Improving IVOA registration for new data service providers

Agenda and notes for Zoom meeting

Oct 3 2023 at 4pm CEST (2pm UTC, 10 am ET, 7am Tucson)
Zoom link: https://cnrs.zoom.us/j/96910109418?pwd=VUcrNSttTzJobHVpZXlVZnA3SGUxdz09

Participants

Renaud Savalle - renaud.savalle@obspm.fr
Gilles Landais - gilles.landais@unistra.fr
Destry Saul (tentative, may be listening-only) - destry.saul@noirlab.edu
Hendri zu,, B8 c,c,k Heinl - heinl@g-vo.org
Marco Molinaro - marco.molinaro@inaf.it
Igor Suarez-Sola - igor.suarez-sola@noirlab.edu
Mark Taylor - m.b.taylor@bristol.ac.uk
Markus Demleitner - msdemlei@ari.uni-heidelberg.de
Christophe Arviset - Christophe.Arviset@esa.int
Patrick Dowler - pdowler.cadc@gmail.com
Tamara Civera - tcivera@cefca.es
Robert Nikutta - robert.nikutta@noirlab.edu
Philip Matsson - philip@winterway.eu
Tessa Dower - dower at stsci edu
Sebastien Derriere - sebastien.derriere@astro.unistra.fr
Tom Donaldson - tdonaldson@stsci.edu
Tess Jaffe - tess.jaffe@nasa.gov

Agenda

1. RS: Introduction - The Registry Ecosystem - 5 mins
2. HH: Edu IG group involvement - 5 mins
    a. Workshop addressing new data providers (following the VESPA examples)
    b. Workshop addressing established data providers on improving/correcting their registry records
    c. Establishing an entry point for data publishing in the VO (in the frame of the new webpage)
3. Data Providers - Data/Services Registration experience: Feedback/Stories/Tools
    a. New data providers (ex: NOIRLAB…)
    b. Established data providers (ex: CDS, VizieR, CEFCA…)
4. All - Brainstorming - How to improve the process ?

Ideas/Questions

- Standard registry implementations, or links to open source registries (Destry)
- registry validators (links on https://wiki.ivoa.net/twiki/bin/view/IVOA/IvoaValidatorsSummary are broken) (Destry)
- As far as making it ‘easier’ to implement a publishing registry: removing the `from=` and `until=` args would make it easier to host a simple publishing registry in a static site (e.g. github/gitlab pages) (but using query args even for simple navigation isn’t supported by gitlab pages anyways) (Destry)
- Do VO services all return VOResource records that the registry needs to publish? If not, that would simplify creating the VOResource records. (Destry)

Collaborative Notes

RS: Introduction to registry eco system

- Registry
- Publishing in the VO page is a bit outdated and could need some love

Experiences from data providers

RN: Looking into the how to register page is quite opaque and it’s not easy to make sense. The process seems to be complicated to us. Maybe this is due to a grown system, but maybe it could be improved.
IS-S: The documentation is great, but a lot. It’s missing simple examples to start from. Two questions

1. Who do we tell that our XML works ? Where to go to publish ?
2. Are there available tests out there to validate the service ?

RN (in chat): Of course we’d like to stand up a publishing registry, because down the line we’d be adding other datasets and services.

CA: For validation: https://registry.euro-vo.org/evor/#info
IS-S: It would be great to have a local test, e.g. some python code.
MD: The validator does a lot that’s actually hard to do locally. The code is available on github though. But the validator is safe in the regard of “not publishing yet”.
As for local validation: do the schema validation, if these are OK, the RofR validator will check for issues linked to remote connection.
IS-S: It’s helpful to have these test available locally.

MD : For local validation, I think it’s enough to simply run an XSD validation; when you break things, they typically make things XSD-invalid, too. It’s rather tricky to do the XSD validation without pulling all the schema files from ivoa.net all the time.

If you have a Debian-derived Linux, the most convenient way may be to
apt install gavodachs2-server
And run
dachs adm xsd <yourfile>
(or check how DaCHS builds a joint validator from local sources; that would be in gavo.helpers.testtstricks, starting at XSDResolver going all the way through getXSDErrorsLXML. For casual use, there's now also stilts xsdvalidate, but that will pull files from ivoa.net.

PD: We have code for validation to check changes to services over time. If it's useful for other people we'd make this available. Will post it over the registry mailing list.

TC: The documentation indeed is sometimes difficult to follow. It's difficult to know where to start. In case of CEFCA I used other registry records as examples, which was very helpful to follow. Especially hard to figure out which datarecords are needed for different VO services and data releases.
When validating the VO resources, I found errors thanks to the validators, so I could correct this in our publishing registry. Also sometimes it's hard to understand error messages and link it to the actual error on our side. Eventually it would be useful to have a standalone tool instead of a web page, because for the latter you have your publishing registry already running.

CA (in chat): the Euro-VO Registry now also offers (as requested by Markus), a rest API to the validators https://registry.euro-vo.org/evor/#infopublishersrest or a GUI version at https://registry.euro-vo.org/evor/#newstandalonevalidation

MD: For standalone validation I run a XML schema validator. Dachs comes with an XML validator.
dachs adm xsd <resource>
Dachs is available as debian package.
apt install dachs

DS: I agree with Tamara. We are at the same spot in registering our VO resources and that is hard to figure out.

TC: At CEFCA we make a VO resource for every data release, so having a standalone validator would be very feasible.

SD: At CDS we are running a publishing registry, mostly our resources are catalogs coming from Vizier, today this are more than 20,000 VO resources. We are running this registry since some 20 years.
Currently our process consist of three parts:

1. The factory, which reads metadata from ASCI files, some come from XML files
2. Local validation of the XML schema and check if some resources have disappeared
3. OAI-PMH Images visible from the outside, which are accessible for harvesting registries.

GL: For CDS the publishing registry is not only for publishing in the VO, but also for other registry harvesters, these days it's important for B2FIND/EOSC.
That is linked to a problem with B2FIND and DOIs.

— after verification: DOI is in Eudat and in EOSC (sorry , my info was not up to date !)

Also I agree with Tamara that it's not always easy to find your way around.

SD: For validation we do not only rely on our local validation, but also get input from partners harvesting our metadata and give feedback about problems.

SD (in Chat): An interesting library if you want to provide an OAI-PMH interface to your collection of XML resource:
https://pypi.org/project/pyoai

RN: How picky are the validators ?

CA: They are very picky.

MD (in chat): In case you'd like a bit of philosophy on how picky validators ought to be:
https://blog.g-vo.org/requirements-and-validators.html

TD: At MAST we do XML validation when uploading new VO resources. Our validator is differently picky, because of the underlying code. Would it be useful to raise awareness when our validator complains during the harvesting of others registries ?

RS: The ROFR validator: http://rofr.ivoa.net/regvalidate/

RS: To conclude for today: It seems that implementing OAI-PMH is the crucial step, don't invent the wheel here, there are existing packages for this available in different languages.

IS-S (in chat):
https://ws.cadc-ccda.hia-iha.nrc-cnrc.gc.ca/reg/

DS (in chat): Ours [validator] is/will be Open source
https://gitlab.com/nfs-noirlab/csdc/vo-registry

MD (in chat): Sample VO-Resource records are few pages down on:
https://dc.ah.uni-heidelberg.de/purx/enroll/info

Validators

Registry of Registries Validator
http://rofr.ivoa.net/regvalidate/

Euro VO
https://registry.euro-vo.org/evor/#info
A REST API: https://registry.euro-vo.org/evor/#infopublishersrest
as GUI version https://registry.euro-vo.org/evor/#newstandalonevalidation

CADC:
https://ws.cadc-ccda.hia-iha.nrc-cnrc.gc.ca/reg/

For local testing:

Dachs
On Debian derived Linuxes:
apt install gavodachs2-server
Then run:
dachs adm xsd <yourfile>

Examples for VO Resource records

https://dc.ah.uni-heidelberg.de/purx/enroll/info

Links

1. https://wiki.ivoa.net/twiki/bin/view/IVOA/GettingIntoTheRegistry
2. https://wiki.ivoa.net/twiki/bin/view/IVOA/PublishingInTheVO
3. https://www.ivoa.net/documents/IVOAArchitecture/20211101/index.html - IVOA Architecture Version 2.0
4. https://ui.adsabs.harvard.edu/abs/2014A%26C.....7..101D/abstract - Review Paper “The virtual observatory registry” by Demleitner et al.