



Data Origin in the VO

Contributors:

G.Landais, M.Demleitner, R.Savalle, G. Muench

Thank to all participants
(Data Origin splinter, IVOA October 2022)

Data Origin



Definition : Information on the origin of distributed data that may result from selections, changes in data flows or in a user query.

Decorate VO results with metadata that describes the Data Origin

Metadata in result :

- Description of the content (eg: VOTable, Field description, UCD, etc)
- Description of the query (protocol, query, etc.)
- Description, of the Dataset queried (publication, authors, links, ...)

Example of metadata :



Dublin Core

- Identifier
- Authors
- Licence
- ...

Meta-data for access (~reproducibility)

- Data Center
- URL + parameters
- Execution Date
- ...

Motivations



- Improve Data understanding for end users
- Reproducibility
- Citation

- Report origin information in all step of the curation workflow
Improve the curation workflow « Data center → users »
- Datasets proliferation
The same Dataset distributed in many Data Centers that operate Data curation
(Ex : Gaia in Gavo, CDS, ESA, ...)

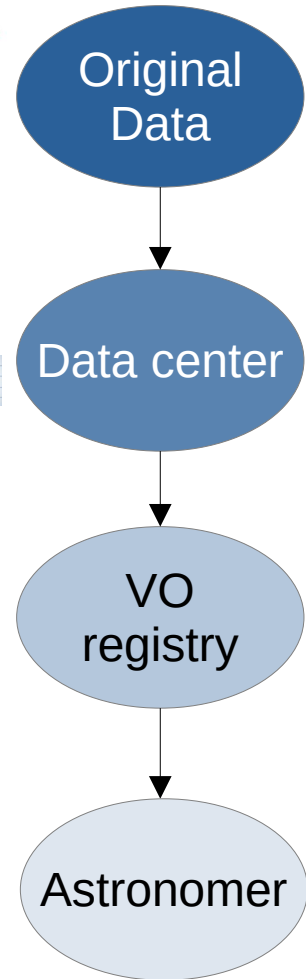
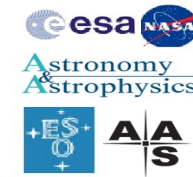
”Same in content, customized serialization ”

eg of customization :

- Added values: links, crossmatch
- Column selections, column format
- Provide metadata to trace origin
 - Make a bibtex
 - Contact the Data Center

Difficulties to get the information :

- Find the way to query the resource « landing page »
- The metadata access of each Data center depends of its implementation



Where to find Data Origin in the VO ?



- **VO registry**
includes DublinCore, +- Datacite compatible
 - Identifiers, authors, publication date, rights, ...
 - References to other resource (eg: bibliographic links, etc..)

- **Data Models**
 - ProvDM
 - LastSetpProvenance (not a standard)
 - DatasetDM (not a standard)

- **VOTable**
 - No standard way to put ivoId, DOI, bibcode, authors, ...
 - Mivot (in RFC)

- **Protocols ?**
 - DALI: minimal information: QUERY_STRING, citation, standardID (see G.Mantelet talk)
 - TAP ? SCS ? Etc.
 - regTAP

Meta-data proposal



<https://github.com/gilleslandais/ivoa-dcp-data-origin>

- Meta-data list
 - Meta-data used for reproducibility
 - Meta-data used to trace Origin (Dublin Core extension)
- Data Origin in the Registry
- VOTable serialization



Data Origin in the VO Version 1.0

IVOA Note 2022-10-30

Working Group

DCP

This version

<https://www.ivoa.net/documents/data-origin/20221030>

Latest version

<https://www.ivoa.net/documents/data-origin>

Previous versions

This is the first public release

Author(s)

G.Landais, G.Muench, M.Demleitner, R.Savalle, looking for contributors

Editor(s)

G.Landais

Abstract

The goal of the document is to make the Data Origin more visible in the query results executed in the Virtual Observatory. The document lists metadata required to provide sufficient traceability to end-users in order to improve the understanding of the resultsets and enabling its reuse and its citation.

NOTE in work - template for a possible IVOA note.

Status of this document

Meta-data proposal



Meta-data used for reproducibility

Meta-data	Description	Mandatory
ivoid	ivoid identifier to link registry	yes
publisher	Data center that provides the VOTable	yes
version	Service/software version (or release date)	yes
service_protocol	Protocol access with version	
request	Request URL	
request_post	(POST Request) POST arguments	
request_date	Query execution date	
contact	email or URL contact	
landing_page	Dataset landing page	

Meta-data proposal



Meta-data used to trace the Origin

Metadata	Description	Level	Dublin Core	Registry
publication_id	Dataset identifier that can be used for citation	M	identifier	altIdentifier
curation_level	Controlled vocabulary (IVOA rdf, content_level)			contentLevel
resource_version	Dataset version or last release	R		version
rights	Licence URI (standard spdx form preferred : https://spdx.dev/)		rights	rights
rights_type	Licence type (eg: CC-by, CC-0, ...)			
copyrights	Copyright text			rights
creator	The person or organization primarily responsible for creating the intellectual content of the resource.	R	creator	creator
editor	Editor name			
relation_type	An identifier of a second resource and its relationship to the present resource. Controlled vocabulary		relation	relationship
related_resource	Information about a second resource from which the present resource is derived.		source	relatedResource
publication_date	Date of publication	R		date
resource_date	Date of the original publication	R		

VOTable serialization



- Simple serialization based on <INFO> tag
- Do not required standard modification
- Well integrated in clients :
 - TOPcat
 - Fork Astropy (M.Demleitner)



TOPCAT(3): Table Parameters

Window Parameters Display Help

+ - ? X

Table Parameters for 3: J_AJ_161_36_table8.xml

Name	Value	Description
Name	//AJ/161/36/table8	Table name
Column Count	19	Number of columns
Row Count	6	Number of rows
Description	Planet candidate properties	
ivoid	ivo://cds.vizier/j/aj/161/36	
publisher	doi:10.26093/cds/vizier.51610036	
landing_page	https://cdsarc.cds.unistra.fr/viz-bin/cat//AJ/161/36	
publication_id	doi:10.26093/cds/vizier.51610036	
curation_level	Research	
resource_version	2022-10-07	
rights	https://cds.unistra.fr/vizier-org/licences_vizier.html	
creator	Bryson S.	
related_resource	2021AJ....161...36B	
editor	Astronomical Journal	
publication_date	2021-03-16	
resource_date	2021	
version	7.294	
protocol	Simple Cone Search 1.03	
request_date	2022-10-30T12:08:00	
request	https://vizier.cds.unistra.fr/viz-bin/conesearch//AJ/161/36/...	
contact	cds-question@unistra.fr	

Name: ivooid
Class: String
Shape:
Units:
Description:
UCD:
Utype:
Value: ivo://cds.vizier/j/aj/161/36

Serialization example : VOTable resulting of a VizieR Simple Cone Search (SCS)

https://github.com/gilleslandais/ivoa-dcp-data-origin/blob/master/tests/J_AJ_161_36_table8.xml

VOTable serialization



Aladin output (version beta 12.057)

Aladin v12.0 *** BETA VERSION (based on v12.057) ***

File Edit Image Catalog Overlay Coverage Tool View Interop Help

Available data → 33165
71 VIEW ● OUT VIEW
DSS #PanSTARRS #S
DSS2 color

Command
DSS2 color

Properties
Properties of the plane "info.xml"
PlaneID: info.xml
Color: [Color palette]
Default shape: square
Source: 6
Table information: [Column information...] [Parsing report...]
Specific drawing method
.projection center: 19 01 27.97306 +39 16
.projection: Aitoff
Scaling factor: 0 50 100 200 250 300
Overlay opacity: 0 20 40 60 80 100
[Close]

Frame ICRS Projection Aitoff ALADIN

select
pan
dist
phot
draw
tag
moc
spec
filter
cross
epoch
size
dens.
cont
opac.
zoom
pixel
prop
del

info.xml
DSS2/P/DSS2/color

Search
x 6.706"

Rad.	e_R	Per	Ins.	E_I	e_I	T ₁
0.08	0.08	112.3	1.01	0.08	0.07	47
0.1	0.21	331.55	0.89	0.07	0.07	57
0.12	0.22	425.65	1.7	0.16	0.15	60
0.16	0.09	214.88	1.14	0.1	0.09	57
0.11	0.07	46.18	2.24	0.24	0.22	46
0.27	0.53	160.02	2.13	0.16	0.15	46

19:00:32.01 +39:27:35.4
8.819" x 6.706"
-180
-90
sky
no time filter
6 sel / 6 src 91fps / 502Mb

Catalog information
VOTable format
resource information:
.VERSION: 7.294
.protocol: Simple Cone Search 1.03
.request_date: 2022-10-30T12:08:00
.request: https://vizier.cds.unistra.fr/viz-bin/conesearch/J/AJ/161/36/table8?RA
.contact: cds-question@unistra.fr
.version: 7.294
resource information:
.ivoid: ivo://cds.vizier/j/aj/161/36
.publisher: CDS
.landing_page: https://cdsarc.cds.unistra.fr/viz-bin/cat/J/AJ/161/36
.publication_id: doi:10.26093/cds/vizier.51610036
.curation_level: Research
.resource_version: 2022-10-07
.rights: https://cds.unistra.fr/vizier-org/licences_vizier.html
.creator: Bryson S.
.related_resource: 2021AJ...161...368
.editor: Astronomical Journal
.publication_date: 2021-03-16
.resource_date: 2021
Table J/AJ/161/36/table8
-assuming RADEC in degrees column 18 for RA and 19 for DEC
[RA=17 (proba=100.0%) DE=18 (proba=100.0%) PMRA=-1 (proba=0.0%) PMDEC=-1 (proba=0.0%)
-Coordinate system references found:
ID="J2000_2000.000" => eq_FK5 Eq=J2000 Ep=2000.000
ID="J2000" => eq_FK5 Eq=J2000
=> RA/DEC coordinate conversion not required: ref="J2000_2000.000" => "FK5(J2000_2000.000)"
-Table loaded & parsed in 8ms for 6 objects
Catalog queried, loaded and parsed in 38ms

Extract acknowledgment



Python code: <https://github.com/gilleslandais/ivoa-dcp-data-origin/tree/master/tests>

Template

We extract data published in <related_resource> (<creator>, <resource_date>),
via <publisher> services (ivoa resource=<ivoid>, <publication_date>)
using <service_protocol> (version <version>, executed at <request_date>)

Example

We extract data published in bibcode:2021AJ....161...36B (Bryson S., 2021),
via CDS services (ivoa resource=ivo://cds.vizier/j/aj/161/36, 2021-03-16)
using Simple Cone Search 1.03 (version 7.294, executed at 2022-10-30)



- Light provenance, well integrated in current VO framework
- VizieR beta version (conesearch & ASU access)
eg:
<http://viz-beta.u-strasbg.fr/viz-bin/conesearch/J/ApJ/712/585?RA=0.8&DEC=-18.4&SR=0.1>
- Meta-data in DALI ?