International
Virtual
Observatory
Alliance

Data Origin in the VO
Version 1.0

UNKNOWN DOCUMENT (fix DOCTYPE) ???

Working group
DCP
This version
https://www.ivoa.net/documents/NOTE-ivoa-data-origin/???
Latest version
https://www.ivoa.net/documents/NOTE-ivoa-data-origin
Previous versions
This is the first public release
Author(s)
G.Landais, G.Muench, 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 meta-data required to provide sufficient traceability to end-users in order to improve the understanding of the resultsets and enabling its reuse and its citation.

Status of this document

UNKNOWN DOCUMENT (fix DOCTYPE)

A list of current IVOA Recommendations and other technical documents can be found at https://www.ivoa.net/documents/.

Contents

Acknowledgments

???? Or remove the section header ????

Conformance-related definitions

The words "MUST", "SHALL", "SHOULD", "MAY", "RECOMMENDED", and "OPTIONAL" (in upper or lower case) used in this document are to be interpreted as described in IETF standard RFC2119 (**?**).

The *Virtual Observatory (VO)* is a general term for a collection of federated resources that can be used to conduct astronomical research, education, and outreach. The International Virtual Observatory Alliance (IVOA) is a global collaboration of separately funded projects to develop standards and infrastructure that enable VO applications.

1 Introduction

Data origin is required for end users to understand data, for citation and for reusability. The provenance is cited as a mandatory criterion in the EOSC or in RDA FAIR definition.

The virtual observatory provides an advanced framework to search and consume data provided by Data Centers or Space Agencies who apply curation in different level. In this context, Data Origin in output includes meta-data from the data producer (author, space agency) and the Data Center which hosts the resource. Depending of the implementation, the users can find the origin information in the Data center web pages (landing pages) or in the Registries of the Virtual Observatory. For citation, the ADS search engine provides citation capabilities offering bibtex. There are no VO standards today to get the information easily. For instance, the origin meta-data are not included neither in output format, nor in protocols used to access the data. A list of basics meta-data added in strategic location (as result output or resource listing) would give easier the authors search who is looking for how to cite VO resources. Tracing data origin, from the producer to the final query enables also to report to end users the different agents implied in the data preservation (authors, data center, space agencies, journal)- especially when data can be subject to a curation which depends of the different agents. We propose to list the meta-data which responds to the need of Provenance and methods available today for their implementations.

1.1 Role within the VO Architecture

2 Use cases

- To get provenance information easier
  example: Query the Gaia catalogue using VO services (for instance with topcat or any other VO-software). The registry lists Data Center (eg: Gavo, VizieR, ESA) which provides Gaia tables using TAP. The results returns VOTable having information in the header. However, the information depends of the implementation.
  e.g.:

  - VizieR, GAVO, ESA provide the ADQL
  - GAVO provides the license and ivoid in the output.
  - ESA provide the license, url to a datalink service

  To get citation information require the user to query the providers web sites or to use ADS.

- Relevant meta-data for final users to cite resources
  Provide Data origin output or with API to get all meta-data needed to build bibtex?

- Relevant meta-data for final users to understand data origin
  Table provided by a Data Center can be a copy of an existing resource. For instance, a table published in a journal or by a Space Agency is also hosted in a Data Center like CDS, GAVO, etc. The data curation depends of the Data Center which can add associated data, enrich meta-data (eg: add filter for magnitude) or make a sub-selection of columns.

- Give me a bibliography of everything I've used in the workflow", where we'd probably have handwave away "workflow" and say "you can do it per query result now"

The basics meta-data should contain the data origin (space agency or authors, article references), the data center providing the resource, the date of publication ...

3 State of the art

Data origins consists in a collection of meta-data that can be consumed by users. Data origin is partially included in Data Models or Data Access Layers. We lists a non exhaustive state of the arts of their usage.

3.1 Data origin in Data Models

3.1.1 ObsCore

Obscore is an example which includes in the model Data identifiers and instruments (mandatory) that can be used to put Data origin.

3.1.2 ProvDM

ProvDM is dedicated for workflows. It is an astronomy extension of the W3C Provenance. The result is a graph which is composed by Entities, Activities and Agents linked together.
eg. VizieR workflow can be helpful to get Data Origin:
https://cdsarc.cds.unistra.fr/viz-bin/cat/J/AJ/154/57
https://cdsarc.cds.unistra.fr/viz-bin/provenance?filter=true&out=prov:png&cat=J/AJ/154/57

Note: a Python library enables to trace the Provenance of a workflow in a graph. It is available for users to trace their python activity (M.Servillat)

3.2 Data origin in Access layers

3.2.1 The registry

VO registries,based on OAI-PMH, are harvested by RoR and clients. The registry's schemas allows Data centers to fill data origin in a rich meta-data collection: Identifiers, authors, dates, comments, etc.

List of meta-data useful to trace data origin:

- altIdentifier
- curation.publisher
- curation.creator
- date: role= Updated, role=Created
- contact
- content.subject
- content.description
- content.source : format=bibcode
- content.referenceURL
- content.relationship

3.2.2 TAP

The TAP protocol includes a schema which describes the tables provided in a service. The TAP schema doesn't provide any meta-data about origin.

3.3 Data Origin in VOTable serialization

DALI proposes the additional information “citation” and "IVOA standard ID" to be in the VOTable header (https://ivoa.net/documents/DALI/20170517/REC-DALI-1.1.html)

To fill the data origin, each Data Center adopts its own solution :

- VizieR catalogues: the first author and the publication year are added in the catalogue title description
- Gavo: use the DALI proposal to add ivoid + licences
- other example ?

Currently in development status, the Mango Data Model is a container used to annotate VOTable. MangoDM uses Mivot to serialize Data Models instances. This serialization enables links and can include Data Models in the VOTable header. It could be used to export rich Data Models like Provenance.

3.4 DOI

The DOI is a persistent identifier which includes meta-data. We can extract from the Datacite schema as well as the Crossref schema, information which enables to cite the data, to get authors, dates or licenses information.

The richness of information depends on the implementation of the data producers. The information is not homogeneous.

4 Expected Data Origin

4.1 Link Provenance DM with Data origin

Translated to Provenance "Entites" and "Agents" are sufficient to establish bibtex – In a full Provenance diagram, Entities can be linked each other with Activities. In the Data origin context, we will focus on "Entities" and "Agents".

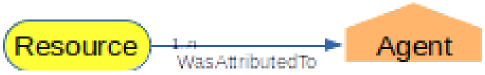

Figure 1: Entities ans Agent in Provenance schema

Note: Why this limitation ? Full provenance is adapted to describe workflows. The result is a graph with many branches and its serialization depends of the need of the producers. The Provenance DataModel is encouraged for any datacenter to describe the workflows using the serialization adapted for its purpose. To extract from the Provenance model only "Entites" and "Agents" enables homogenization of its usage and finally simplify the client parsing (no recursivity needed).

Note: applied to tables, "Entities" and "Agents" meta-data can by assigned to dataset, collection , table or column and could be serialized into VOTable or TAP schema extension.

4.2 List of meta-data

We list meta-data expected to Data origin :

(context: table or dataset resulting from a query) (in bold, the ProvDM equivalent)

4.2.1 Agents

Agents (authors, institutes): any agents included in the workflow

- Author (**prov: Agent.type=Person**)
  - name (**prov: Agent.name**)
  - Orcid (**prov: Agent.id**)
- Institute (**prov: Agent.type=Organization**)
  - Name (**prov: Agent.name**)
  - URL (**prov: Agent.url**)
- Data Center (**prov: Agent.type=Organization**)
  - Name (**prov: Agent.name**)
  - URL (**prov: Agent.url**)
- Original Data Center (in case of copy/mirror) (**prov: Agent.type=Organization**)
  - Name (**prov: Agent.id**)

  - URL (**prov: Agent.url**)

- Editor: journal editor or publisher (**prov: Agent.type=Organization**)
  - name (**prov: Agent.name**)
  - URL : Journal web page (**prov: Agent.url**)

4.2.2 Resources

Any resources used to build the result (**prov: Entity**)

- Query
  - Date of execution (**prov: Entity.generatedAtTime**)
  - query (adql, ...) (**prov: Entity.comment**)
- Table(s) or catalogue(s) (**prov: Entity**)
  - Persistent Identifier (**prov: Entity.id**)
  - VO identifier (ivoid) (**prov: Entity.?**)
  - URL (landing page) (**prov: Entity.url**)
  - Date of publication (**prov: Entity.generatedAtTime**)
  - Curation level (**prov: ActivityDescription.type or Entity.comment (?)**)
  - Curation activity : simple text – eg: "it is a subset from original data ..." (**prov: Activity.comment or Entity.comment (?)**)
- Article(s) (**prov: Entity**)
  - Persistent Identifier (**prov: Entity.id**)
  - Journal name (**prov: Entity.name**)
  - Date of publication (**prov: Entity.generatedAtTime**)
- Any resources completing the original data (**prov: Entity**)
  - resource type: filter , computed (**prov: Entity. ?**)
    eg: computed column for position, time, etc.
  - name (**prov: Entity.name**)
  - comment (**prov: Entity.comment**)
- Software used to generate the query (**prov: Agent=SoftwareAgent**)
  - user-agent (**prov:Agent.comment**)
- Licence (**prov: Entity**)
- Copyrights (**prov: Entity**)

5 Implementation tracks

Several tracks could be explored to provide Data Origin (that we have to define for the context of this document)

5.1 Data Origin included in VOTable output

We dress a non exhaustive list of possible implementation. Tracks using INFO tags are ready-to-client and don't need any development in sofwtare/api.

- DOI-based: delegate Data Origin support to DOI and limit the Data origin implemention with adding DOI in <INFO> in the VOTable header
- URL-based: delegate citation capability, and other information to specialized services.
  The URL could be added in <INFO> in the VOTable headers
  Eg:
  - use link to ADS for citation, license, ...
  - URL to landing page (human readable)
  - URL to service providing machine-readable output (json/XML)
- Semantic-based: define a semantic (mapping with ProvDM could be possible) that can be exploited with <INFO> in VOTable
  e.g.:
  - <INFO name="author" values="name=G. H. Rieke orcid=https://orcid.org/0000-0003-2303-6519"/>
  - <INFO name="article" value="pid=10.3847/1538-3881/ac3b5d date=2022 text=The Astronomical Journal, Volume 163, Number 2"/>
- MiVOT-based: use Mivot capabilities to add rich Provenance into VOTable header

5.2 Data origin using dedicated services

Delegate Data origin capability to services like ADS or VO registries and provides tools/api (for instance in pyVO) based on ivoid to get the information.

5.3 Data origin in Access layer

Delegate Data origin capability to TAPschema which could describe the Data origin of the tables with table like: $TAP\_SCHEMA.prov\_table$, $TAP\_SCHEMA.prov\_agent$ As the previous case, this track needs to provide api/tools to get the information and facilitate the usage.

A Changes from Previous Versions

No previous versions yet.