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 decriebe 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 »

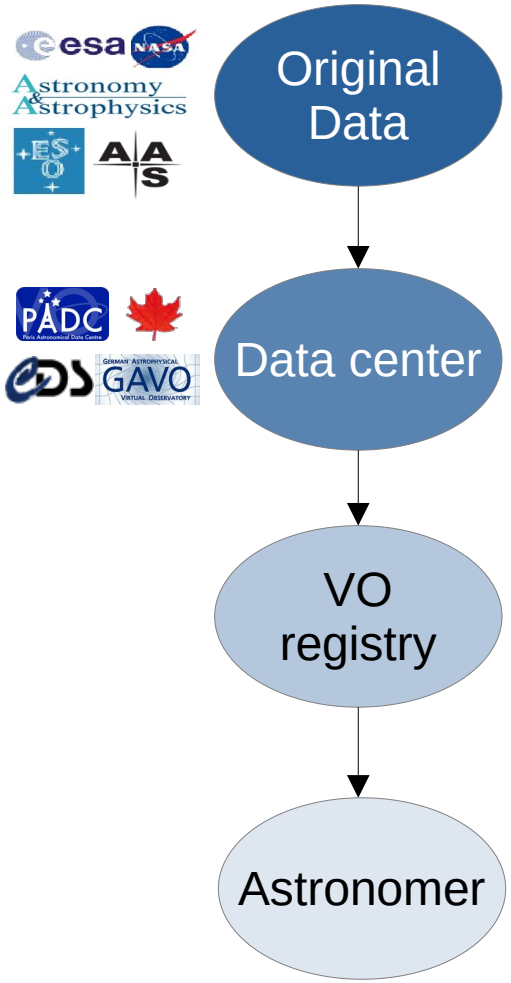
- **Datasets proliferation**
  The same Dataset distributed in many Data Centers that operate curation on the output (Ex : Gaia in Gavo, CDS, ESA, …)

  ”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



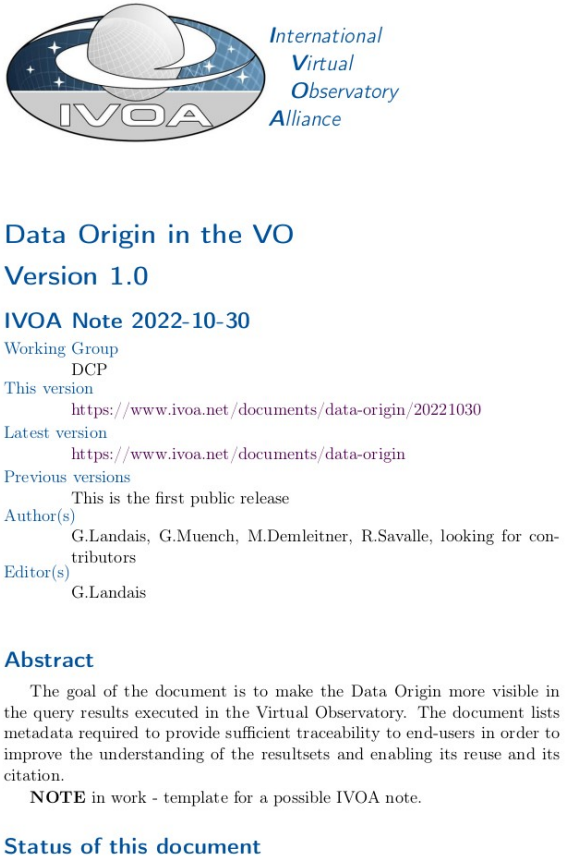
Where to find Data Origin in the VO ?

- **VO registry**
  includes DubinCore, +- Datacite compatible
  - Identifiers, authors, publication date, rights, ...
  - References to other resource (eg: bibiliographic 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

Proposed list of meta-data

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 serialisation



International Virtual Observatory Alliance

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

Proposed list of meta-data

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	

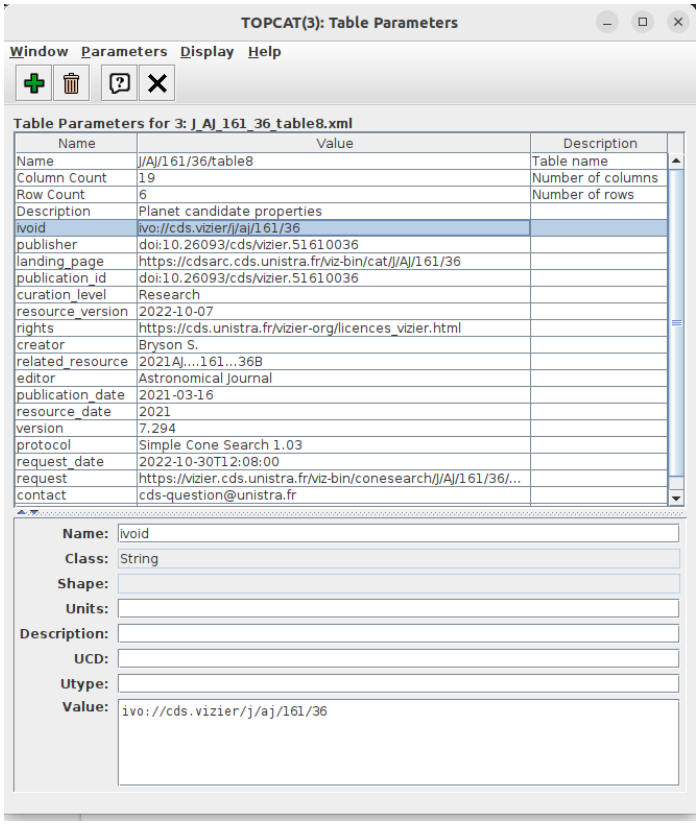
Proposed list of meta-data

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	Controled vocabulary (IVOA rdf, content_level)			contentLevel
resource_version	Dataset version or last release	R		version
rights	Licence URI (standard spdx form prefered :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 serialisation

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

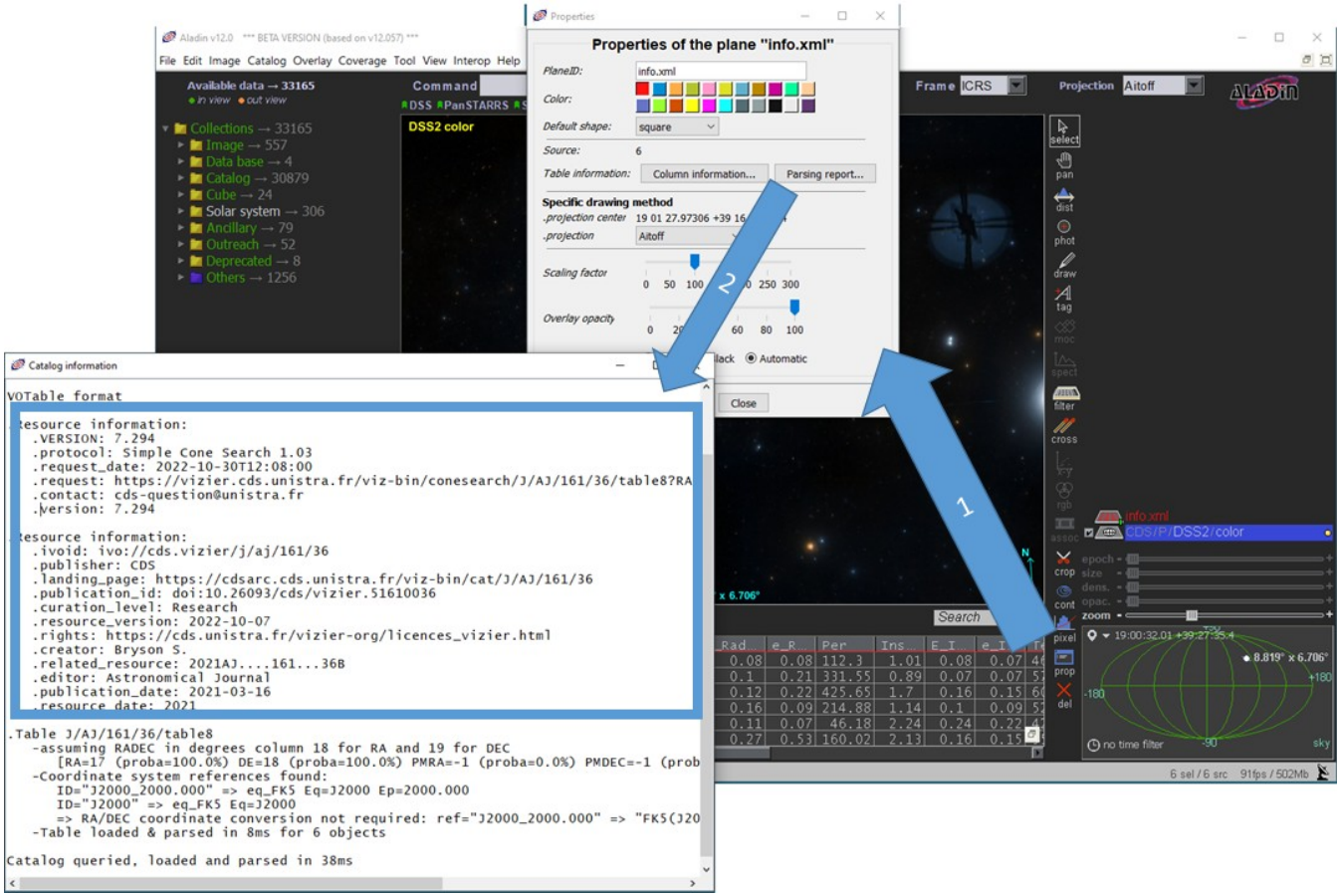


Serialisation example : VOTable resulting of a VizieR Simple Cone Search (SCS)

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

VOTable serialisation

Aladin output (version beta 12.057)



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)
  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 ?