Fig. 1

Fig. 2

1. The VO And Why It Matters

(vgl. Fig. 1)

Markus Demleitner
msdemlei@ari.uni-heidelberg.de

(vgl. Fig. 2)

- What is the VO?
- The VO's answers
- What can you do?
- Publish data or let it perish

2. What's the VO?

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.

3. "comprehensive"

The VO intends to allow access to basically all astronomical data, present and past.

Right now: About 15000 resources like

- VizieR catalogs
- Lots of space missons
- Many observatory collections
- Theory data like synthetic spectra
- Much more

But **much** is still missing (e.g., much of ESO's data).


Fig. 3

4. Data and Services

While the VO is about data, much of it is concerned with services.

A service is

- a piece of software accessible via a network
- with a well-defined interface
- allowing access to some data collection.

Important: Service users ("clients") must be able to figure out how to operate the service and find out as much as possible about the data contained.

5. Astronomy

Well, of course...

(vgl. Fig. 3)

...but we also pave the way for similar endeavours in other fields; cf. the current ASTROTROP project for VO tech in tropical rainforest research.

6. Clients and Choice

"Web pages" aren't really what the VO is about. It is about standard interfaces to data.

This means: A single program (possibly web-based) can operate all kinds of archives and services. Many such programs are listed at http://ivoa.net.

It also means: A given service can be operated by any client speaking the VO languages – you get to choose or use libraries like pyVO in your own programs.

7. The big equalizer

It used to be that you had to go to the big observatories to get top-notch data.

Converseley, chances someone would see and use your data if you weren't there weren't terribly good.

The VO already delivers excellent data to anywhere and anyone in the world.

And with All-VO searches and increased adoption of Registry use, everyone gets a more uniform view of the data taken.

8. Preservation

We're currently losing historical observations at an unprecedented rate: All the tapes from the 80ies and 90ies are deteriorating.

Linus Torvalds:

> Only wimps use tape backup: real men just upload their important stuff on ftp, and let the rest of the world mirror it.

If data is to survive, it must be in living services not far from spinning disks.

(Yes, there's more to it, but the living part is vital)

9. Challenge: 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.

10. The VO way

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).

11. 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: WIRR[1], VO Desktop[2], In-Application interfaces.

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://dc.g-vo.org`
[2] `http://www.astrogrid.org/wiki/Install/Downloads`

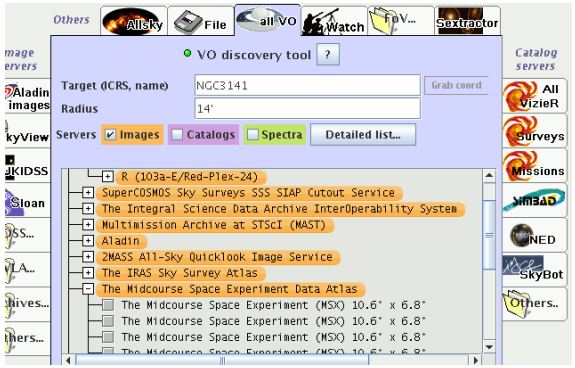

Fig. 4

12. 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.

13. An All-VO Image Search

(vgl. Fig. 4)

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).

[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/

14. 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. See also Astropy and many other libraries

15. Challenge: Size

Some modern data collections are too large to move – smarts must come to the data.

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]

16. Challenge: 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.

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

17. Demo time

Put together

- Registry
- SCS
- TAP
- SAMP

for a nice visual encounter with sources with infrared excess around galactic OH masers not too far from the Galactic center.

Here's what to do:

Getting the OH masers

Start TOPCAT, select VO/Cone search

Put in `OH masers` as keywords. You should get back about 20 resources, if it's less, use a different registy (this is using the Registry standards).

In the lower part of the dialog, enter `Sgr A` as object, hit Resolve, enter 30 as Radius.

In the service list, look for the service with short name `engels_ohmasers` and double click it (this is using the Cone Search standard).

Plot the resulting table on a sphere to make sure it looks plausible.

Getting objects close to the masers

We're going to use the supercosmos data (that's basically objects within DSS). You could locate them using the registry, but we're taking a short cut.

Still in TOPCAT, do VO/TAP Query. In TAP URL, enter `http://dc.g-vo.org/tap`, hit "Enter Query".

In the top part of the dialog, select the table `supercosmos.sources`. Then, from "Examples" below, select "Upload Join" and edit the resulting query to look like this:

```
SELECT
   TOP 100000
   db.*
   FROM supercosmos.sources AS db
   JOIN TAP_UPLOAD.t11 AS tc
   ON 1=CONTAINS(POINT('ICRS', db.raj2000, db.dej2000),
                 CIRCLE('ICRS', tc.raj2000, tc.dej2000, 20./3600.))
```

(i.e., raise the match limit in TOP, change the select list to `db.*`, and raise the match radius to 20 arcsec).

You may need to adapt TAP_UPLOAD.t11; the 11 must be the little number in front of the table with the masers in TOPCAT's table list (if the table you want to upload was selected when you generated the query, TOPCAT has already done the right thing).

Execute the query. This might take a few 10s of seconds as it is inspecting the vicinities of 4000 objects in a catalog with roughly 10^9 rows). The result is about 60000 objects.

Add infrared magnitudes

Again, we're taking a shortcut by using 2MASS from a known location. Try using WISE by asking the registry.

In TOPCAT, again to VO/TAP Query. Select `twomass.data` fromt the table list, pick "Upload join" from the examples. This time, change the match size to 1/3600. (both catalogs have good astrometry, and we don't expect our objects to move fast) and the match limit to 100000. Send away the query. Here, there's quite a bit of data transfer involved, so be prepared to wait for two minutes or so, longer if you're on a slow network.

Inspect the result

Plot gcormagb agains kmag, flip kmag.

Now have a look at the weird objects in the upper part of the display: Start Aladin, in TOPCAT, do Interop/Send Table to/Aladin. Also set the activation action to: "transmit coordinates".

In Aladin, activate "Optical" and zoom in. Then click on suspicious points and gawk at them in Aladin.

For the curious: Can you get Spectra for these?

18. Your Contribution

Do you have data that others could re-use? No? You're sure?

19. Common Excuses

Shamelessly stolen from http://datapub.cdlib.org/closed-data-excuses-excuses/[14]

- People will contact me to ask about stuff – well, science is about exchange, and you'll usually notice that most of those questions are actually quite clever, so answering them is a good use of your time.
- People will misinterpret the data – good documentation and standards mitigate this. The rest is just as with publishing prose, isn't it?
- My data is not very interesting – leave that decision to others. You'd be surprised how much „boring data" people click-and-type from printed graphs and tables each week.
- I might want to use it in a research paper – well, if you've not done so so far, will you? When? Too much data is gathering dust, waiting for the „real soon now". Be fair to the world and publish, if need be with an embargo.
- I'm not sure I own the data – that sucks. The original source has some advice for you.
- My data is too complicated – if it's too complicated to explain: are you sure you've understood it yourself? Try explaining anyway, you won't regret it.
- My data is embarrassingly bad – everyone's is. Good data is just bad data that more eyes have seen and more hands have improved.
- It's not a priority and I'm busy – ah-ha! Here we're coming to a real kicker. Rewarding data publishing is something we're working on (e.g., the Thomson Reuters has started a data citation index). Then again, publishing doesn't need to be so terribly painful...

[14] `http://datapub.cdlib.org/closed-data-excuses-excuses/`

20. Data Publishing

There is nothing like Journals for publishing data yet (though VizieR comes close for tables).

See: http://ivoa.net, "Publishing in the VO" – either:

- Ask a data center (VizieR, us, ...) to do it for you, or
- Use a publishing toolkit on your own machine, or
- Write your own software using libraries

21. Conclusion

The Virtual Observatory is there for you.

Use it.

Contribute.

Thanks!