International

Virtual

Observatory

Alliance

The IVOA in 2006: Assessment and Future Roadmap

IVOA Technical Coordination Committee

Version:
2006 June 6

Author:
Roy Williams

In consultation with:
IVOA Technical Coordination Committee:
Mark Allen, Francoise Genova, Bob Hanisch, Gerard Lemson, Tony Linde, Jonathan McDowell, Reagan Moore, Maria Nieto-Santisteban, Francois Ochsenbein, Masatoshi Ohishi, Andrea Preite Martinez, Guy Rixon, Yuji Shirasaki, Doug Tody

Abstract:
This document is the result of a study by the IVOA Technical Coordination Committee with the intention of coordinating the IVOA Working Groups and Interest Groups. The study was commissioned by the Executive Committee in May 2005 and revised in June 2006, with the objectives of:

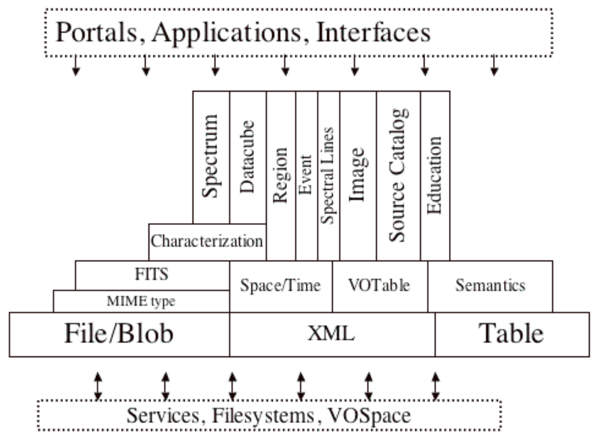
- Elaborating the IVOA Architecture
- Building a roadmap for the IVOA that is a union of roadmaps for the Working Groups and Interest Groups.
- Ensuring productive crosstalk of the WG/IG so that workpackages cover relevant ground, but also do not overlap.
- Evaluating dependencies of one WG/IG on another and minimizing impact.
- Attaching milestones to the WG/IG roadmaps, representing planned achievements and target dates.
- Ensuring an effective evaluation of proposed standards during the RFC period.
- Providing a continuous reporting checkpoint to the IVOA Executive Committee on roadmap status.

IVOA Architecture

The IVOA architecture is patterned on a combination of three elements: View, Service, and Registry. Views are how people perceive data, Services deliver the data and metadata, and the Registry allows publishing and discovery of services.

A View is the human side of the data, it has a data model, it is a way of thinking of astronomical data, a frame of reference, a focus. The diagram at the right shows this hierarchy. Often a user portal will be built around one or more views, for example Catalog and Image. Thus in Figure 1, we see the human interaction at the top, the portals, applications, and interfaces, each relying on several views, the principals being *Image*, *Catalog*, and *Spectrum*. These in turn are built with lower-level views, such as Space Time Coordinates (STC), standard controlled vocabulary (Semantics), and standardized rich format for expressing tabular data. These are built in turn on simpler concepts such as FITS files and XML.

Figure 1 : Views

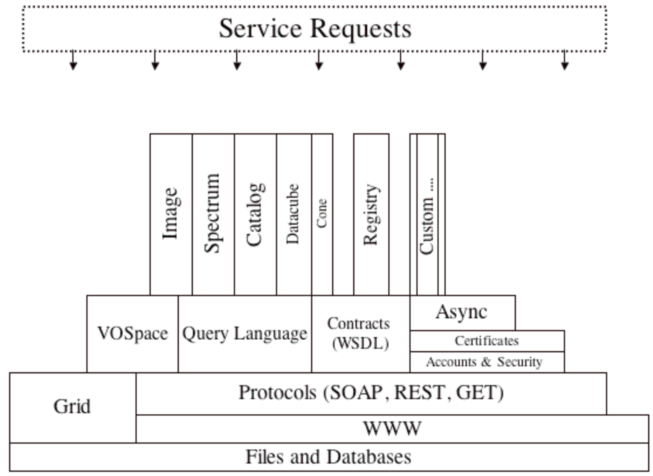


Specifically, a view is a data model, with interfaces and file formats for representing it, together with associated data models/interfaces/formats; these act as concrete containers for data -- in shapes that atsronomers already understand. For example, a Source Catalog view includes positions, fluxes, and classifications of sources, which brings in coordinates, bandpasses, standard vocabulary, provenance in the literature, the table format, and a host of other data models and standards. Understanding what the IVOA means by a particular view and its associations: that is what allows an astronomer to construct and resolve scientific questions in the IVOA framework.

Characterization is a view where datasets exist in an abstract multi-dimensional space of position, wavelength, time, polarization, etc. A *Region* means a compact subset of the sky manifold, perhaps representing the coverage of an image or survey. *Event* is a view of transient sources where rapid changes have position and time measurements. The *Image* view sees sets of scalar images, that cover a part of the sky, distinguished by the name of a bandpass. The Datacube view implies a multi-dimensional array of data, where two of the axes represent sky position.

The Services (Figure 2) are constructed to implement and expose the astronomical data objects that astronomers want, the most important being Image, Catalog, and Spectrum. The Cone search is a simple exemplar of a service that has proved to be a useful pedagogical example of a standard service. The Registry exposes a number of services for publishing and discovery that are elaborated below. Services can also be custom, so that data can be delivered in new ways. The IVOA service infrastructure relies on some common elements, such as VOSpace, that allows users to store, retrieve, and share data; the formal WSDL contracts that a service exposes so that clients can automatically configure an interaction with the service. A standard query language based on SQL allows effective use of database technology. Accounts and security allow access to private data, and the asynchronous services allow long-running batch jobs to be executed and monitored. Logging of services is essential to gauge how services are utilized, and thereby improve the infrastructure. All of these services rely on a bedrock of protocols, from the simple GET to the sophisticated SOAP, which in turn rely on the web (http) protocol. The data Grid provides bulk access to data, where parallel streams and other technologies allow massive processing and delivery.

Figure 2: Services

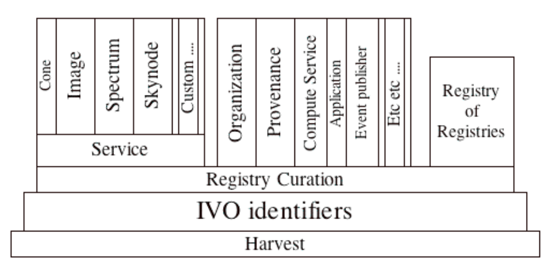


The IVOA supports a distributed global Registry (Figure 3) to allow publishing and discovery of standard service providers and other resources. Registry nodes are harvested regularly from each other, so that each contains the whole IVOA: not only its own published resources, but also those published at every other registry node. A registry of registries will allow an overall view of all registry nodes. The IVOA supports a global identifier system; these URI strings can be used to refer to any IVOA registered resource, and any registry can resolve the identifier to the metadata that it represents. Each registry entry contains standard curation information based on the ubiquitous Dublin Core standard, defining who published the record and why. A major thrust of the registry is to contain service definitions as discussed above; other registry resources may represent organizations, applications that can be used in a workflow, etc. A group of specialist astronomers may define non-standard resource types, and such records will be harvested by all registry nodes, although such nodes may see these resources as opaque, i.e. not queryable. There is a distinction in the IVOA between a dataset and the service that delivers the dataset: thus the provenance of the dataset is represented as a distinct record: who took the data and why and how, with citations to the literature.

It would be valid to consider the Registry as just another View of the Virtual Observatory -- it has data models, interfaces, and file formats – but it is sufficiently different in several ways that it is justified to think of it separately. The registry is a special view of the VO because it is the mechanism for communication within the IVOA, so that a published service can be found by another; we think of it as the glue that connects the desired views – the scientific payload -- to the services that provide the data to instantiate those views.

In general, the Views are elaborated through the Data Models and Semantics Working Groups; the Services through the Data Access Layer, Query Language, and Grid/Web Services Working groups; and the Registry through the Registry Working Group.

Figure 3: Registry



Now that the essential nature of standards and schema have become clear, it is time to concentrate on fixing those standards and implementing an operational global Virtual Observatory.

General Recommendations

(1) Crossmatch standards (TCC): The Interop meeting in Victoria (May 2006) generated a wide-ranging discussion on the nature of astronomical crossmatch – meaning a fuzzy join of source catalogs in order to associate multiple observations of the same astrophysical object. It is a core operation of statistical astronomy, and a widely-implemented crossmatch standard is essential if catalogs are to be richly federated. The standard crossmatch cannot, of course, solve a scientific problem to its conclusion, but should be a way to obtain a "good selection of candidates" to which the scientist can apply custom processing. Therefore we **recommend** that the Technical Coordination Committee set up a small international "tiger tem" to investigate and report on crossmatch with a deadline of October 2006, specifically covering use-case, algorithm, complexity, and notation:

- Produce a number of scientific use cases for crossmatch algorithms, such that the convex hull of these use cases covers most practical applications;
- Research and enumerate the different algorithm components, including distance and the chi-squared (Szalay et al) algorithm, including algorithms that use non-spatial information;
- Evaluate the strengths and weaknesses of these algorithms with respect to the scientific use cases;
- Research and understand the technical implementation of these algorithms with relational database technology, assigning computational complexity to the algorithms;
- Evaluate the conceptual complexity of the algorithms by building suitable notation based on ADQL or SQL, and writing the use-case queries in terms of this notation;

The report should be online at the IVOA wiki, with a discussion section to allow the use-cases, algorithms, complexity measures, and notations to be evaluated by the whole IVOA.

(2) Simple Image Access (DAL): One of the early successes of the IVOA standards process was the Simple Image Access Protocol, a standard service based on a view of an image survey as a covering of the sky with a small number of named filters. This protocol has been widely accepted and implemented. Emerging in 2006 is a more sophisticated "Datacube Access" version, where the image can be a multidimensional datacube, the metadata aligns with the characterization data model, and other enhancements. We hope that the older, simple level can be retained as "Simple Image Access", in addition to the new protocol for datacubes, because (a) many sites will continue using the original standard, and (b) the simpler protocol can often do everything that is needed. We **recommend** continued support of "Simple Image Access" in the same form as it has been, and welcome the addition of the new datacube protocol.

(3) Spectral Model/Interface/Access (DM, DAL): The spectrum data model and access services have appeared as Working Drafts in the Data Models and Data Access Layer groups. We **recommend** accelerated implementations of these standards, and experiments on interoperability between these, which will lead to accelerated approval to Recommendation by the IVOA. In this way, the exposure of spectral data by data centers can be brought to the same level of maturity as image and catalog views.

(4) Source Catalog View (DM, VOQL): Many astronomical databases (but not all) are object catalogs. A choice of views is emerging between the "Table" and "Catalog" concepts. In the Table view, any database table can be exposed and its relational schema used to create queries; and in the Catalog data model, a table of astronomical sources is exposed in a standard data model – so that, for example, "positional error" is always written the same way rather than according to an arbitrary name chosen by the table author. We **recommend** (hope for) a vigorous discussion within the IVOA of this Source Catalog Data Model, with the objective of international agreement on a standard representation for a source catalog.

(5) Registry Implementation (Registry): As with many IVOA standards, it is time to finalize the schema for the Registry to enable a clear path to implementation. A new plan has been agreed at the May 2006 Interop, that elaborates the idea of *Service* into a family: the parent *Service* contains *Interfaces* and *Capabilities*.

- We **recommend** that this change in registry schema should be the last for a long time – at least the last schema change that would invalidate old records.
- We also **recommend** that the registry WG define and reach agreement on the scope of the registry in terms of the variety and granularity of metadata. Registries can cache detailed metadata on a regular basis, or maintain limited (but valid) metadata and fetch detail only when required.
- We also **recommend** that the "Registry of Registries" should be created immediately and/or advertised on the IVOA website, even if it is informal (a web page), so that information can be gathered at the same time as the formal specification is built.
- We hope to clarify and define closely the idea of annotation/augmentation of existing registry records by an entity that is not the author. We **recommend** that the Registry group provide use-cases for this concept.

(6) Registry Query Language (Registry, VOQL): Querying a registry of services is rather different, semantically, from querying a star catalog. The former may involve small data in complex schemas, and the latter large data in simple schema. The star catalog query is helped by specific language constructs (eg. Region of the sky) that may mean nothing in the context of the registry query. We **recommend** a sub-committee of the Registry and VOQL groups should examine the case for and against a separate query language for registry, that would be customized for registry queries and independent of future development of the catalog query language.

(7) Table and Catalog Access (VOQL, DAL): The Query Language group has made considerable advances in generalizing and standardizing levels of compliance and utility. For general catalog access, cone search, trivial as it is, has proven to be a good start as it provides easy access to data via a simple interface. SkyNode addresses the much harder problems of providing a general query language, crossmatching of large catalogs, and distributed cross matches. What is needed -- an intermediate approach for basic catalog access -- is something which provides both a language-based interface (ADQL) as well as a parameter-based interface more sophisticated than cone search, and eventually, data model mediation via standard catalog data models.

- In the language interface, a relational database is exposed through relational schema: the table names and table attributes, together with the ability to build a query using that metadata. We **recommend** the creation of a standard interface called Simple Table Access Protocol (STAP) that implements this view. It would be derived from the basic Skynode interface and the core ADQL language.
- In the catalog view, queries can be created within the Catalog Data Model, language-based and/or parameter-based, so that the same query can be sent to multiply-authored source catalogs, and the results returned in the context of that view. We **recommend** the creation of a standard catalog access service interface to support his view.
- The most sophisticated queries involve distributed cross-match, where multiple source catalogs generate associations of observations of the same physical object. See above for recommendation relating to crossmatch.

(8) VOSpace (GWS): The VOSpace effort within the Grid/Web services working group is building semantics, schema, interface, and prototype. The view and capabilities of the VOSpace is revealed at three levels of depth:

- Data are stored as files/blobs, but MIME types are recorded against them so that them may be understood after being fetched out of VOSpace.
- MIME types are used to allow access to parts of a file, or to allow dynamic reformatting during output from VOSpace.
- Data are stored in some way that makes their internal structure accessible through an alternate interface on the same logical service. E.g., data put into VOSpace as a VOTable become accessible via a sky-node interface.

The spaces themselves may be structured in three levels of federation:

- Data objects are siblings with no hierarchy and spaces are not linked.
- Data objects can be grouped and arranged in a hierarchy of directories.
- There are symbolic links between VOSpaces, allowing global federation.

We **recommend** the formation of a *VOSpace Use Cases* document, to more closely define the direction of this fine effort, and to differentiate it from related efforts in the grid community.

(9) Interoperable Security (GWS): Security and authentication is being implemented in several new efforts. The UK Astrogrid project has built a sophisticated workflow system for asynchronous computations and is adding authentication; a complementary project from the US NVO project is exploring the idea of “graduated security” for giving community access to high-performance computing. We **recommend** a study of these and other “grid” projects to promote interoperability.

(10) Space Time Coordinates (DM): An effect of a sophisticated data model can be the impression in the community that all levels of complexity must be understood before any part of it can be used. It would be better to have data models that can be used at different levels of sophistication. A jewel of the IVOA is the Space-Time Coordinate system specification, because of its rigor and accuracy. While it has become immensely more usable over the last year, it could be improved further by presenting a “toolkit” for expressing coordinates that *allows* rigor and accuracy rather than forcing a scientist to use accuracy and rigor even when there is are reasons against this.

(11) Units (DM): Most scientific quantities carry units, and data returned from IVOA services should also carry explicit unit information when not clear implicitly. Units should follow the IAU recommendation[1], which follows the SI convention. When a user makes a query based on a quantity, units can either be user-defined or fixed. In the former case, the user has the freedom to express the quantity in arbitrary units (eg. *calories per square furlong per hour*!), or an enumerated choice (eg. *Angstroms OR nanometers*). In the case of fixed units, the data model of the query is bound to specific units (eg *all angles must be in decimal degrees*). We **recommend** a study by the Data Model Working Group of how units are used in IVOA views and services, where it would be appropriate to simply fix the units, and where it is necessary to allow freedom of choice. In the latter case, the report should also recommend on how unit conversion is implemented: who is responsible and the nature of the software.

[1] Recommendations Concerning Units, http://www.iau.org/Units.234.0.html

Working Group Chair Responsibilites

- Each WG must have a clear Roadmap in a standard form - with planned achievements versus target dates (i.e. milestones)
- WGs should pay close attention to the top-level Technical Milestones, making sure each relevant milestone is inside the WG roadmap.
- There should be a checkpoint at each Exec Meeting and at each Interop Meeting
- For each checkpoint, the WG chair should provide (i) a very short text report (1-2 paras) (ii) a progress statement on each element of their roadmap
- The above reports will be requested 2 weeks in advance from the IVOA.

In addition to the above responsibilities for her own Working Group, the Chair is also responsible for active comment (1-3 paragraphs) on each request for comment (RFC) that has been issued by another Working Group.

Interest Group Chair Responsibilities

- Reporting by IGs should be relatively low key and informal. This informality is a key distinction between WGs and IGs. WGs are much more work, and need to deliver a product.
- IGs should provide verbal reports at each Interop meeting.
- The Interop organising committee should request these several weeks before the Interop Meeting.

Roadmaps for the Future

The current roadmap situation (May 06) is summarized in Table 1, the Working Groups and Interest Groups, and Table 2, the proposed roadmap for each WG/IG. Since one of the main objectives of the IVOA is production of standards documents, the status of these documents is called out in terms of what type of document is being produced and the stage it has reached in that production.

In Table 2, documents that are in progress or in the future are labeled by their status in the IVOA document sequence:

- **inWG**: Preparation within WG, meaning that a draft is being circulated among a subset (or all) of the WG, and that action is on the WG chair to ensure progress
- **WD**: A Working Draft is available on the IVOA Documents page, at level 1,0 or greater.
- **PR**: The chair of the Working Group has notified the Technical Coordination Committee and the IVOA Document Coordinator, and a 4-week comment period has started, with proper instructions for how to comment. This cycle can happen several times. Two interoperable implementations are needed for a standard to be considered as a PR.
- **REC**: The Executive Committee of the IVOA has moved this to a Recommendation.

In addition to the above document categories, working groups or other groups can also submit a **Note**, which is not an explicit part of the standards process.

Table 1: **IVOA Working Groups and Interest Groups**

Working/Int. Group		Chair	Current priorities
Applications	IG	Mark Allen	Various application news.
Data Access Layer (DAL)	WG	Doug Tody	Spectral Energy Distribution (with DM). Simple Spectral Access, Level 2 Image Access (datacube), Characterization and "generic dataset".
Data Curation and Preservation (DCP)	IG	Francoise Genova, Reagan Moore	Metadata formats and methods. Evaluating Preservation environments (eg Dspace, Fedora). Curation/maintenance of registries?
Data Models (DM)	WG	Jonathan McDowell	Spectral Energy Distribution (with DAL) Characterization (of observations) DM Space-Time coordinates (STC). Catalog DM Provenance (of observations) DM Spectral line (atomic line) DM
Event	WG	Roy Williams	Event Semantics WD 1.0 and schema. Prototypes and transport.
Grid-Web Services (GWS)	WG	Guy Rixon	Security, trust, single sign-on. Prototypes. VOStore and VOSpace. Asynchronous services and WSRF. Logging and support for services.
Query Language (VOQL)	WG	Maria Nieto Yuji Shirasaki	Astronomical Data Query Language (ADQL) as XML and script. SkyNode Interface methods. Integration with DAL
Registry	WG	Tony Linde	Resource Metadata, semantics and schema. Service Interfaces. Registry of registries Registering general services and applications. Query languages for the registry.
Semantics/UCD	WG	Andrea Preite-Martinez	Updating and agreeing UCD list. Workflow for changes to list. Role of ontology. Standard vocab for Process/Objects
Standards and Documents (SD)	WG	Bob Hanisch	Improved workflow for RFC process
Systems Architecture & Technical Coordination (TCC)		Roy Williams	Technical Coordination Committee: overlap, dependencies, RFC process.
Table	WG	Francois Ochsenbein	Parsers, implementations and bug fixes.
Theory	IG	Gerard Lemson	Data Modelling and Formats (Lemson et al); Access Protocol – N-body and mesh simulations Semantics and UCDs for Theory (Shaw et al).

Table2: **IVOA WG Roadmap May 2006**

Date	WG/IG	Standard	Status	Responsible
May-06	DAL	Simple Numerical Access	inWG	Lemson
Oct-06	DAL	Simple Image Access-V1.0	PR	Tody, Plante
Oct-06	DAL	Simple Cone Search 1.0	PR	
Dec-06	DAL	Spectral Line Access-V1.0	WD	Dubernet, Osuna
Dec-06	DAL	Simple Spectral Access-V1.0	WD	Tody, Dolensky
May-07	DAL	Datacube (SIA-Level2)	WD	Tody
2007	DAL	Simple Cone Search 1.1	inWG	
2007	DAL	Simple Catalog Access	PR	
2007	DAL	Generic Dataset	inWG	Tody
2007	DAL	SSA 1.1, includes Spectral Energy, Time Series	inWG	
Aug-05	DM	Atomic Line Lists-v1.0	WD	Dubernet, Osuna
Jul-06	DM	Spectrum-v1.0	WD	
Jul-06	DM	Space Time Coordinates-V1.3	PR	Rots
Sep-06	DM	Characterisation-V1.0	WD	Bonnarel, Louys
Dec-06	DM	Spectrum-v1.0	PR	
2006	DM	VOQuantity-V1.0	Note	McDowell, Berry, Dowler, Thomas
May-07	DM	Spectral Energy Density-V1.0	PR	McDowell, Tody
Mar-07	DM	Source Catalog Model	PR	
2008	DM	Dataset Registry Resource		
Jul-05	Event	VOEvent 1.0	WD	Seaman, Williams
Sep-06	Event	VOEvent --Transport	Note	Seaman
Oct-06	Event	VOEvent 1.1	PR	Williams, Seaman
Aug-06	GWS	VO Basic Profile -V1.0	PR	Schaaf
Aug-06	GWS	Common services	WD	Rixon
Aug-06	GWS	Universal Worker Service pattern	WD	Rixon
Jul-06	GWS	Logging and harvest	WD	Thakar
Sep-06	GWS	Single Signon Authentication V1.0	PR	Rixon
Sep-06	GWS	VOSpace V1.0	PR	Graham, Morris, Plante, Rixon
Jan-07	GWS	Universal Worker Service contract	WD	Rixon
2007	GWS	VOSpace V2.0	WD	Graham, Morris, Plante, Rixon
Oct-03	SD	IVOA Document Standards 1.0	REC	

Date	WG/IG	Standard	Status	Responsible
Jul-06	VOQL	Astronomical Data Query Language -v1.04	PR	Nieto, Shirasaki
Jul-06	VOQL	Skynode Interface 1.02	PR	Nieto, Shirasaki
Dec-06	VOQL	SkyNode Extensions	WD	Nieto, Shirasaki
Jan-07	VOQL	Asynchronous skynode	inWG	
Jun-05	Registry	VO-Identifiers V1.1	REC	Plante
Jun-05	Registry	Resource Metadata V1.1	REC	Hanisch, Linde
May-06	Registry	VOResource-V1.0	WD	Plante, Linde
Jul-06	Registry	Registry Of Registries	inWG	Plante
Jul-06	Registry	Registry Interface-V1.0	WD	Benson, Linde
Sep-06	Registry	VODataService v1.0	WD	
Sep-06	Registry	VOApplication v1.0	WD	
Oct-06	Registry	VOResource-V1.0	REC	Plante, Linde
Oct-06	Registry	Registry Interface-V1.0	REC	Benson, Linde
Nov-06	Registry	CEAApplication v1.0	WD	
Aug-05	Semantics	Unified Content Descrptors, V-1.10	REC	Derriere, Preite Martinez, Williams
Dec-05	Semantics	The UCD1+ controlled vocabulary Version 1.11	REC	Derriere, Preite Martinez
Jun-06	Semantics	Maintenance of the list of UCD words	REC	Derriere, Preite Martinez
Oct-06	Semantics	The IVOA Standard Vocabulary	WD	Preite Martinez
May-07	Semantics	Ontology of astronomical object types	WD	Derriere, Preite Martinez
Aug-04	Table	VOTable-V1.1	REC	Ochsenbein
Aug-06	Table	VOTable-V1.2	REC	Ochsenbein
Oct-06	Theory	UCD Extensions for Theory	Note	
Oct-06	Theory	Simple Numerical Access Protocol –v0.1	WD	Lemson