DOI : for data provider in astronomy and elsewhere

Harry Enke,

EScience & SuperComputing

DOI : what?

- notion of a persistent identifier (PI) for digital entities led to

  * various handle systems (Handle, PURL, ARK …)

  * one variant is the DOI (Digital Object Identifier)

".. International DOI Foundation (IDF), [is] a not-for-profit <u>membership organization</u> that is the governance and management body for the <u>federation of Registration Agencies</u> providing Digital Object Identifier (DOI) services and registration, and is the registration authority for the ISO standard (ISO 26324) for the DOI system. The DOI system provides a technical and social infrastructure for the registration and use of persistent interoperable identifiers, called DOIs, for use on digital networks. (<u>www.doi.org</u>)

- adoption of PIds first in context of librarian efforts to cope with digitised entities,

- thus the well known DOI application for publications.

DOI : what?

for data

- addressing major problems:

* digital objects are by nature volatile, not bound to any

real location or physical realisation

* moving of a digital object leads to difficulties of retrieving,

finding and verifying it again (link rot)

* changing references to such digital objects are expensive

and should be avoided

- digital data plays the fundamental role for most of today's scientific work,

- persistent identification of the data sets is crucial for reliability and reproducibilty

DOI : how?

DataCite was founded in 2009, (European and US Libraries)

* goal: extending DOI to scientific data sets

* registering with DataCite incurs fee (moderate)

(e.g. in Germany academic organisations don't pay, TIB takes care)

* contract between institution and DataCite

* the institution gets its own DOI prefix

By joining a contract with DataCite the institution commits to

* guarantee the validity of its DOI

* update the DataCite registry, when digital

objects change their addresses, or undergo changes

* keep objects with DOI stable

DataCite

* guarantees resolving of the DOI to the actual address of the object

* keeps a basic set of metadata for each data set

DOI : how?

If an institution provides DOI for its data

- definition access policy

- definition of the data set and it's elements

- agreement on licenses ("rights" for data is inappropriate)

- description of the data set with metadata using the (XML) schema by DateCite

- selection of suitable metadata schema of the discipline

- description of the data set by a landing page (for humans)

- definition of a procedure to provide data set and DOI in a concise way

(data curation)

- plan for long term maintenance of the data set

The decision to provide DOI for data leads to a better understanding of the roles and responsibilities in this process.

Thus: only data centers or archives planning for more than casual delivery of some data are able to mint DOI

DOI : Example?

Gaia@AIP | Query | Documentation ▾ | Database tables ▾ | Blog | FAQ | Contact | Harry Enke ▾

gdr2.gaia_source

Description

This table has an entry for every Gaia observed source as listed in the Main Database accumulating catalogue version from which the catalogue release has been generated. It contains the basic source parameters, that is only final data (no epoch data) and no spectra (neither final nor epoch).

Attribution

If you have used Gaia data in your research, please use the following acknowledgement:

> This work has made use of data from the European Space Agency (ESA) mission Gaia (http://www.cosmos.esa.int/gaia), processed by the Gaia Data Processing and Analysis Consortium (DPAC, http://www.cosmos.esa.int/web/gaia/dpac/consortium). Funding for the DPAC has been provided by national institutions, in particular the institutions participating in the Gaia Multilateral Agreement.

More information can be found on the credit and citation instructions page.

Columns

Name	Type	UCD	Unit	Description
solution_id	long	meta.version		Solution Identifier
designation	char	meta.id meta.main		Unique source designation (unique across all Data Releases)
source_id	long	meta.id		Unique source identifier (unique within a particular Data Release)

Access

You can access this table using the query interface and download the results afterwards.

Digital object identifier

You can use the following DOI to cite this table in a publication:

https://doi.org/10.17876/gaia/dr.2/1

DOI : Example?

```xml
<resource xsi:schemaLocation="http://datacite.org/schema/kernel-4 http://schema.datacite.org/meta/kernel-4/metadata.xsd" xmlns="http://datacite.org/schema/kernel-4"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
    <identifier identifierType="DOI">10.17876/gaia/dr.2/1</identifier>
    <creators>
        <creator>
            <creatorName>Gaia Data Processing and Analysis Consortium</creatorName>
        </creator>
    </creators>
    <titles>
        <title xml:lang="en-us">Gaia: Second Data Release</title>
    </titles>
    <publisher>European Space Agency (ESA) Gaia misson and Gaia Data Processing and Analysis Consortium (DPAC)</publisher>
    <publicationYear>2018</publicationYear>
    <subjects>
        <subject subjectScheme="DDC">520 Astronomy</subject>
        <subject>Gaia Data Release 2</subject>
    </subjects>
    <contributors>
        <contributor contributorType="ResearchGroup">
            <contributorName>The Gaia Data Processing and Analysis Consortium (DPAC)</contributorName>
            <affiliation> </affiliation>
        </contributor>
        <contributor contributorType="DataCurator">
            <contributorName>Gaia Archive @ ESAC</contributorName>
            <affiliation>ESAC</affiliation>
        </contributor>
        <contributor contributorType="DataCurator">
            <contributorName>Leibniz Institute for Astrophysics Potsdam (AIP)</contributorName>
            <affiliation>AIP</affiliation>
        </contributor>
    </contributors>
    <dates>
        <date dateType="Updated">2018-04-18</date>
    </dates>
    <language>en-us</language>
    <resourceType resourceTypeGeneral="Dataset">Database</resourceType>
    <relatedIdentifiers>
        <relatedIdentifier relatedIdentifierType="URL" relationType="HasMetadata" relatedMetadataScheme="UCD1+" schemeURI="http://www.ivoa.net/documents/latest/UCDlist.html"
schemeType="Text">https://gaia.aip.de/metadata/gdr2/gaia_source</relatedIdentifier>
        <relatedIdentifier relatedIdentifierType="URL" relationType="HasMetadata" relatedMetadataScheme="Units in the VO" schemeURI="http://ivoa.net/documents/VOUnits/20140523/VOUnits
20140523.pdf" schemeType="Text">https://gaia.aip.de/metadata/gdr2/gaia_source</relatedIdentifier>
        <relatedIdentifier relatedIdentifierType="URL" relationType="IsCitedBy">https://gaia.aip.de/metadata/gdr2/gaia_source</relatedIdentifier>
    </relatedIdentifiers>
    <sizes>
        <size>1692919168 rows</size>
    </sizes>
    <formats>
        <format>csv</format>
        <format>votable</format>
        <format>hdf5</format>
    </formats>
    <version>1.0</version>
    <rightsList>
        <rights rightsURI="https://gea.esac.esa.int/archive/documentation/GDR2/Miscellaneous/sec_credit_and_citation_instructions/">Acknowledgements</rights>
    </rightsList>
```

DOI : FAIR connection

Assigning DOI to data sets helps to make them

Findable

Accessible

Interoperable

Re-usable (reproducible)

DOI : Virtual Observatory approach

DOI for data sets are not used in Virtual Observatory

VO protocols/standards do not really recognize **data set** as an important entity:

VOTable does not have a data set identifier XML tag (and could also use an explict "query" tag to deliver the generating query with the data set)

VOResource starts with citing :

> ".. VO element that can be described in terms of who curates or maintains it and which can be given a name and a unique identifier. Just about anything can be a resource: it can be an abstract idea, such as sky coverage or an instrumental setup, or it can be fairly concrete, like an organisation or a data collection. "

DOI : VO DOI

in a preparatory draft for VOResource 1.1 the desire to 'incorporate'

the DOI schema is expressed,

but as yet:

Data is only perceived as subsumed by service

vr:Resource Type Schema Documentation
Any entity or component of a VO application that is describable and identifiable by an IVOA Identifier

vr:IdentifierURI Type Schema Documentation

A reference to a registry record.

This type should only be used if what is referenced must actually be a true Registry

record; vr:IdentifierURI does not allow query or fragment parts and is hence not suitable

for everything defined by IVOA Identifiers, in particular not standard keys (which are

used for versions of standards, for instance) or dataset identifiers.

When something does not need to be locked down to a reference to a single registry

record, **xs:anyURI** *should be used.*

DOI : VO DOI

Data is only perceived as subsumed by service

- vs:DataService Inheriting from vr:Service, this type is for services that access astronomical data. It adds the ability to describe the data's <u>coverage</u> of the sky, frequency, and time.

- vs:DataCollection This resource declares the existence of a collection of data, what it represents, and how to get it. The access to the data may be limited to a human-readable web page (given by content/referenceURL); however, if the contents of the collection are available statically via a URL (e.g. an FTP URL to a directory containing all the files), that URL can be provided. It can also provide pointers to other IVOA registered services that can be used to access the data.

- vs:CatalogService Inheriting from vs:DataService, this type specifically refers to a service that accesses tabular data. In addition to the coverage information, this type adds the ability to describe the tables and their columns. This is intended for describing services that support the "simple" IVOA data access layer protocols such as Simple Image Access [SIA] and Simple Cone Search [SCS].

no alternative schema for defining data sets (e.g. using a DOI schema) is allowed

DOI : VOResource

DOI occurs only once: (as Example)

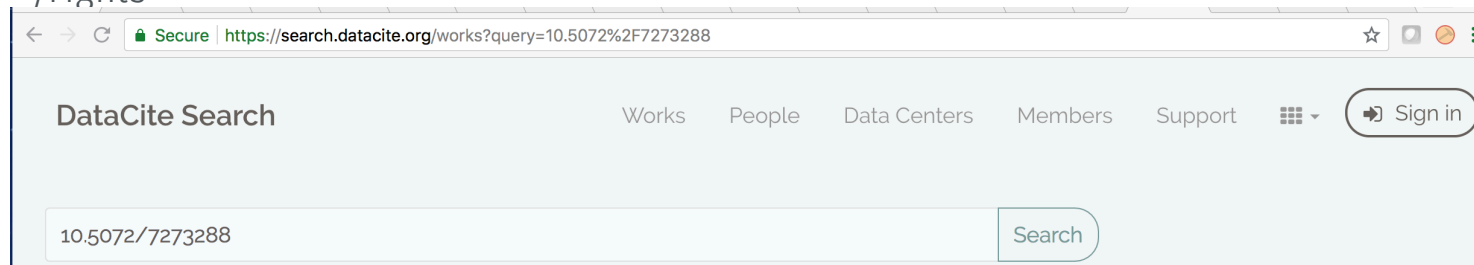`<rights rightsURI="http://creativecommons.org/licenses/by/3.0">`

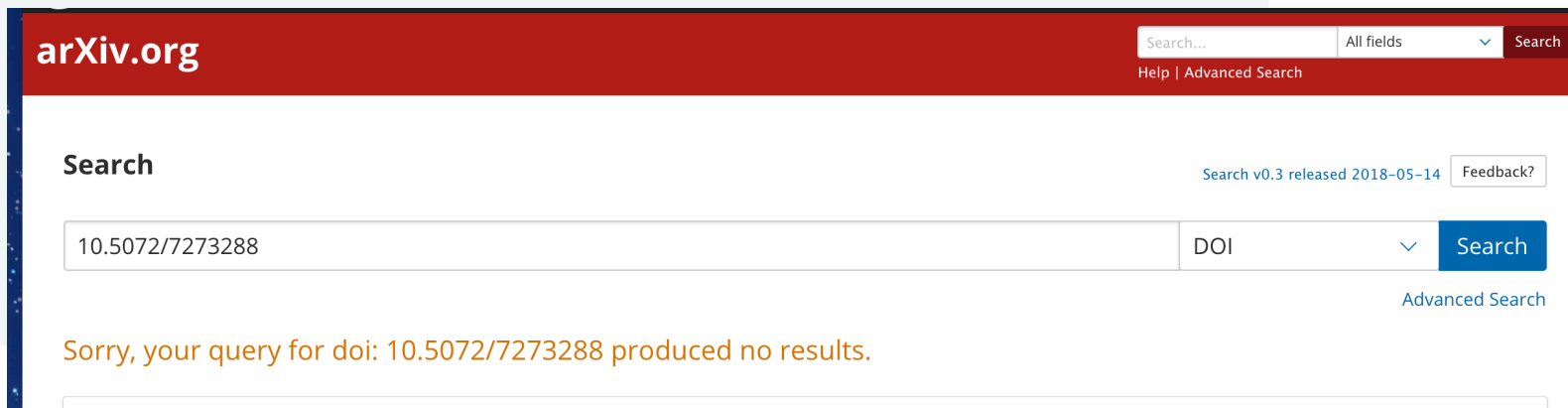The images from the X survey are copyright 2016, the X project.
They are published under the creative commons attribution 3.0 unported
license. If you use this data, please cite doi:10.5072/7273288.
Images are under embargo for one year after their addition to the repository.

`</rights>`

DOI in VO :   There is hope though!

GERMAN ASTROPHYSICAL
GAVO
VIRTUAL OBSERVATORY

Help
Service info

Metadata

Description
The services offered here a

Keywords
DOI

Creator
Demleitner, M.

Created
2016-01-18T12:20:00

Data updated
2018-05-18

Try ADQL to query our data.

Privacy | Disclaimer
Log in

Information on resource 'VO in DOI'

The services offered here allow the assignment of DOIs to VO resources; they also maintain the landing pages.

VOiDOI ("VO in DOI") lets you mint a Digital Object Identifier for a registered VO service.

Why would I want a DOI?

The main purpose of obtaining a DOI for a VO resource is to enhance its citability. Many popular citation styles let you give DOIs, which in turn facilitates later identification of the service, even if the service's URL changes, the service is operated by a different data center, or the service disappears.

The IVOID (the ivo-URI you get when registering with the VO) could, in principle, work almost as well for citation purposes, but few people know what to do with it, and few citation styles let you use IVOIDs meaningfully. Also, the IVOID changes when the publisher changes; a DOI can remain constant (although VOiDOI right now doesn't support that yet; ask us if this becomes an issue for you).

All DOIs are resolvable to "landing pages" (giving metadata and access options) when you prepend http://dx.doi.org/ to them; that is true for VOiDOI-generated DOIs as well.

Current practice in the VO is to preferably cite a paper associated with a resource if there is such a paper. Hence, VOiDOI is targeted mainly at services that either are not closely associated with a paper (e.g., observatory collections) or that collect data emerging from multiple papers (thing Simbad-like services).

How do I operate VOiDOI?

DOI in VO :   There is hope though!
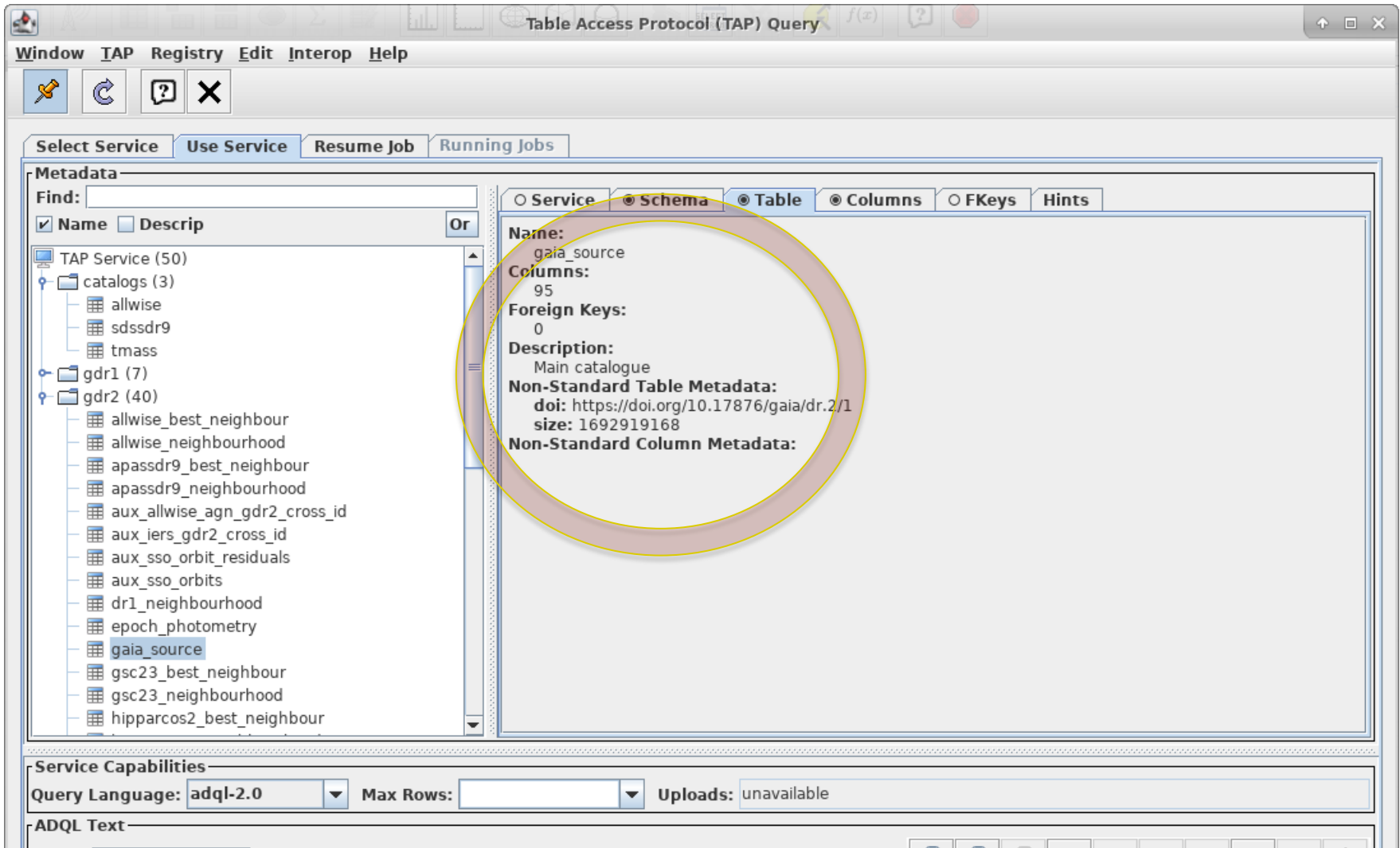
DOI : used by data providers?

Outside Astronomy : plenty (e.g. CERN) *http://opendata.cern.ch/docs/about*

How to re-use and cite these datasets

All datasets and other material available in this portal are minted with a persistent identifier, a so called DOI (Digital Object Identifier) that allows permanent linking to the records. The CERN Open Data Portal endorses the FORCE 11 Joint Declaration of Data Citation Principles. Thus, we ask you to cite the data provided in the portal when you re-use them. To make this easier for you, we provide you with a citation recommendation for every dataset as well as output formats (e.g. BibTex) for common reference programs. Citing datasets in the reference list of your paper will allow other platforms such as INSPIRE to track citations to these datasets and measure their impact.

In Astronomy : few

- ESA is working on minting DOI for their data collections

- ESO is working on minting DOI for their data collections

It's time to use DOI by all atronomy data providers (not only in speeches on RDA conferences)