DataOrigin

June 2, 2025

1 Exploit DataOrigin in VOTable

G.Landais (CDS), June-2025

Note: https://www.ivoa.net/documents/DataOrigin/

DataOrigin goal: list of metadata that describes both query and datasets basic provenance.

- Semantic based on Registry (VOResource) described in RDF documents
- Basic Provenance can be serialized in VOTable result with simple "INFO" tag.

Example: https://vizier.cds.unistra.fr/viz-bin/conesearch/J/MNRAS/473/4130?RA=10.7039167&DEC=41.25666

1.1 Extract basic information from a VizieR query

Library used in this notebook: - module astropy.io.votable, version 7.1.0 (requires python3.11 or later) - module (**pull request**) astropy.io.votable.dataorigin https://github.com/gilleslandais/astropy/

[1]:
```
import astropy.io.votable
import astropy.io.votable.dataorigin as dataorigin
```

Query VizieR catalogue (using CDS protocol ASU).

Catalogue: J/AJ/164/20 : RGB radial velocities in M31 northeast shelf fields (Escala+, 2022)

[2]:
```
aj_votable = astropy.io.votable.parse("https://vizier.cds.unistra.fr/viz-bin/
↪votable?-source=J/AJ/164/20/table2")
aj_votable.get_table_by_index(0).to_table()[:10]
```

[2]:
```
<Table length=10>
recno   ID     RAJ2000     DEJ2000     Vmag  …  e_Imag   RVel   e_RVel Simbad
                                        mag   …   mag    km / s  km / s
int32 int32     str11       str11    float32 … float32 float32 float32  str6
----- ------ ----------- ----------- ------- … ------- ------- ------- ------
    1 133576 00 52 36.06 +42 08 14.3   23.17 …    0.03  -343.9     3.6 Simbad
    2 137351 00 52 36.24 +42 08 57.0   22.81 …    0.02  -346.0     4.0 Simbad
    3 137797 00 52 36.80 +42 09 04.3   23.75 …    0.03  -210.1     8.9 Simbad
    4 140180 00 52 25.13 +42 09 28.9   23.44 …    0.04  -306.0     5.5 Simbad
    5 140853 00 52 28.23 +42 09 37.8   22.62 …    0.05  -270.6     6.6 Simbad
    6 141301 00 52 33.86 +42 09 45.4   23.63 …    0.06  -252.1     3.7 Simbad
```

```
  7 142012 00 52 30.51 +42 09 54.3   23.60 …    0.03  -372.1     5.8 Simbad
  8 143936 00 52 41.88 +42 10 15.7   22.43 …    0.01  -284.3     3.5 Simbad
  9 145261 00 52 37.33 +42 10 30.0   22.98 …    0.02  -298.4     7.3 Simbad
 10 146884 00 52 27.56 +42 10 49.0   23.84 …    0.05  -325.5     3.2 Simbad
```

[3]:
```
aj_dataorigin = dataorigin.extract_data_origin(aj_votable)
print(aj_dataorigin)
```

```
publisher: CDS
server_software: 7.5.1
service_protocol: ASU
request: https://vizier.cds.unistra.fr/viz-
bin/votable?-source=J/AJ/164/20/table2
request_date: 2025-06-01T13:09:07
contact: cds-question@unistra.fr

ivoid: ivo://cds.vizier/j/aj/164/20
citation: doi:10.26093/cds/vizier.51640020
reference_url: https://cdsarc.cds.unistra.fr/viz-bin/cat/J/AJ/164/20
rights_uri: https://cds.unistra.fr/vizier-org/licences_vizier.html
creator: Escala I.
editor: Astronomical Journal (AAS)
cites: bibcode:2022AJ…164…20E
original_date: 2022
publication_date: 2024-11-12
```

The API separates (basic) Provenance and Query informations

[4]:
```
# get query information (reproducibility)
print("Query:")
print(aj_dataorigin.query)
# Get Basic Provenance
print("\nOrigin:")
print(aj_dataorigin.origin[0])
```

```
Query:
publisher: CDS
server_software: 7.5.1
service_protocol: ASU
request: https://vizier.cds.unistra.fr/viz-
bin/votable?-source=J/AJ/164/20/table2
request_date: 2025-06-01T13:09:07
contact: cds-question@unistra.fr

Origin:
ivoid: ivo://cds.vizier/j/aj/164/20
citation: doi:10.26093/cds/vizier.51640020
reference_url: https://cdsarc.cds.unistra.fr/viz-bin/cat/J/AJ/164/20
```

```
rights_uri: https://cds.unistra.fr/vizier-org/licences_vizier.html
creator: Escala I.
editor: Astronomical Journal (AAS)
cites: bibcode:2022AJ…164…20E
original_date: 2022
publication_date: 2024-11-12
```

Query VOTable containing 2 resources Query 2 catalogues having data in M31 - J/AJ/164/20: RGB radial velocities in M31 northeast shelf fields (Escala+, 2022) - J/MNRAS/473/4130: A SITELLE view of M31's central region. I Calibrations and radial velocity catalogue of nearly 800 emission-line point-like sources. (2018)

[5]:
```
m31_votables = astropy.io.votable.parse("https://vizier.cds.unistra.fr/viz-bin/
↪votable?-source=J/AJ/164/20/table2,J/MNRAS/473/4130")
```

[6]:
```
# Table J/AJ/164/20/table2
m31_votables.get_table_by_index(0).to_table()[:5]
```

[6]:
```
<Table length=5>
recno   ID      RAJ2000     DEJ2000      Vmag   …  e_Imag   RVel   e_RVel Simbad
                                          mag   …   mag    km / s  km / s
int32 int32     str11       str11      float32 … float32 float32 float32  str6
----- ------ ----------- ----------- ------- … ------- ------- ------- ------
    1 133576 00 52 36.06 +42 08 14.3   23.17 …    0.03  -343.9     3.6 Simbad
    2 137351 00 52 36.24 +42 08 57.0   22.81 …    0.02  -346.0     4.0 Simbad
    3 137797 00 52 36.80 +42 09 04.3   23.75 …    0.03  -210.1     8.9 Simbad
    4 140180 00 52 25.13 +42 09 28.9   23.44 …    0.04  -306.0     5.5 Simbad
    5 140853 00 52 28.23 +42 09 37.8   22.62 …    0.05  -270.6     6.6 Simbad
```

[7]:
```
# Table J/MNRAS/473/4130
m31_votables.get_table_by_index(1).to_table()[:5]
```

[7]:
```
<Table length=5>
recno          ID           RAJ2000    … XmMOIII Comments   SimbadName
                                         …
int32        str19           str11      …   str4   str13       str14
----- ------------------- ----------- … ------- -------- --------------
    1 S004248.94+411524.0 00 42 48.94 …    Sat.          [MMD2006] 1336
    2 S004253.55+411734.3 00 42 53.55 …    Sat.          [MMD2006] 1261
    3 S004246.47+411702.7 00 42 46.47 …    Sat.          [MMD2006] 1330
    4 S004231.03+411625.3 00 42 31.03 …    Sat.          [MMD2006] 1151
    5 S004249.75+411803.9 00 42 49.75 …    Sat.
```

[8]:
```
# Data Origin
m31_dataorigin = dataorigin.extract_data_origin(m31_votables)
print(m31_dataorigin)
```

```
publisher: CDS
```

```
server_software: 7.5.1
service_protocol: ASU
request: https://vizier.cds.unistra.fr/viz-
bin/votable?-source=J/AJ/164/20/table2,J/MNRAS/473/4130
request_date: 2025-06-01T13:11:14
contact: cds-question@unistra.fr

ivoid: ivo://cds.vizier/j/aj/164/20
citation: doi:10.26093/cds/vizier.51640020
reference_url: https://cdsarc.cds.unistra.fr/viz-bin/cat/J/AJ/164/20
rights_uri: https://cds.unistra.fr/vizier-org/licences_vizier.html
creator: Escala I.
editor: Astronomical Journal (AAS)
cites: bibcode:2022AJ…164…20E
original_date: 2022
publication_date: 2024-11-12

ivoid: ivo://cds.vizier/j/mnras/473/4130
citation: doi:10.26093/cds/vizier.74734130
reference_url: https://cdsarc.cds.unistra.fr/viz-bin/cat/J/MNRAS/473/4130
rights_uri: https://cds.unistra.fr/vizier-org/licences_vizier.html
creator: Martin T.B.
cites: bibcode:2018MNRAS.473.4130M
original_date: 2018
publication_date: 2024-08-17
```

Extract infos - get contact - get DOI(s) - Make an APA citation

[10]:
```
print(f"Contact: {m31_dataorigin.query.contact}")

for origin in m31_dataorigin.origin:
    print(f"\nivoid: {origin.ivoid[0]}, citation={origin.citation[0]}\n")

    # Citation in APA style
    vo_elt = origin.get_votable_element()
    title = vo_elt.description if vo_elt else ""
    print(f"APA: {','.join(origin.creator)} ({origin.publication_date[0]}).␣
 ↪{title} [Dataset]. {m31_dataorigin.query.publisher}. {origin.citation[0]}")
```

```
Contact: cds-question@unistra.fr

ivoid: ivo://cds.vizier/j/aj/164/20, citation=doi:10.26093/cds/vizier.51640020

APA: Escala I. (2024-11-12). RGB radial velocities in M31 northeast shelf fields
(Escala+, 2022) [Dataset]. CDS. doi:10.26093/cds/vizier.51640020

ivoid: ivo://cds.vizier/j/mnras/473/4130,
citation=doi:10.26093/cds/vizier.74734130
```

APA: Martin T.B. (2024-08-17). M31 center emission-line point-like sources (Martin+, 2018) [Dataset]. CDS. doi:10.26093/cds/vizier.74734130

DataOrigin and Provenance (ProvDM)

Use voprov Python package (M.Sanguillon): https://github.com/sanguillon/voprov

In this part, we will use: - module astropy.io.votable - module astropy.io.votable.dataorigin - pyvo - voprov - module dataorigin_prov : **home-made library not planed to be share yet**

[11]:
```
#from dataorigin import provenance
import dataorigin_prov.provenance as provenance
from voprov.visualization.dot import prov_to_dot
from IPython.display import Image
import sys
```

Extract Provenance from DataOrigin and print in Turtle (rdf)

[12]:
```
pdoc = provenance.extract_prov_from_dataorigin(aj_dataorigin,␣
 ↪entity_name="aj_resultset")
pdoc.serialize(sys.stdout, format='rdf', rdf_format='ttl')
#print(pdoc.get_provn())
```

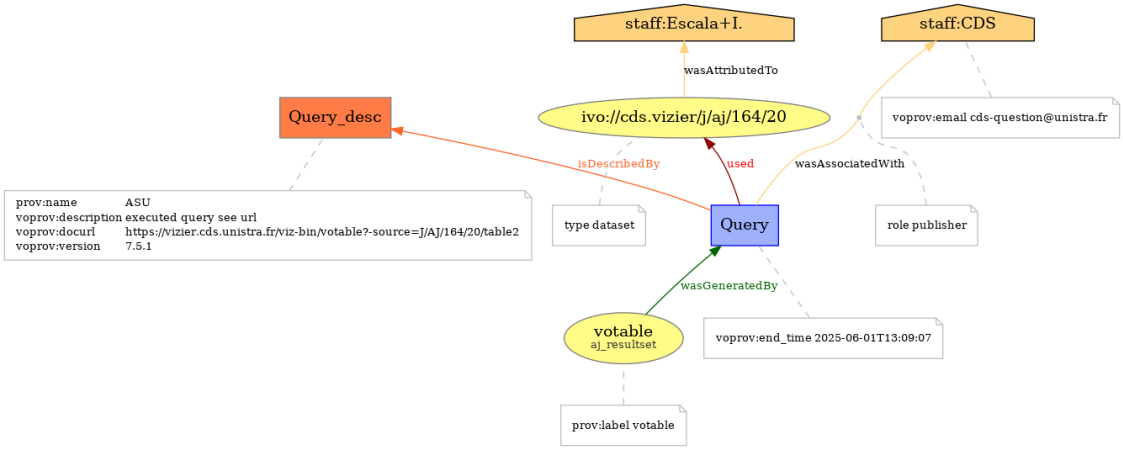
```
@prefix prov: <http://www.w3.org/ns/prov#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix staff: <https://orcid.org/orcid-search/search?searchQuery=> .
@prefix voprov: <http://www.ivoa.net/documents/ProvenanceDM/index.html#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

voprov:aj_resultset a voprov:Entity ;
    rdfs:label "votable"^^xsd:string ;
    prov:wasGeneratedBy voprov:Query .

voprov:Query a voprov:Activity ;
    voprov:end_time "2025-06-01T13:09:07"^^xsd:string ;
    prov:isDescribedBy voprov:Query_desc ;
    prov:qualifiedAssociation [ a voprov:Association ;
            voprov:role "publisher"^^xsd:string ] ;
    prov:used <https://ivoa.net/#//cds.vizier/j/aj/164/20> ;
    prov:wasAssociatedWith staff:CDS .

voprov:Query_desc a voprov:ActivityDescription ;
    voprov:description "executed query see url"^^xsd:string ;
    voprov:docurl "https://vizier.cds.unistra.fr/viz-
bin/votable?-source=J/AJ/164/20/table2"^^xsd:string ;
    voprov:version "7.5.1"^^xsd:string ;
    prov:name "ASU"^^xsd:string .
```

```
<https://ivoa.net/#//cds.vizier/j/aj/164/20> a voprov:Entity ;
    voprov:type "dataset"^^xsd:string ;
    prov:wasAttributedTo <https://orcid.org/orcid-
search/search?searchQuery=Escala+I.> .

staff:CDS a voprov:Agent ;
    voprov:email "cds-question@unistra.fr"^^xsd:string .

<https://orcid.org/orcid-search/search?searchQuery=Escala+I.> a voprov:Agent .
```

Make a graph

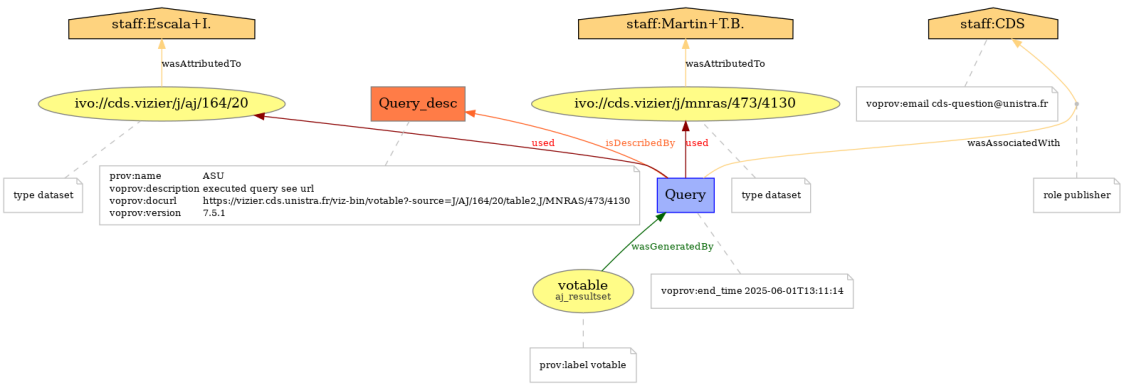
[13]:
```
dot = prov_to_dot(pdoc, use_labels=True)
dot.write_png('/tmp/voprov_example.png')
Image('/tmp/voprov_example.png')
```

[13]:



[14]:
```
pdoc = provenance.extract_prov_from_dataorigin(m31_dataorigin,␣
 ↪entity_name="aj_resultset")
dot = prov_to_dot(pdoc, use_labels=True)
dot.write_png('/tmp/voprov_example.png')
Image('/tmp/voprov_example.png')
```

[14]:



1.2 Use ProvDM to serialize an ivoa workflow

In the example we will make a conesearch on VizieR (j/aj/164/20) and then reuse the result to query by TAP the GAVO ObsCore table

[15]:
```
import pyvo
```

[16]:
```
# Make a VizieR conesearch
vizier_resource = pyvo.registry.search(ivoid="ivo://CDS.VizieR/j/aj/164/20")[0]
vizier_conesearch = vizier_resource.get_interface(service_type='conesearch').
  ↪to_service()
vizier_records = vizier_conesearch.search(pos=(13.1502500, 42.1373), sr=2/60.)
```

Note: comptible with pyvo

[17]:
```
vizier_dataorigin = dataorigin.extract_data_origin(vizier_records.votable)
print(vizier_dataorigin)
```

```
publisher: CDS
server_software: 7.5.1
service_protocol: ivo://ivoa.net/std/ConeSearch/v1.03
request: https://vizier.cds.unistra.fr/viz-bin/conesearch/J/AJ/164/20/table2?RA=
13.15025&DEC=42.1373&SR=0.03333333333333333&VERB=2
request_date: 2025-06-01T13:15:58
contact: cds-question@unistra.fr

ivoid: ivo://cds.vizier/j/aj/164/20
citation: doi:10.26093/cds/vizier.51640020
reference_url: https://cdsarc.cds.unistra.fr/viz-bin/cat/J/AJ/164/20
rights_uri: https://cds.unistra.fr/vizier-org/licences_vizier.html
creator: Escala I.
editor: Astronomical Journal (AAS)
cites: bibcode:2022AJ…164…20E
original_date: 2022
```

```
publication_date: 2024-11-12
```

Query Gaia in GAVO

[18]:
```
tap_service = pyvo.dal.TAPService("http://dc.g-vo.org/tap")
sql = '''
SELECT TOP 100  *
  FROM ivoa.obscore LEFT JOIN TAP_UPLOAD.vizier
    ON 1=CONTAINS(POINT("_RAJ2000", "_DEJ2000"), CIRCLE(s_ra,s_dec,0.1))
'''
resultset = tap_service.run_sync(uploads={'vizier': vizier_records}, query=sql)
```

[19]:
```
resultset_dataorigin = dataorigin.extract_data_origin(resultset.votable)
print(resultset_dataorigin)
```

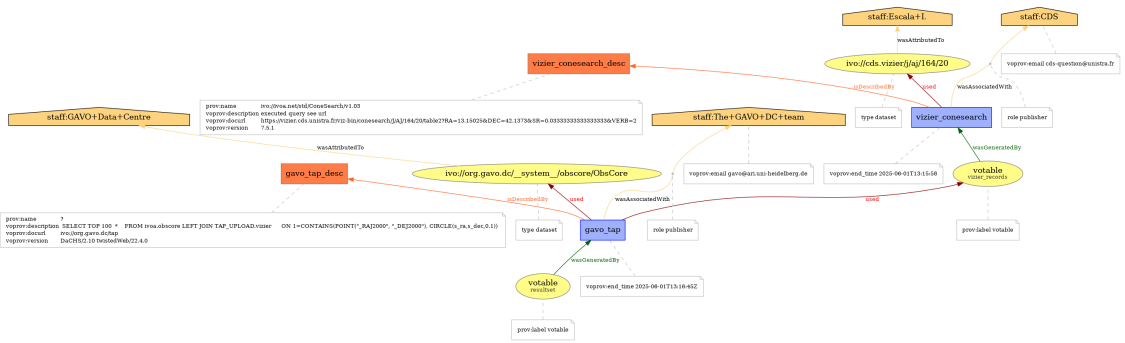
```
ivoid_service: ivo://org.gavo.dc/tap
publisher: The GAVO DC team
server_software: DaCHS/2.10 twistedWeb/22.4.0
query:  SELECT TOP 100  *    FROM ivoa.obscore LEFT JOIN TAP_UPLOAD.vizier
ON 1=CONTAINS(POINT("_RAJ2000", "_DEJ2000"), CIRCLE(s_ra,s_dec,0.1))
request_date: 2025-06-01T13:16:45Z
contact: gavo@ari.uni-heidelberg.de

ivoid: ivo://org.gavo.dc/__system__/obscore/ObsCore
citation: http://dc.g-vo.org/__system__/tap/run/howtocite,http://dc.g-
vo.org/tableinfo/ivoa.ObsCore#ti-citing
reference_url: http://dc.g-vo.org/__system__/tap/run/info,http://dc.g-
vo.org/tableinfo/ivoa.ObsCore
creator: GAVO Data Centre
```

Join the 2 graphs. - make Provenance for the 2 steps (vizier_prov and resulset_prov) - Update the final Provenance Activity with the result Entity of the first step (vizier conesearch) using *"used"* relation

[20]:
```
vizier_prov = provenance.extract_prov_from_dataorigin(vizier_dataorigin,␣
 ↪entity_name="vizier_records", activity_name="vizier_conesearch")
resultset_prov = provenance.extract_prov_from_dataorigin(resultset_dataorigin,␣
 ↪entity_name="resultset", activity_name="gavo_tap")
provenance.link_prov_activity(prov_final=resultset_prov,
                              prov_used=vizier_prov,
                              id_activity_final="gavo_tap",
                              id_entity_used="vizier_records")
```

[21]:
```
dot = prov_to_dot(resultset_prov, use_labels=True)
dot.write_png('/tmp/voprov_example.png')
Image('/tmp/voprov_example.png')
```

[21]:



1.3 Limits on for DataOrigin serialisation

Make a TAP uery that crossmatches 2 tables: - J/AJ/164/20/table2 - I/355/gaiadr3

[22]:
```
tap_service = pyvo.dal.TAPService("https://tapvizier.cds.unistra.fr/TAPVizieR/
 ↪tap")
sql = '''
SELECT TOP 100  *
  FROM "J/AJ/164/20/table2" aj LEFT JOIN "I/355/gaiadr3" as gaia
    ON DISTANCE(POINT('ICRS', aj.RAJ2000, aj.DEJ2000), POINT('ICRS', gaia.
 ↪RA_ICRS, gaia.DE_ICRS))<5/3600.
'''
xmatch = tap_service.run_sync(query=sql)
```

[23]:
```
xmatch
```

[23]:
```
<DALResultsTable length=100>
recno   ID        RAJ2000       … e_RAJ2000 e_DEJ2000 RADEcorJ2000
                    deg         …    mas         mas
int32 int32       float64       …  float64   float64    float32
----- ------ ------------------ … --------- --------- -------------
  306 359932 12.451958333333332 …        --        --            --
  309 375396 12.447166666666664 …        --        --            --
  314 382592 12.479791666666666 …        --        --            --
  318 390421  12.48670833333333 …        --        --            --
  315 383898 12.467749999999999 …        --        --            --
  341 439818 12.521374999999999 …        --        --            --
  347 453850 12.531249999999998 …        --        --            --
  359 470020 12.540291666666665 …        --        --            --
  362 472882 12.530041666666666 …        --        --            --
  …     …                  … … …       …         …             …
   25 158008 13.217458333333331 …        --        --            --
   29 161854 13.207333333333331 …        --        --            --
```

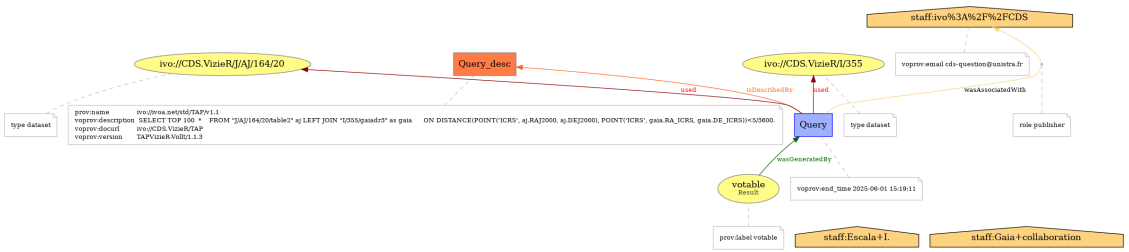
```
 38 163964 13.229916666666664 …        --        --          --
  8 143936 13.174499999999998 …        --        --          --
 16 151915 13.161499999999998 …        --        --          --
 19 154525 13.186749999999998 …        --        --          --
 14 151036 13.142083333333332 …        --        --          --
 15 151337 13.146083333333332 …        --        --          --
 20 154782 13.136166666666664 …        --        --          --
```

[24]:
```
xmatch_dataorigin = dataorigin.extract_data_origin(xmatch.votable)
print(xmatch_dataorigin)
```

```
ivoid_service: ivo://CDS.VizieR/TAP
publisher: ivo://CDS
server_software: TAPVizieR-Vollt/1.1.3
service_protocol: ivo://ivoa.net/std/TAP/v1.1
query:  SELECT TOP 100  *    FROM "J/AJ/164/20/table2" aj LEFT JOIN
"I/355/gaiadr3" as gaia      ON DISTANCE(POINT('ICRS', aj.RAJ2000, aj.DEJ2000),
POINT('ICRS', gaia.RA_ICRS, gaia.DE_ICRS))<5/3600.
request_date: 2025-06-01 15:19:11
contact: cds-question@unistra.fr

ivoid: ivo://CDS.VizieR/J/AJ/164/20,ivo://CDS.VizieR/I/355
reference_url: https://cdsarc.cds.unistra.fr/viz-
bin/cat/J/AJ/164/20,https://cdsarc.cds.unistra.fr/viz-bin/cat/I/355
rights_uri: https://cds.unistra.fr/vizier-
org/licences_vizier.html,https://cds.unistra.fr/vizier-org/licences_vizier.html
creator: Escala I.,Gaia collaboration
article: 2022AJ…164…20E,2022yCat.1355…0G
publication_date: 2022-11-04 08:10:20+01,2022-06-20 08:37:42+02
```

[25]:
```
xmatch_prov = provenance.extract_prov_from_dataorigin(xmatch_dataorigin)
dot = prov_to_dot(xmatch_prov, use_labels=True)
dot.write_png('/tmp/voprov_example.png')
Image('/tmp/voprov_example.png')
```

[25]:



1.4 Issues with “INFO” tag to distinguish metadata by resources.

1.5 In preparation (RFC) a mivot/Mango response

1.5.1 astroquery

No interface available for astroquery yet because astroquery API provides astropy.table and not astropy.io.votable

[ ]: