# Practical Interoperability in the Virtual Observatory

Markus Demleitner[1,2]

[1]Universität Heidelberg, Zentrum fürAstronomie;
[2]German Astrophysical Virtual Observatory GAVO

The Virtual Observatory (VO) is an international effort to run and develop a federated data infrastructure in Astronomy that is held together by a set of data and protocol standards. Consisting of a Registry, some 30000 interoperable services (which roughly correspond to data collections) comprising hundreds of millions of datasets (spectra, images, and the like) and hundreds of billions of table rows, as well as a set of clients and libraries consuming these services, it is widely used in the astronomical community.

In this contribution, I will give a condensed overview of the technologies behind the VO, with a particular emphasis on how the VO is not just a platform but a truly global infrastructure jointly shaped by data providers, software authors, and science users.

## 1 Introduction

The Virtual Observatory (VO) as a project started in the early 2000s (e.g., [3]) and has grown to become a global data infrastructure for Astronomy and Astrophysics, now encompassing data centers in about 30 countries[1].

The VO is *not* a website ("platform", "portal"), nor some sort of network of websites, nor a programme that tries to do everything astronomers might want to do with a computer. It is the main objective of this contribution to explain what it is instead and why it was designed in this way.

The VO could be defined as

- A few dozen standards for finding, accessing, using, and describing data, authored and agreed upon under the auspices of the International Virtual Observatory Alliance IVOA.
- The astronomy data centers publishing data using these standards; this includes almost all the major players like NASA, ESA, or ESO.

[1] There are 34 top-level domains in the access URLs of VO-registered services in March 2021.

Publiziert in: Vincent Heuveline, Nina Bisheh (Hg.): E-Science-Tage 2021.
Share Your Research Data. Heidelberg: heiBOOKS, 2022.
DOI: https://doi.org/10.11588/heibooks.979.c13716 (CC BY-SA 4.0)

- Some volunteer institutions running infrastructure services[2], where the design is such that clients can easily be written without dependencies on specific instances of central components.
- Authors of client software, libraries, and web pages making these resources available to astronomers. Of the clients programmes listed on the IVOA's web site[3], non-astronomers might want to look at TOPCAT and Aladin.

The VO Text Treasures[4] service lists worked-out use cases that may give closer insights into what this actually means.

## 2 Data Discovery in the Virtual Observatory

In the VO, data discovery very typically is a two-step process, in which a client first queries the Registry for services that might have relevant data. The Registry is the collection of the metadata records on the level of data collections and will be looked at in some detail in Sect. 3.7.

The result of the Registry query will in general be a set of access URLs of machine APIs together with an identifier of what standard the API implements. With this information, a client programme can then visit the APIs in turn, running discovery queries in the end yielding references to datasets matching the constraints.

To make this a bit more concrete, consider a researcher looking for images of Barnard's star in X-rays. In this scenario, the researcher will run a client programme that will ask the Registry: What "resources" (services or data collections) are available that:

- serve or contain images
- have data in the X-ray part of the spectrum
- have data around $\alpha = 269.45$, $\delta = 4.693$ (the current position of Barnard's star in ICRS coordinates)?

In a second step, the client visits each service found (provided it supports the advertised communication protocols) and will post one request to each for images that

- cover the position $\alpha = 269.45$, $\delta = 4.693$,
- intersect the spectral range $0.1 \cdots 120\,\mathrm{keV}$ of photon energy.

The on-the-wire serialisation of these constraints depends on the protocol and may even have to happen client-side; for instance, the image discovery protocol SIAP in its version 1 (which is still widely used) cannot express spectral constraints.

[2] This primarily includes components of the service registry, but also the the IVOA web page at `https://ivoa.net` and the associated document repository and collaboration wiki.

[3] `http://ivoa.net/astronomers/applications.html`

[4] `https://dc.g-vo.org/VOTT`

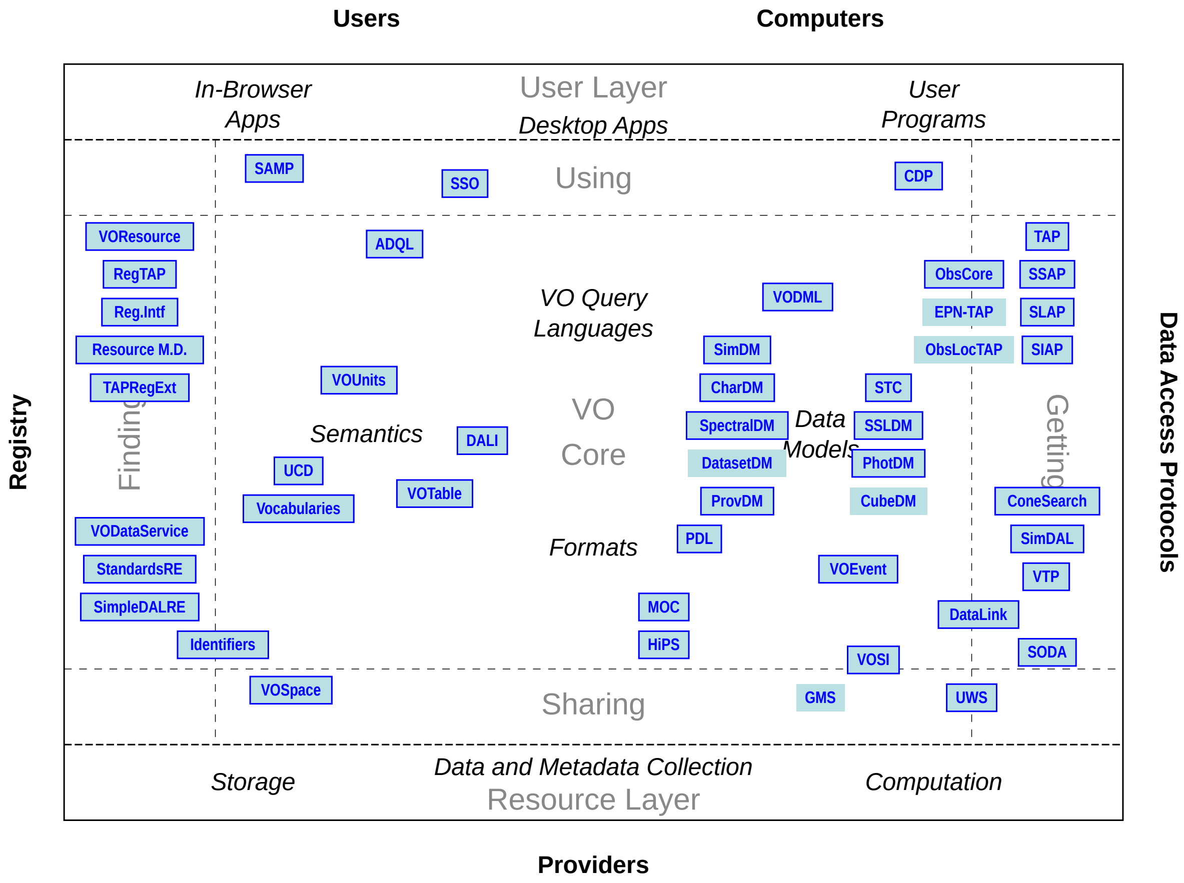

Figure 1: The Virtual Observatory's "Architecture Diagram", showing one box for each standard contributing to the ecosystem, grouped by their function. Some of the standards shown are discussed in the text; all of them are available through the IVOA's document repository.

The client programme then presents the resulting metadata records to the researcher in some suitable form – the Aladin client, for example, will show image footprints on a sky display in addition to a tabular rendering –, who in turn decides which data to retrieve (and in what way).

In practice, this procedure is a lot more natural than it might sound here, mostly because much of the interaction can run behind the scenes thanks to the standardised APIs in use in the VO.

## 3 Standards

Essentially all VO procedures depend on machines querying each other and understanding the responses without human intervention. This requires a significant effort in standardising protocols and formats.

Ever since the Architecture Note [1], in the Virtual Observatory the standards are visualised as shown in Fig. 1. Since the sheer number of standards involved may be somewhat daunting, let us have a look at some of the more salient ones and their functions. Words written like this correspond to standards readily found in the IVOA document repository[5] In order to keep the bibliography reasonably compact, we do not include full citations for them.

## 3.1 Finding Data and Datasets

The discovery protocols on the right side of the architecture diagram include “typed” protocols specialised on images (SIAP), spectra (SSAP), objects (SCS), or spectral lines (SLAP). All of these essentially define some query parameters and some basic requirements on the tabular structure of the response.

As the VO matured, tools were in place to define a table schema (ObsCore) that, together with the Table Access Protocol (TAP) to be discussed presently, enables flexible and powerful dataset discovery largely independently of their types. In a re-design of the VO, it is likely that one would think hard whether there actually is a need for the typed protocols – which, on the other hand, were a lot simpler to design than the generic discovery protocols and thus kept up the momentum while the years required to develop something like TAP (and its implementations) went by.

Another standard we should have started with but only introduced relatively late in the VO’s evolution is the Data Access Layer Interface DALI giving common patterns for VO standards concerned with finding and accessing datasets.

## 3.2 Advanced Data Access

Where datasets (such as images, spectra, or time series) are discovered, the services return metadata rather than the datasets themselves because very typically only a small fraction of the discovered datasets turn out to be necessary for a given analysis, and the datasets may be large.

For large and complex datasets, the typical access mode of simply dereferencing an (http) URI may not be suitable; in particular, clients may want to only retrieve parts of the dataset. To cater for such cases, the VO has defined Datalink, which allows data providers to declare relationships between various artefacts (e.g., raw data, calibration files, and reduced data) making up a dataset, and giving separate access to them.

On top of Datalink, SODA defines some standard operations on array-like data (e.g., cutouts), which in many science cases can reduce the amount of data to be transferred by several orders of magnitude.

[5] `https://ivoa.net/documents`.

## 3.3 Interacting with Databases

The definition of the Table Access Protocol (TAP) in 2010 enabled many interesting use cases; in particular, the built-in table metadata inspection and table upload facilities are central to the protocol's success.

Equally important was the definition of a common language available across all TAP services, the SQL-derived Astronomical Data Query Language ADQL. It is now no longer unusual to see ADQL fragments in scientific publications, and an increasing fraction of astronomers becomes proficient in expressing science questions in SQL-like languages.

## 3.4 Formats

Rich metadata is a precondition of re-usability of data as well as interoperability of services. Hence, the XML-based table and metadata format VOTable was the first standard defined in the VO and keeps being regularly evolved.

The VO had a head start because FITS [4], a standard for images and several other data types, was already widely accepted within the field and could simply be re-used. Still, some additional formats, for instance for complex spherical geometries (MOC) or for large, multi-resolution, possibly full-sphere images and tables (HIPS), had to be defined.

Related to these data formats are several syntactic aspects, such as agreeing on a well-defined way to serialise physical units into ASCII strings (VOUnit).

## 3.5 Desktop Interoperability

From the VO's start it was clear that no single software would ever be sufficient to serve the needs of the various communities using VO facilities. Instead, the design called for multiple, independently developable applications that, however, can closely interoperate among each other. Given that no widely adopted standard for cross-platform desktop programme communication existed in the mid-2000 (or, really, exists today), the Simple Application Messaging Protocol SAMP was developed and enjoys great popularity among VO users.

## 3.6 Semantics

Early on, a relatively complex labeling scheme called UCD was devised to enable machine-readable annotation of table columns with what sort of physics they represent (e.g., "radio flux" versus "distance").

Later, RDF-compliant, hierarchical vocabularies were defined for many different purposes, from the relationships between different artefacts in Datalink to time scales to content levels of resources (Vocabularies).

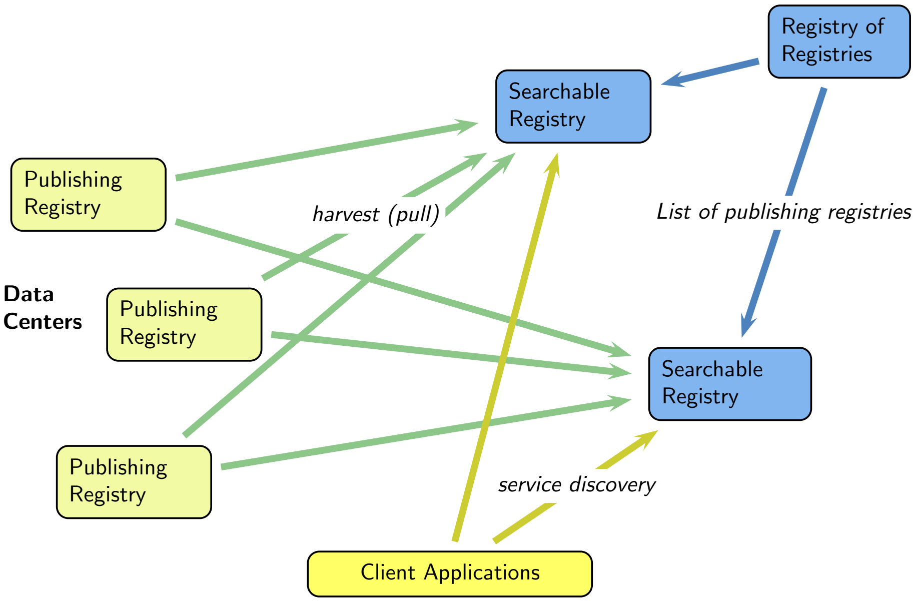

Figure 2: A sketch of the registry architecture of the VO: Data centers operate OAI-PMH endpoints ("publishing registries"), which are harvested by searchable registries, which in turn are what client applications talk to. The searchable registries learn what publishing registries to harvest from a central list kept at the registry of registries.

## 3.7 Registry

The part of the VO that is probably most readily re-usable in large parts in other disciplines is the Registry (cf. [2]), the distributed data collection metadata store for the VO's resources. Let us therefore take a closer look at its architecture.

The basic scheme, as defined in Registry Interfaces, is shown in Fig. 2. It again follows the fundamental VO principle that no single central component should be required for client applications. The searchable registries – all of them ideally serving identical data – can be operated by anyone and liberally scaled, thus providing substantial resilience against failures. In practice, three major searchable registries are operated in the VO, one each by NASA, ESA, and GAVO.

The components not easily reproducible, the Registry of Registries and the publishing registries, can in principle be down for days without VO users noticing; it is only registry updates – to the list of publishing registries and the records published by a single publisher, respectively – that will not happen during the downtime.

Standards contributing to this system are a lightweight identifier scheme, IVOA identifiers, a set of rules for the VO-specific application of OAI-PMH, Registry Interfaces, and the metadata schema based in VOResource and extended to various resource types in VODataService, SimpleDALRegExt, and TAPRegExt.

The now-dominant registry client interface RegTAP builds on TAP and essentially just defines a relational mapping of the metadata scheme.

This mapping yields moderately-sized tables ranging from $\sim 30000$ records for the table of resources to $\sim 1.2$ million records for the table of columns in published tables.

## 4 Concluding Remarks

Compared with just writing some custom web page, building a system like the Virtual Observatory may seem like a daunting task. On the other hand, many publishers rebuilding the same kinds of tools on their custom web pages over and over is substantially more work on the long run, and the benefits of being able to share the load of developing client software provides a large benefit to both data producers and data consumers.

An alternative to our current design – that predates the “platform economy” – would be a single, giant, central platform keeping essentially all astronomical data. This is a model that in astronomy works quite well for literature in the form of NASA’s Astrophysics Data System ADS[6]. For data, with its much greater variety, volume (at least when measured in bytes), and demands on machine-readability, it would probably be very difficult to work out a global funding scheme for such and establishment.

More importantly, however, with multiple interoperable data centers, the VO can “grow from the edges”, much like the internet (at least in its early days). This means that everyone is free to run services and to improve the tools and, within reason, also the standards – there is no single entity determining what can and cannot be done.

For users, this means that they are free to choose whatever tools they want to use, something a platform would severely limit. Having data uniformly presented and described also not only greatly facilitates working with cross-instrument (and hence multi-wavelength, possibly even multi-messenger) data, it also significantly increases the chances that workflows will be reproducible (again, within reason) years down the road: Our API endpoints have proved to be a lot more stable than web pages.

## Acknowledgements

The German part of the Virtual Observatory, GAVO, has been supported by BMBF under several grants. Current grant: e-inf-astro, FKZ 05A20VH5.

## Bibliography

[1] Christophe Arviset, Severin Gaudet, and IVOA Technical Coordination Group. IVOA Architecture Version 1.0. IVOA Note 23 November 2010, November 2010.

[6] https://ads.harvard.edu

[2] M. Demleitner, G. Greene, P. Le Sidaner, and R. L. Plante. The virtual observatory registry. *Astronomy and Computing*, 7:101–107, November 2014.

[3] R. J. Hanisch. The Virtual Observatory: I. *Astronomy and Computing*, 7:1–2, November 2014.

[4] D. C. Wells, E. W. Greisen, and R. H. Harten. FITS - a Flexible Image Transport System. *Astronomy and Astrophysics Supplement*, 44:363, June 1981.