

Confessions of a Registry Janitor

GEFÖRDERT VOM



Bundesministerium
für Bildung
und Forschung

Markus Demleitner

msdemlei@ari.uni-heidelberg.de

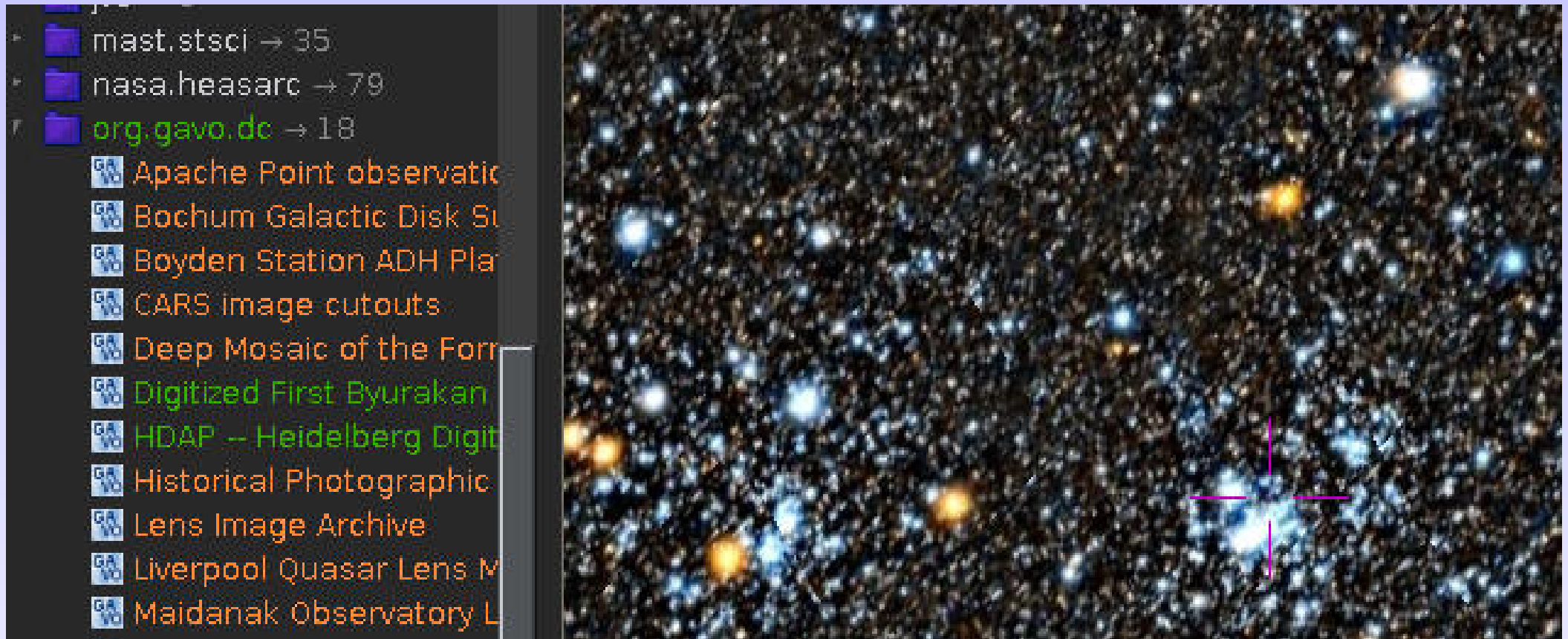


- Coverage!
- Free-form person names
- Personal addresses as contacts
- Non-UAT subjects
- Short descriptions
- Missing Tablesets
- Column descriptions too short or too long
- Bad column UCDs



Coverage!

Tell the world where in space, time, and spectrum your data is.
That's the number one thing needed for blind discovery. It also lets aladin colour matching resources:



The image shows a screenshot of a file browser interface on the left and a corresponding astronomical image on the right. The file browser lists several directories and files, each with a small icon and a count of items. The directories are:

- mast.stsci → 35
- nasa.heasarc → 79
- org.gavo.dc → 18

Below these are several files, each with a small icon and a label:

- GA VO Apache Point observatic
- GA VO Bochum Galactic Disk Su
- GA VO Boyden Station ADH Pla
- GA VO CARS image cutouts
- GA VO Deep Mosaic of the Forn
- GA VO Digitized First Byurakan
- GA VO HDAP -- Heidelberg Digit
- GA VO Historical Photographic
- GA VO Lens Image Archive
- GA VO Liverpool Quasar Lens M
- GA VO Maidanak Observatory L

The astronomical image on the right shows a dense field of stars, with a prominent bright star in the lower right quadrant. A red crosshair is overlaid on this star, indicating its position. The stars are of various colors, including blue, yellow, and white.

Broken Author Names

If I'm looking for data created by Gerard de Vaucouleurs, it would be great if I could just write `role_name` like `'de Vaucouleurs, G%'` and not have to worry about any of:

- `van der Waarden, B; de Vaucouleurs, G.`
- `Gerard de Vaucouleurs`
- `G. de Vaucouleurs`
- `Vaucouleurs, G. de`

And don't get me started on `"et al. MNRAS (in press)"`, `"⟨NOT GIVEN⟩"`, or `"under a cooperative agreement with the NSF on behalf of. . ."` as creator names.

Fragile Contact Info

Curation/contact is an important item, both for janitors and for our users (for bug reports and/or help requests) – have you noticed the `get_contact` method on pyVO registry results?

Help make it more robust by avoiding personal addresses. Instead, use addresses you can hand on when you move on.

Try talking your IT department into giving you `vo-support@your.institute.edu`.

Free-form Subjects

Since last year, it's finally clear that subjects should *really* be the identifiers from <http://www.ivoa.net/rdf/uat> (“abell-clusters”).

So, please no:

- “DR1, SIAP, Images”
- “???”
- “optical lines equivalent width catalog”
- “Galaxy:general”

There's much cool stuff you can do with machine-readable subjects – don't miss out!

Short Resource Descriptions

A resource description should be enough to give a non-specialist astronomer an idea if the data can be useful for what they're trying to do. And it should work for full-text searches.

So, please no:

- “Survey (LoLSS).”
- “USET Group form”
- “Convolution of normalized synthetic stellar spectra”

Guideline: If you briefly address each of What, From where, What for, that's probably already a useful description.

Relationships

If a table underlying your data is queriable in a TAP service, don't forget to declare a ServedBy relationship there.

And please don't use related-to any more. It's deprecated for a reason.

Check out our relationship_type vocabulary (and have it amended if it's missing what you need).

Missing Tablesets

If your registry record has this:

```
</capability>  
</ri:Resource>
```

you're probably doing it wrong: Try to define a tableset, i.e., the table(s) you return or that can be queried.

For TAP services, I need it so I can one day shut down GloTS. In general, it's great for discovery ("Where can I get position angles of ellipsoids fitted to galaxies?").

Bad Column Descriptions

As for resource descriptions: Non-specialist astronomers should have a chance to figure out what it is. So, please no:

- ?, ??, or ???
- “RA”
- “This is the excess [about 2k chars follow] Algorithms, and Software Implementation.”
- NULL

On descriptions of excessive length: Let’s work on column footnotes in VOTables!

Bad Column UCDs

It'd be great if I could reliably discover all position angle columns in the VO by constraining `ucd='pos.posAng'`.

But: there are 117 columns in the VO called `pa` without a UCD.

And: 30342 times `??`, 41 times `???`, and 70 times `????`.

Oh, and don't use pre-1+ UCDs any more (`"POS_EQ_RA"`).

Trivia: the longest UCD in the VO has 74 characters.

Thank You For...

...improving the VO Registry!

The long version of this talk is at

<http://blog.g-vo.org/registry-a-janitor-speaks-out.html>