

DataOrigin scope

Data Curation & Preservation



Describe Origin != Describe the data

Definition: A set of information about the origin of distributed data, which may have been subject to selections, combination or transformation in data streams or in a user request.



Dublin Core

- Identifiers
- Authors
- License
- ...

Reproducibility Metadata

- Data Center (data location)
- Standard, URL, parameters
- Execution date
- ...

VO framework

A very advanced interoperable framework based on a robust architecture allowing discovery, query, manipulation of datasets of different type and origin.

Based on fine grained data entities, but **no Origin (Provenance) in data transfer.**

DataOrigin motivations

Data Curation & Preservation



Motivations

Add a minimum of Provenance in VO standard workflows

- Improve the dataset understanding (Bibliography, ...)
- Allows to execute again a query (QUERY, PROTOCOL, StandardID, ..)
- Allow to extract citation (need authors, DOI, ...)
- Keep the Origin in serialized resultsets
(allows to execute again and to have the origin of a saved result)

Data Origin note

Light Provenance (based on VOResource) in VOTable to improve Data understanding, reproducibility and citation

- A basic provenance metadata
- Information on Query execution
- VOTable sterilization

```
<INFO name="service_protocol" value="ivo://ivoa.net/std/ConeSearch/v1.03"> IVOID of the protocol through which the data was retrieved</INFO>
<INFO name="request_date" value="2025-10-03T07:26:14"> Query execution date</INFO>
<INFO name="request" value="https://vizier.cds.unistra.fr/viz-bin/conesearch/I/355/gaiadr3?RA=0&DEC=0&SR=0.05"> Full request URL</INFO>
<INFO name="contact" value="cds-question@unistra.fr"> Email or URL to contact publisher</INFO>
<INFO name="server_software" value="7.5.2"> Software version</INFO>
<INFO name="publisher" value="CDS"> Data centre that produced the VOTable</INFO>
<INFO name="ivoid" value="ivo://cds.vizier/i/355"> IVOID of underlying data collection </INFO>
<INFO name="creator" value="Gaia collaboration"> First author or institution </INFO>
<INFO name="cites" value="bibcode:2022yCat.1355....0G"> Bibcode of the dataset </INFO>
<INFO name="original_date" value="2022"> Year of the article publication </INFO>
<INFO name="reference_url" value="https://cdsarc.cds.unistra.fr/viz-bin/cat/I/355"> Dataset landing page </INFO>
<INFO name="citation" value="doi:10.26093/cds/vizier.1355"> Dataset identifier that can be used for citation </INFO>
<INFO name="publication_date" value="2024-12-12"> Date of first publication in the data centre </INFO>
<INFO name="rights_uri" value="https://cds.unistra.fr/vizier-org/licences_vizier.html"> Licence URI </INFO>
```

DataOrigin History

Data Curation & Preservation



History

- IVOA 2022: discussion on metadata expected
 - Use case: citation, reproducibility – services impacted
 - Establish a list of metadata
- 2022/10/30: version 1.0 with a VOTable serialisation based on <INFO>
- 2022... : implementation in CDS, in Dachs (Gavo)
- 2024/01/26: version 1.1 minor update
- 2025/04 : astropy module `astropy.io.votable.DataOrigin` (v7.1.0)

IVOA Note

International Virtual Observatory
Alliance

IVOA Documents

Data Origin in the VO
Version 1.1

IVOA Note 26 January 2024

Interest/Working Group:
<http://www.ivoa.net/wiki/bin/view/IVOA/IvoaCP>

Author(s):
Gilles Landais, August Muench, Markus Demleitner, Renaud Savalle

Editor(s):
Gilles Landais

Query Execution metadata (~reproducibility)

Key	Description	Level	Dublin Core
ivoid	IVOID of underlying data collection	R	
publisher	Data centre that produced the VOTable	R	publisher
server_software	Software version (*)		
service_protocol	IVOID of the protocol through which the data was retrieved	R	
request	Full request URL including a query string (**)	R	
query	An input query in a formal language (e.g. ADQL)		
request_date	Query execution date	R	
contact	Email or URL to contact publisher		
(*) Operators are encouraged to follow Demleitner and Taylor (2021) in this item			
(**) For "Simple" protocols (regardless of the HTTP method), put the application/x-www-form-urlencoded form of the query parameters in the query part of the URL here. More complex scenarios like UWS are not covered by this document.			

Table 1: INFO names available for specifying the query that generated a VOTable

DataOrigin metadata

Data Curation & Preservation



Light Provenance

Key	Description	Level	Dublin Core
citation	Dataset identifier that can be used for citation (e.g. dataset DOI)	R	identifier
reference_url	Dataset landing page		
resource_version	Dataset version	R	
rights_uri	Licence URI (*)	R	rights
rights	Licence or Copyright text		rights
creator	The person(s) mainly involved in the creation of the resource; generally, the author(s).	R	creator
editor	Editor name (article)		
article	Bibcode or DOI of a reference article		relation
cites	An Identifier (ivoid, DOI, bibcode) of second resource using relation type "cites" (**)		relation
is_derived_from	An Identifier (ivoid, DOI, bibcode) of second resource using relation type "is_derived_from" (**)		relation
original_date	Date of the original resource from which the present resource is derived (DALI timestamp)		
publication_date	Date of first publication in the data centre (DALI timestamp) (***)	R	
last_update_date	Last data centre update (DALI timestamp) (****)	R	date
(*) Following Registry practice, this should come from SPDX https://spdx.org/licenses/ , though Creative Commons URLs https://creativecommons.org are also admitted			
(**) https://www.ivoa.net/rdf/voresource/relationship_type/			
(***) Equivalent to curation/date[@role='created'] in registry			
(****) Equivalent to curation/date[@role='updated'] in registry			

Table 2: INFO names available for specifying information related to the origin of the data set(s) a VOTable was generated from

ASUCC 2023, Trieste - Data Origin in VO

DataOrigin implementations

Data Curation & Preservation




The provider implementations

- CDS : VizieR Catalogues
Simbad (Query execution)
- Gavo (Dachs)
- Paris Observatory (Dachs),
- Other ?

The Client implementations

Data Origin is available in :

- Topcat
- Aladin
- astropy



TOPCAT(2): Table Parameters

Window Parameters Display Help

Table Parameters for 2: gaiadr3?RA=13.1502500&DEC=42.137&SR=0.05

Name	Value	Description
Name	/355/gaiadr3	Table name
Column Count	226	Number of columns
Row Count	99	Number of rows
Description	Gaia DR3 source catalog (1811709771 sources)	
ivoid	ivo://cds.vizier/I/355	IVOID of underlying data collection
creator	Gaia collaboration	First author or institution
cites	bibcode:2022yCat.1355....0G	Bibcode of the dataset
original_date	2022	Year of the article publication
reference_url	https://cdsarc.cds.unistra.fr/viz-bin/cat/I/355	Dataset landing page
citation	doi:10.26093/cds/vizier.1355	Dataset identifier that can be used for citation
publication_date	2024-12-12	Date of first publication in the data centre
rights_uri	https://cds.unistra.fr/vizier-org/licences_vizier.html	Licence URI
matches	99	matching records
service_protocol	ivo://voa.net/std/ConeSearch/v1.03	IVOID of the protocol through which the data was retrieved
request_date	2025-10-03T09:20:48	Query execution date
request	https://vizier.cds.unistra.fr/viz-bin/conesearch/I/355/gaiadr...	Full request URL
contact	cds-question@unistra.fr	Email or URL to contact publisher
server_software	7.5.2	Software version
publisher	CDS	Data centre that produced the VOTable
MaxTuples	50000	

DataOrigin usage

Data Curation & Preservation



Using DataOrigin

Presentation in IVOA 2025 CollegePark

- Extract citation
- Extract information from VOTable in RDF/Provenance

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&SR=0.1>

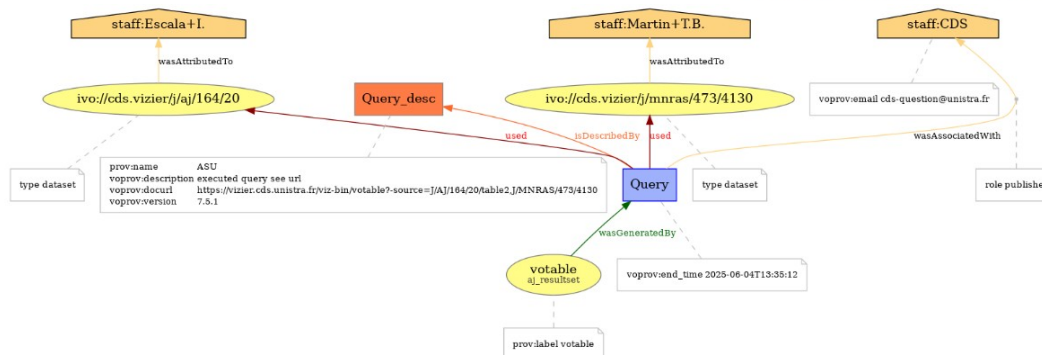
Extract basic information from a Vizier query

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
```



DataOrigin usage

Data Curation & Preservation



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

[6]: aj_votable = astropy.io.votable.parse("https://vizier.cds.unistra.fr/viz-bin/conesearch/J/AJ/164/20/table2?RA=13.1502500&DEC=42.137&SR=1")
do = dataorigin.extract_data_origin(aj_votable)
print(do)
```

```
publisher: CDS
server_software: 7.5.2
service_protocol: ivo://ivoa.net/std/ConeSearch/v1.03
request: https://vizier.cds.unistra.fr/viz-bin/conesearch/J/AJ/164/20/table2?RA=13.1502500&DEC=42.137&SR=1
request_date: 2025-10-03T08:12:51
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
```

Citation

```
[11]: origin = do.origin[0]
vo_elt = origin.get_votable_element()
title = vo_elt.description if vo_elt else ""

apa_citation = f"APA: {' '.join(origin.creator)} ({origin.publication_date[0]}). {title} [Dataset]. {do.query.publisher}. {origin.citation[0]}"
print(apa_citation)
```

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




With DataOrigin, resultset containing Origin information can be serialized and reuse later !

And next ?

- Promote DO (eg: ESA SCS provides information matching with DO)
- Astroquery implementation
- Ask for implementation in standards
 - SCS 1.1
 - SCS 2.0
 - DALI
- Investigation for improvement
 - Other metadata: ivoid for service, instruments ?
 - Limitation for complex ADQL query (join operation)
 - Investigate for a record granularity (eg: ObsCore) (reported by P.Dowler)
→ needs a DataModel

Currently, we propose to add a **simple reference to the note in standards.**

Proposal for an Endorsed note (discussion with M.Demleitner, P.Dowler, G.Landais, M.Molinaro)