Upgrading the CDS/VizieR publishing registry

S. Derriere, G. Landais, A. Guyot

CENTRE DE DONNÉES
ASTRONOMIQUES DE STRASBOURG

CDS/VizieR publising registry

Current version:

- Created in 2007 - ivo://CDS.VizieR/registry

- OAI endpoint
  http://cdsweb.u-strasbg.fr/reg-bin/vizier/oai.pl

- Contains 20,305 Resources (+1200 new/year)

  – Represents a significant fraction of all VO Resources

  – Mostly VizieR catalogues described as vs:CatalogService, with ConeSearch capability if a table contains coordinates

CDS/VizieR publising registry

- Programmed in PERL
  - Uses the VTOAI library (VirginiaTech OAI-PMH2 PERL Implementation)
    http://dlrl.cc.vt.edu/projects/OAI/software/vtoai/vtoai.html

- A few hard-coded ressources
  - Registry, Authority, TAP, SIAP, SSAP...
  - Stored as PERL associative arrays

- <ri:Resource xsi:type="vs:CatalogService"> dynamically generated from VizieR metadata when harvested
  - RDBMS : Sybase, then PostgreSQL
  - Additional information from ReadMe ascii files

Problems

- Designed with ~6,500 Resources, now 20,000+
  - Not so problematic with PMH ResumptionToken

- Dynamic generation of Resources
  - Failures in DB access or reading ReadMe files → incomplete records

- System has no memory – no deleted ressources
  - Changes in the DB could cause a Resource to disappear

- Hard to maintain/evolve

☐ A brand new version

Make a new publishing registry

In Python.

Separate :

- Resource factory : XML generation

- Content validation

- OAI-PMH harvesting

Resource Factory

- Generate VOResource .xml files for all Resources at a given date
  - Possible to keep the last N versions
- IVOA .xsd schemas to Python classes with PyXB
  - https://pypi.org/project/PyXB/
- Still a few static local ressources, stored in XML : Registry, Authority, SIAP…
- Metadata for VizieR catalogues are collected from :
  - ElasticSearch queries
  - API on ReadMe files returning JSON
  - Direct access to PostgreSQL VizieR metadata

Validator

- Identify Factory errors while retrieving metadata

- Check that Resources have been successfully generated, and they are valid

- Check for Resources that might have disappeared (not implemented yet)

Harvesting

- OAI endpoint : simple script, parsing of parameters (verb, metadataPrefix, from, …)

- Handling of OAI-PMH queries (verbs, …)

- Uses pyoai library
https://pypi.org/project/pyoai/
+ adaptation to VOResource (namespaces, …)

- Wrap Resources in OAI envelope

Improvements, upgrades

- Better management of UCDs

- New <relationship>s

```
-<relationship>
    <relationshipType>IsServedBy</relationshipType>
    <relatedResource ivo-id="ivo://CDS.VizieR/TAP">TAP VizieR generic service</relatedResource>
 </relationship>
-<relationship>
    <relationshipType>IsPreviousVersionOf</relationshipType>
    <relatedResource ivo-id="ivo://CDS.VizieR/VII/234">VII/234</relatedResource>
 </relationship>
```

- Description of mirrors

```
-<capability standardID="ivo://ivoa.net/std/ConeSearch" xsi:type="cs:ConeSearch">
  -<description>
     Cone search capability for table J/A+AS/135/133/redcat (DENIS I-band extragalactic catalog)
  </description>
  -<interface role="std" xsi:type="vs:ParamHTTP">
    -<accessURL use="base">
       http://vizier.unistra.fr/viz-bin/conesearch/J/A+AS/135/133/redcat?
    </accessURL>
    -<mirrorURL title="VizieR at IUCAA: Pune, India">
       https://vizier.iucaa.in/viz-bin/conesearch/J/A+AS/135/133/redcat?
    </mirrorURL>
    -<mirrorURL title="VizieR at SAAO: SAAO, South Africa">
       http://vizieridia.saao.ac.za/viz-bin/conesearch/J/A+AS/135/133/redcat?
    </mirrorURL>
```

- DOIs <dc:identifier>doi:10.26093/cds/vizier.33100008</dc:identifier>

- Better handling of DublinCore output

Conclusions & Roadmap

- New publishing registry available for testing
  https://cds.unistra.fr/registry/beta/

- Final tests and validation

- Crontab for Factory and Validation reports

- Announcement to the Registry mailing-list

- Use the new OAI service endpoint

  – Update ivo://CDS.VizieR/registry

  – Redirect queries from previous to new endpoint

  – Update RofR

- Full registries will have to re-harvest most records