Fig. 1


Fig. 2

1. VO – what's in it for you?

(cf. Fig. 1)

Markus Demleitner
msdemlei@ari.uni-heidelberg.de

(cf. Fig. 2)

- A definition
- The data paradise: big and diverse
- Challenge: Finding data
- Challenge: Using data
- Challenge: Dealing with size
- Challenge: Contributing

For links grab http://docs.g-vo.org/vowf.pdf

Well, in times of The Crisis, you know, challenges are opportunities, so let's start again:

2. A Definition

The Virtual Observatory (VO) is (or will be), a
comprehensive set of
data and **services**
relevant to **astronomy**
accessible from **clients** of **your choice**
regardless of where you are and
preserving products of digital astronomy.


Fig. 3

3. Data Paradise I: Diversity

There's tens of thousands of data collections somewhere online, and more should be.

To unlock the treasures hidden there, you have to be able to

- Find the data
- Get it
- Use it

VO jargon: A dataset is understood to be an "individual data item with included metadata", which could be a table, a spectrum, an image, a data cube, or yet something else. Since a set of such things needs a name, too, and dataset is not available, we call that a data collection.

Also, maybe the word "metadata" deserves a brief comment: Metadata is "data on data". For an image, that could be "When was it taken?", "What filter was used?", "Where does it point?", "What does it show?", etc.

(cf. Fig. 3)

Remember Astronomy and Astrophysics Abstracts? The VO is to old data management what the ADS is to AAA.

4. The VO answer

In theory, all those data collections could reside in one, professionally managed place.

This would be like ADS; the publishers deliver their data, and the ADS staff unifies and "curates" this.

In reality, such a place doesn't exist. Although for tabular data, VizieR comes pretty close.

The VO way: Let there be many data centers, but have them speak common languages ("protocols") and make it so their metadata can be collected and interpreted by machines.

This is a bit like the Web, where there's lots and lots of web servers, but google's robots can harvest what's on them and provide an index (only there's more webservers and far less structure in the Web).

5. Finding Services

The union of the metadata of all the data centers in the VO is called the **registry**. There, users can issue queries like:

- Where are image services specialized on radio?
- What data sets are out there containing x-ray fluxes and proper motions?
- What services are out there dealing with time standards?
- What services expose the data associated to a paper?

Clients: web interfaces[1], VO Desktop[2].

You can also query the registry using the TAP/ADQL clients mentioned below using the TAP access URL http://dc.g-vo.org/tap. If I (as the author of the respective standards) may say so, this is probably the way to go if you're planning advanced stuff with the registry.

[1] http://nvo.stsci.edu/vor10
[2] http://www.astrogrid.org/wiki/Install/Downloads

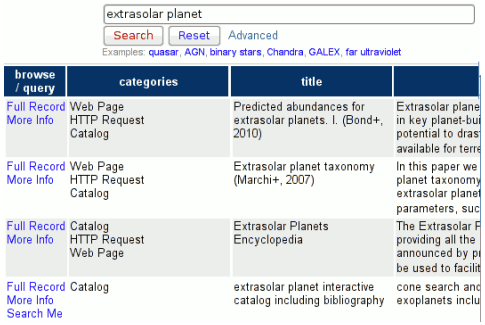

Fig. 4

6. A Web-based Registry Search

(cf. Fig. 4)

This is an edited screenshot of the VAO registry-based discovery tool (http://nvo.stsci.edu/vor10).

7. Finding Data Sets

The VO has defined "**typed interfaces**" that let you talk to all services in the same fashion. "Typed" means literally types of data. There is, for example, "Simple Cone Search" (SCS) for tables with sky positions in them, the "Simple Image Access Protocol" (SIAP) dealing with images of the sky, and "Simple Spectral Access Protocol" (SSAP) for accessing spectra.

The common language lets programs query many servers at one click. So, you can ask questions like:

- Find all images containing NGC3141
- Are there infrared spectra of a source at 271.8281, +23.42?
- What is known about sources within 2 arcminutes of Geminga?

Clients: TOPCAT[3] for tables, Aladin[4] for images, Splat[5] for spectra, and more.

Upcoming, there's ObsTAP that lets you post even more expressive queries against database tables.

These protocols also usually say how you can get the data once you have located it. There is work in progress on server-side manipulations, though (cutouts, cube cuts, etc) – but standards for that are hard.

The interfaces are called "typed" since they have been written for a specific *type* of data product:

- SCS for tabular data with positions ("object lists"),
- SIAP for images, preferably with astrometrical calibration
- SSAP for spectra (and potentially time series)
- SLAP for lists of spectral lines

[3] http://www.star.bris.ac.uk/~mbt/topcat/
[4] http://aladin.u-strasbg.fr/aladin.gml
[5] http://star-www.dur.ac.uk/~pdraper/splat/splat-vo/

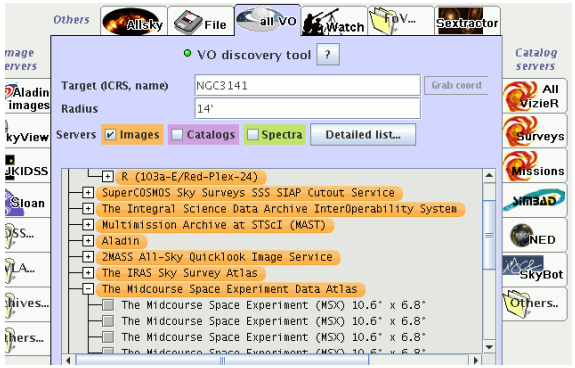

Fig. 5

8. An All-VO Image Search

(cf. Fig. 5)

This is Aladin's load dialog. You could also use, for example TOPCAT's SIAP dialog to do the same thing, or write your own simple client in a couple of lines of python (and a library, of course).

9. Using Data

The VO uses existing data formats where they are appropriate (e.g., FITS for images). Where they aren't it uses its own: **VOTable**, containing rich metadata. This saves you from having to write code everytime you want to use a new data source.

And it comes with descriptions, units, UCDs, and more. UCDs are a VO thing, too: spelled out, it's unified content descriptors. They are short strings that say what kind of physics a column represents: pos.eq.ra is a right ascension, phot.mag;em.opt.V is a visual magnitude, etc.

The VO also defines data models (e.g., for spectra) that say what metadata items are necessary for a useful description.

Clients: TOPCAT[6] and STILTS[7] for generic VOTables; the clients for typed interfaces also consume VOTables.

[6] http://www.star.bris.ac.uk/~mbt/topcat/
[7] http://www.star.bristol.ac.uk/~mbt/stilts/

Fig. 6

10. Data Paradise II: A Sea of Bytes

We now have 10^9 sources in catalogs, the time domain is coming. Ftping ASCII tables of those is tedious, and mostly you only want some 10^3 of those objects in the first place. So:

- Keep the data on the servers as long as possible
- have servers exchange data on the user's behalf
- only download data when most of the stuff irrelevant to your research is weeded out.

(cf. Fig. 6)

That's the Hippcarcos catalog. With about 10^5 objects, it's positively small by today's standards.

11. Trawling the Sea of Bytes

ADQL lets you write simple programs, **TAP** lets you run them on remote servers, upload your tables, and retrieve the results.

If you know CASJobs: about the same thing, only with a solid standard and supported by more services.

Clients: TOPCAT[8] and STILTS[9], tapsh[10], seleste[11], TAPHandle[12]

Learn it: ADQL course[13]

[8] http://www.star.bris.ac.uk/~mbt/topcat/
[9] http://www.star.bristol.ac.uk/~mbt/stilts/
[10] http://vo.ari.uni-heidelberg.de/soft/tapsh
[11] http://neo.cfa.harvard.edu/seleste/
[12] http://saada.unistra.fr/taphandle
[13] http://docs.g-vo.org/adql

12. User Paradise: Freedom of Choice

The VO is about standards. Any client implementing a standard can query any server implementing a standard. This gives users a choice of software, and using libraries or frameworks, they can simply write their own clients.

Plus, most VO software interoperates – you can send tables, selections, etc. from one program to the next using a protocol called **SAMP**. Try it, it's fun.

Clients: Almost all of them. You won't even notice.

13. See You in Paradise

This is real – but it's still missing much data. Such as, in all likelihood, yours. Let us publish it in the VO (http://g-vo.org)

If you have ideas or wishes: Let us know.

If you have problems: Complain, to us or the operators of the service you're using.

Stay up to date:

- http://dc.g-vo.org/regrss
- http://twitter.com/germanVO