InterOpMay2024 Registry Session

Registry Session: Tuesday May 21 2024 @ 16:00-17:30 (Session #7) Room C122

Program: https://wiki.ivoa.net/twiki/bin/view/IVOA/202404InteropRegistry

Audience: ?? on side + ~15 online

Tess Jaffe - Introduction

Registry Session Orientation: what are we talking about ?
- View from afar diagram: Publisher publish metadata describing services, Registry harvest metadata, Client query RegTAP service to discover data from published metadata
- RegTAP: relational database and endpoint using DAL, TAP/ADQL, VOTable
- Today's program and what does it concern in the above diagram

Q&A
- Comment by Christophe A: publisher can also use EURO-VO Registry to avoid generating their metadata.

Pat Dowler - Update on a re-usable OAI publishing registry

OAI-PMH registry service is used to publish resources by operating an OAI publishing registry to maintain your resource records
OpenCADC registry service: reg a simple web service driven entirely from files
Code on github, docker image avail (https://github.com/opencadc/reg)
TODO list:
- more documentation and examples/template XML files to start from
- tools to create/update resources ? editors, deployable web UI
- open to suggestions but need a plan to help people
- dev resources limited at CADC

Q&A:
- meaning of "claim an authority"? => A first VOR in the the RofR
- managing several authorities useful + =>

Renaud Savalle - Update on linking capabilities with tablesets

Gilles: Changing the granularity creates curation issue in Vizier which is more important than just factorization. VizieR have to provide whole catalog to authors. Currently some metadata is assigned to whole cat not tables (doi, keywords).

Collections in VizieR remain important. Relationships are interesting but require clients to adapt. Modification of the workflow would be necesary for VizieR.

Not a CDS issue! it is also the same for UKIDSS, ESA, etc.
In any cases, Collections are a good option to gather datasets with common topic (for instance dataset coming from the same mission/release (ESA), from the same article (VizieR)).
When coming from article, It is needed for authors to provide a way to list the whole collection with the metadata like DOI (important for citation).
Modifying the granularity in CDS, would be also inconsistent with current workflows (eg; astroquery, VizieR web pages, EOSC, ...)
In the tabset granularity , tablesets can be linked to catalogues using relationships. But these relations are not exploited today by clients, and ther adpations would need (important?) updates.

Markus - productTypeServed in VODataService

Use case: "Give me all svces publishing astro images" used to mean "all services with SIAP services ?"
No longer true: with now ObsTAP SIAP1,2, possibly TAP
Also: ObsTAP can publish anything and SSAP has been used for publishing TimeSeries
Proposed solution: add a productTypeServed element to VODataService's DataResource type
Several values possible, from a vocabulary
User interface: Since getting data providers to adopt will take long, RegTAP 1.4 (or maybe 1.3?) should automatically label images to SIAP1 and 2, spectrum to SSAP UNLESS they define productTypeServed
NB: Obscore Resources don't exist yet (cf TableReg DRAF note about the registration of tables
For RegTAP we could have a separate table, but we can use rr. res_detail (RegTAP extension mechanism for simple metadata)so a simple query to it can find resources publishing a certain sort of data. The UDF gavo_vocmatch avail at reg.g-vo.org could even reply to "give anything which is spatially resolved".
Implem status: DaCHS 2.9.4 can already produce productTypeServed, the rest (GAVO registry and pyVO) should not be hard.
Q&A
- Should Datalink (ObsCore) 's dataproduct_type and access_format be in the Registry ? => dataproduct_type applies to a row... havent' thought about it yet
- Precedent for filling in (possibly incorrect) metadata on behalf of providers ? => We do it for translating deprecated terms in relationships. It's justified by the time it would take for providers to update their resources. This is only encoding current practices (SIAP=>image)

Tess Jaffe - Registry spring cleaning

"Following up on Markus' Confessions of a Registry Janitor, I propose a yearly spring cleaning effort. I'll describe what I have in mind and show however far I get by then.". Markus was encouraging to fix several issues in the way they publish their resources.
Nobody looks at Registry to see how reliable is the metadata
Notebook
- Check 1 : cross check between number of HiPS, SIA, SCS services (except VizieR) in 3 different registries: NAVO, GAVO, EUVO - they don't match !
- Check 2: UCDs: are they valid according to astropy.io.votable.ucd.check_ucd ? Found 58 invalid UCDs, top 10 sorted by number of instances...
- Check 3: authors : Last F vs Last, F etc...
- Check 4: Subjects and the UAT: 872 res_subject entries are not in the UAT !
- Check 4b: concepts uat_concept values in the subject_uat table
Effort to be expanded, repeated every year.
What to do with the results ?
- Report cross-checks between registries to their admins
- Compile a report and advertise it at Interop registry and ops sessions
- Generate and send yearly reports of issues to publishers about [the quality of] their Registry records ??
Q&A?
- shall IVOA take action when publishers fail to respond to these reports ?
- NL a large effort? should we create validation tool so that users can be proactive and not wait for the report
- validation errors found in RegTAP but we'd need to provide location of errors in the XML VOResources !
- PLS from previous efforts validating services, having such a tool is useful
- Helpful to have a list discrepancies in a repo ?
- MD: thanks for the work, individual publishers could be running it by adding a restriction to an authority in the pyvo query. But doubt many people will do that. Let's gather for a regular Spring Cleaning: identify 2-3 operators and talk to them so than they can ask back, this is better than a message from a robot. NB: BTW this effort is not about failing services nor failing services, it's about FINDING services because when metadata is bad, the services won't be FOUND.
Consistency checks: by running 3 independant harvesting I see that the most common reason for inconsistencies for out-of-sync is that records vanish without warnings. Can provide my checking tooling to help.
TJ: We could have a Registry Cleaning Hackathon