extent that they start to use WSRF, this requirement will be satisfied.

6 Summary and Conclusions

This document has described the VO-enabling of the WSA and VSA. At present the WSA implements those VO standards which have wide acceptance and meet the requirements of WSA users, and the close interaction between WFAU’s science archive and VO development teams ensures that the WSA and VSA continue to be early adopters for AstroGrid software components, once they are stable.

References

[1] WFCAM Science Archive overview document; http://www.roe.ac.uk/~nch/wfcam/VDF-WFA-WSA-001-I1/VDF-WFA-WSA-001-I1.html

[2] WFCAM/VISTA Science Archive Development, http://www.roe.ac.uk/~nch/wfcam/

[3] The UKIDSS Proposal; http://www.ukidss.org/sciencecase/sciencecase.html

[4] The UKIDSS Infrared Deep Sky Survey (UKIDSS); Lawrence, A. et al., 2006, MNRAS, submitted; astro–ph/0604426

[5] Usages of the WFCAM Science Archive (Hambly, N.C. & Bond, I.A., 2003); http://www.roe.ac.uk/ nch/wfcam/misc/wsausage.html

[6] VEGA: UK Excellence in Data Processing and Archiving — providing VO Processing for VISTA and GAIA (a revised proposal submitted in response to the 2003 PPARC e–Science Announcement of Opportunity)

[7] The AstroGrid Project Vision; http://wiki.astrogrid.org/bin/view/Astrogrid/RbProjectVision

[8] The International Virtual Observatory Alliance website; http://www.ivoa.net

[9] The World Wide Web Consortium website; http://www.w3.org

[10] The AstroGrid website; http://www.astrogrid.org

[11] The Euro-VO website; http://www.euro-vo.org

[12] The International Astonomical Union; http://www.iau.org

[13] The VOTECH Project website; http://eurovotech.org

[14] IVOA Technical Specifications; http://www.ivoa.net/Documents/

[15] VOTable specification; http://www.ivoa.net/Documents/latest/VOT.html

[16] Flexible Image Transport System (FITS); http://fits.gsfc.nasa.gov/

[17] TOPCAT: Tool for Operations on Catalogues and Tables; http://www.star.bris.ac.uk/ mbt/topcat/

[18] Resource Metadata for the Virtual Observatory; http://www.ivoa.net/Documents/latest/RM.html

[19] Dublin Core Metadata Element Set; http://dublincore.org/documents/dces/

[20] IVOA Astronomical Data Query Language; http://www.ivoa.net/Documents/latest/ADQL.html

[21] IVOA SkyNode Interface; http://www.ivoa.net/Documents/latest/SNI.html

[22] SkyQuery.Net; http://www.skyquery.net

[23] Simple Image Access; http://www.ivoa.net/Documents/latest/SIA.html

[24] Web Services Resource Framework; http://www.oasis-open.org/committees/wsrf/

[25] Web Service Description Language; http://www.w3.org/TR/wsdl

[26] AstroGrid's Astro Runtime; http://www2.astrogrid.org/desktop

[27] Digital Curation Centre; http://www.dcc.ac.uk

7 ACRONYMS & ABBREVIATIONS

ADnn : Applicable Document No. nn
CASU : Cambridge Astronomical Survey Unit
FITS : Flexible Image Transport System
GPS : Galactic Plane Survey (UKIDSS)
JAC : Joint Astronomy Centre
LAS : Large Area Survey (UKIDSS)
SQL : Structured Query Language
UKIDSS : UKIRT Deep Infrared Sky Survey
VDFS : VISTA Data Flow System
VISTA: Visible and Infrared Survey Telescope for Astronomy
WFAU : Wide Field Astronomy Unit (Edinburgh)
WFCAM : Wide–field infrared camera for the UK Infrared Telescope
VO : Virutal Observatory
VOTable : XML format developed for astronomical data for the VO
VSA : VISTA Science Archive
WSA : WFCAM Science Archive
XML : eXtensible Markup Language

8 APPLICABLE DOCUMENTS

AD01	Science Archive Science Requirements Analysis	VDF-WFA-VSA-002-I1 Issue: 1.0 Sept 2006
AD02	Science Archive Database Design	VDF–WFA–VSA–007 Issue: 1.0, September 2006

9 CHANGE RECORD

Issue	**Date**	**Section(s) Affected**	**Description of Change/Change Request Reference/Remarks**
Draft 1	Nov 2005	All	New document based on old WSA SRAD
Draft 2	Sep 2006	3,5,6,7	In preparation of UK VDFS FDR
1.0	Sep 2006	Minor tweaks by NCH	Finalised for VDFS UK FDR

10 NOTIFICATION LIST

The following people should be notified by email whenever a new version of this document has been issued:

WFAU:	P Williams, N Hambly
CASU:	M Irwin, J Lewis
QMUL:	J Emerson
ATC:	M. Stewart
JAC:	A. Adamson
UKIDSS:	S Warren, A Lawrence