International Virtual Observatory Alliance

This is only an outline! Everything is still to be done.

Best practices for the creation of and metadata for digital object identifiers in astronomy archives

Version 1.0

IVOA Note 2024-01-01

Interest Group
Data Curation and Preservation
This version
https://www.ivoa.net/documents/DOI4Archives-v2/20240101
Latest version
https://www.ivoa.net/documents/DOI4Archives-v2
Previous versions
This is a draft
Author(s)
August Muench, Gilles Landais, Raffaele D'ABrusco, Anne Raugh
Editor(s)
TBD

Abstract

Many astronomy archives are producing digital object identifiers (DOI) for datasets and services. This document aims to summarize current workflows for creating and using DOIs, diagnose issues in the metadata of extant DOIs, and develop best practices for worklows and metadata for future DOI deployment. This note is focused on archives in Astronomy, Planetary Science, and Heliophysics. Additional domains may be considered at a later time.

Status of this document

This is an IVOA Note expressing suggestions from and opinions of the authors. It is intended to share best practices, possible approaches, or other perspectives on interoperability with the Virtual Observatory. It should not be referenced or otherwise interpreted as a standard specification.

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

Known major missing pieces: Explicit DataCite metadata keywords; Guidance on DataCite deposit, including JSON markup or Fabrica.

1.1 General observations

this section was in 1.Introduction, items 1+2

We focus on roles of digital object identifiers (DOIs) rather than the full landscape of FAIR data (e.g., **?**). It is beyond the scope of this document to address all aspects of FAIRness. It is also unreasonable we think for DOIs to be laden with the problems of enabling all aspects of FAIRness. Instead we aim for this document to guide archives with achieve compliance for specific aspects of enabling FAIR data.

In this document we are using DataCite Commons [1]. This document adopts DataCite Metadata Schema 4.5. We choose relevant examples from

[1] https://commons.datacite.org

existing DOI taken in approximatively 75000 dataset DOIs in astronomy (See Table **??**).

1.2 DOI landscape

This section has been added by Gilles

It is now in the landscape that the publishing of scientific articles has included DOIs in its process. A good DOI curation allows to provide standard metadata that can be harvested and reused to generate citation with authors and any information required for a bibtex. It includes references and related products that build step by step a graph of linked resources.

DOIs are widely used for data, and although their metadata is slightly different from articles, they work very well together.

The DOI is a sustainable identifier applied to a web resource that emerge today among other identifier such as the ORCID[2], for authors, the RoR[3] for organization or the RAID[4] for projects.

The DOI success is a consequence of an architecture that includes URL resolution (since each DOI is attached to a landing page) and standardized metadata that provides exploitable information. As a unique and standard identifier it allows to make the bridge between different networks which complete each others (more details in **??**)

Data Centers motivated for citation and interconnection, must curate relevant metadata. In this document, we will develop a good practices to provide well citation, interconnections with related resources, including IVOA resources.

1.3 Use case

This section has been added by Gilles

1. Authors and Data Center expect that the resources they created or provide will be included in the citations of each article or resources which used them. It needs well curated Bibtex, linking the Data. Users expect also a visible Bibtex in the Data Center web pages.

2. The Data consumer expects to get information about origin and rights. For access rights, existing license must be visible but also readable by software.

3. Data published are disseminated in networks like the VO registry, DatasetSearch or the EOSC EU node (see **??**). The users as well as the Data Center expect that the entries are consistent.

[2] https://orcid.org/
[3] https://ror.org/
[4] https://raid.org/

4. The Data Center as well as authors needs reports about the usage activity of their work. They query dedicated services (eg: OpenCitation) that make reports and are capable to graph the citation of a specific resource.

5. a workflow using external resource would execute again a process. It expects that the experience is reproducible. If it uses a service or Knowledgebase, it expects that the protocols stay compatible with the original implementation. For Dataset, it could expects that data have not been altered.

1.4 Example of existing DOI

this section was in the original 1.Introduction

This list is not exhaustive; see Table XXX in Appendix A for a complete listing.

Archive	**Prefix**	**ID**	Yo	**Count**
Canadian Astronomy Data Centre	10.11570	nrc.cadc	-	80
Chandra Data Archive	10.25574	si.cda	-	29,449
ESO Science Archive Facility			-	
European Space Agency	10.5270		-	28,858
IPAC (26131 — 26135)	10.2613X	caltech.ipacdoi	-	622
Mikulski Archive for Space Telescopes	10.17909	stsci.mast	-	2,132
NASA Planetary Data System	10.17189	nasapds.nasapds	-	1,976
++Small Bodies Node of PDS	10.26007	sbn.archive	-	4,837
Strasbourg Astronomical Data Center	10.26093	inist.cds	-	17,383

Table 1

2 Role within the VO Architecture

This section has been modified by Gilles

2.1 Status of DOI in the Virtual Observatory

The DOI is not a standard of the Virtual Observatory which use the IVOID as identifier.

Note

Dataset DOIs do not lead to data, or what they do lead to is widely varied.

At the time of writing, the DOI is not well represented in VO standards. They are not part of the metadata of datamodels or protocols created long before the advent of DOIs. However, it starts to appear in Notes or in standards as an alternate identifier.

- the Data Origin (**?**) notes includes DOI among the metadata that a service can provide in the result of a query.
- the registry (**?**) (VOResource 1.2) allows to specify DOI in any ResourceName such as alternate identifier or related resources.
- the BibVO (**?**) is a note that requires DOI to link Data and articles.

add alternate identifier in ivoa and why

2.2 Comparison with the VO Registry schemas

Record in the VO registry has a unique identifier called IVOID. It is used in the Registry, includes a URL in its metadata, but is not really resolvable such as the DOI. However, its syntax which can be extended to a sub-part of the resource is commonly used in VO services. For instance, an IVOID on a tableset resource can be completed to specify a particular record and then to be used to generate a Datalink.

The VO Registry is based on VODataService and VOResource. It includes Resources; Datasets and Services that can be linked with term IsServedBy.

This relation which has no equivalent in DataCite schema allows to provide a strong interoperability. The absence of an equivalent term in DataCite makes it illusory to try and transcribe the same level of interoperability in DOI.

see term like isSourceOf, isCompiledBy that could be used to linkk a service/organization

We propose in section **??** a method to facilitate link between DOI and registry record.

VO Registry and DataCite schemas are based on DublinCore terms extented with their own specifications. Despite a common basis, the vocabularies differs. The DataOrigin note provides in Appendix a mapping between DataCite and VOResource for a sub selection of terms. Note that the endorsement process of DataOrigin has not been initiated.

3 Extant use of Digital Object Identifiers in Archives

This section comes from the original

We observe four different uses for DOIs in astronomy-related archives. Identifiers are being assigned for individual datasets (at various levels of granularity); for collections of datasets; for services; and for what we will call "knowledgebases" or curated metacollections of data and calculated results. These uses are described in more detail below.

3.1 Datasets

Dataset DOIs are the most common usage in astronomy-related repositories, and represent about »XX% of all such DOIs. Individual datasets are being assigned DOIs at various levels of granularity. At the most granular level, every observation in the Chandra Data Archive (CDA) is assigned an individual DOI. Similarly, the European Space Agency assigned DOIs to every individual Herschel, ISO, Planck and XMM-Newton observation. This list also includes every Hubble Space Telescope (HST) observation, which are also archived at the Mikulski Archive for Space Telescopes (MAST) but are not assigned DOIs individually by MAST.

Individual dataset DOIs are implemented mostly consistent with the usage of DOIs by institutional repositories, generalist repositories (e.g., Zenodo), etc. That is, the dataset DOI resolves to a landing page describing the dataset. The landing page provides one or more links to data. The content of the landing page and structure of those data links is not otherwise prescribed. In some cases the landing pages are rich summaries of the dataset, often including much more information than could be gleaned from the DOI metadata alone. In other cases the landing pages pop the user into a web user interface with very little (or almost no) contextual information about the DOI.

Gilles: I propose to remove the sentence

) Dataset DOIs are also being minted for "High-Level Science Products" (HLSP). Such HLSP datasets are not static and can grow over time, accumulating both data revisions and additions. HLSP landing pages for DOIs have no set structure or content and change regularly as the linked data evolve.

Gilles: what follows is interesting; but blurred - see section "Versioning and evolving datasets"

None of these dataset DOIs are versioned. The data discovered by following a dataset DOI to an individual observation will change based upon the version of the archive pipeline used to create it, up real-time or on-demand reprocessing for some archives.

The typical Resource type in DataCite is DataCite:**dataset**

3.2 Collections

Collection-type DOIs direct users to a collection of individual data records in a particular archive. Examples of collection-type DOIs include the DOI

services provided by the Mikulski Archive for Space Telescopes (MAST)[5] (**?**), Chandra Data Archive (CDA)[6] (**?**), and the VAMDC Consortium[7] Query Store service (**?**). There is significant variation in the expected use cases for these Collection DOIs. There are also strong variations in the metadata of Collection-type DOIs.

Considering use-cases, both MAST and Chandra Collection DOIs collect dataset identifiers within their respective databases. In both cases there is no expectation that these Collection DOIs would ever themselves collect attribution (i.e., be cited in the reference list of a corresponding Journal article). However, VAMDC Collection DOIs, are intended to collect and distribute credit to the collection of database identifiers they contain. They attempt to distribute this citation/credit to the collection of resources by "citing" all the related resources in the saved VAMDC query record. Because this depends upon the capabilities of both the chosen DOI minting service (Zenodo) and the DataCite Schema, additional notes on the details and the outcomes of this effort is provided in one of our case studies in Appendix B.

Gilles: this part is interesting: how report citation of individual datasets to the whole collection ?

Similarly the metadata of Collection DOIs vary between MAST and CDA examples. As described in **?**, Chandra Data Collection DOIs create complete records of individual using `relatedIdentifier` tags and predicates. The metadata of MAST Collection DOIs do not provide detailed information on the individual MAST records collated in the collection.

added by Gilles See **??**, **??**, **??**, **??**, **??**

The typical Resource type in DataCite is DataCite:**Collection**

Gilles: I guess, it means that DOI of the collection doesn't include ref to its individual datasets?

3.3 Services

Examples of DOIs for services include IRSA (DUST). Service DOIs lead to query tools. They may lead to query results – I would need examples.

Sometimes Collection DOIs act like Service DOIs but they are not. Collection DOIs may result from queries performed at a Service.

Warning: weeds.

added by Gilles See **??**, **??**, **??**, **??**

The typical Resource type in DataCite is DataCite:**Service**

Gilles: the problem with Services is the meaning of the term Service: is it a Service/organisation (eg VizieR) or is it a program exploit-

3.4 Knowledgebases

A knowledgebase is a collection of material collated from many discrete sources. All of the values contained in a knowledgebase have a provenance traced to other resources and have been curated into a single database for reuse. Examples include: Simbad (as originally described in, **?**), NASA

[5] https://archive.stsci.edu
[6] https://cxc.harvard.edu/cda/
[7] https://vamdc.org

Exoplanet Archive (NEA) (**?**) (as originally described in, **?**), NASA Extragalactic Database (NED) (**?**) (as originally described in, **?**).

Current observations about DOIs for knowledgebases include:

- DOIs for knowledgebases lead users to interstitial landing pages rather than directly to the collated, curated resources.
- DOIs for knowledgebases do not lead to individual values, e.g., the results of a query against that knowledgebase.
- DOIs for knowledgebases never provide information in their *metadata* about the *state* of a knowledgebase: its current version; last update; etc. Nor is this information on the interstitial landing pages of knowledgebases.
- Services that return DOIs for queries against knowledgebase are considered Service DOIs (See Section **??**).

Gilles: I propose to move the item in landingpage

Gilles: idem, is also in section landing-page, dates

added by Gilles See **??**, **??**, **??**, **??**

Warning: weeds.

The typical Resource type in DataCite is DataCite:**Dataset** or DataCite:**Service**

Gilles: I propose to remove this point

3.5 Pathologies

The primary pathology evident in DOI creation today is a mismatch between the metadata created by an archive and the use case (intended to be) implemented by that DOI. Succinctly, the metadata supplied by astronomy archives is often insufficient to ensure the accurate citation of the datasets.

By detailing these pathologies and triaging their less-than-desirable outcomes we can aim to develop empirically-defined best practices to guide repositories forward with the use of identifiers. Here is a topical list of pathologies:

1. Incomplete metadata: missing authors, generic or misstated titles, misunderstood dates;
2. Inconsistent metadata: transmutations of metadata between systems lead to inconsistent metadata deposits. Example: transformation from schema.org to Crossref left ESA DOIs in a nasty state;
3. (Un)versioned data: versioning is mostly nonexistent and when provided it is ill defined and often opaquely transmitted;
4. Misconceptions: DOIs do not lead to data, or what they do lead to is widely varied.

4 Best Practices for DOI Workflows

This section has been added by Gilles (from sections)

4.1 Pre-requisite

Before beginning a DOI workflow, the Data Center has to contact a DOI registration agency. Ten Registration agencies are providing DOIs, Crossref and DataCite are the best known in astronomy. Datacite offers the most appropriate metadata for recording data. We strongly recommend the use of DataCite, on which we focus on in this note.

To generate DOI in DataCite requires to be a DataCite member. Data publisher who is not a member would have to contact this surch of Organization (in general, countries and universities have a contact point).

4.1.1 Providing metadata using the appropriate schema

DOI workflow requires metadata curation based on the DataCite schema[8] which allows to improve the FAIRness of the datasets. Other workflows exist with their own semantic and metadata serialization. These are multiple, they overlap and are also specific.

For instance, the IVOA framework provides a registry with a high level of interoperability and point directly to the resources, whereas DataCite is specialized in data citation and link a human web page called "landing page".

> **Note** DOI implies to maintain a sustainable mechanism to provide a URL. This requirements, even if also in usage in the IVOA registry, is mandatory for DOI.

Datasets distributed in diverses frameworks complement each other and are likeky harvested by platforms or search engine such as ADS, EOSC, Google Dataset. These infrastructures adopt merge mechanisms (for instance OpenAire[9]), often a black box, that depends of their own strategy. Note that DOI, as unique identifier, facilitates the cross operations.

The list of metadata tends to increase, but the most popular are DataCite schema, Dublin Core[10], VOResource [11] (the registry of the Virtual Observatory), schema.org [12] (extends Dublin Core and is used by Google),

[8] https://datacite-metadata-schema.readthedocs.io/
[9] todefine
[10] https://www.dublincore.org/
[11] https://ivoa.net/documents/VOResource/
[12] https://schema.org/

DCAT[13] (linked catalogues, datasets and services. DCAT is a concurrent of the VO registry), OpenCitation[14] (a schema of linked citation), etc.

> **Note** All are specific, and we highlight the importance for Data providers to disseminate consistent metadata (for instance list the whole authors in all output).

In practice, it is better for implementers to think since the beginning about the different output in order to provide consistent workflow.

Note DataCite provides several serialization of the metadata, in particular "schema.org".

> **Note** Maintain the disseminations workflows togethers

See also the presention of H.Enke, Interop 2023, Bologna [15]

4.1.2 DOI syntax

Not sure it should be included in this note?..

The DOI syntax is composed with a prefix followed by a suffix. The prefix is assigned by DataCite or CrossRef and is used for all resources provided by the DOI producer. The suffix is created by the producer to identify the resource.

For instance:
https://doi.org/10.3847/1538-4365/aab76a

- the prefix 10.3847 is the prefix attributed by CrossRef to the AAS journals
- the suffix 1538-4365/aab76a defined the article

DataCite recommends using an opaque syntax. To remove any significance in a name avoids bad interpretation and is more sustainable. For example, the data center where the data are deposited may change and therefore should not be used in the syntax.

Note also the the DOI usage in web pages for which it is recommended to provide clickable links. It is therefore up to the web maintainer whether or not to hide the DOI with appropriate text.

4.2 Citation requirements

TODO : check?

[13] https://www.w3.org/TR/vocab-dcat-3/
[14] https://opencitations.net/
[15] https://wiki.ivoa.net/internal/IVOA/InterOpMay2023RegistryDCP/DOI-AIP-20230510.pdf

4.2.1 Generating Bibtex

DataCite schema includes the required metadata to generate citation in various formats: bibtex, APA, ADS, etc. The quality, generally fixed by journals, ADS or Data publisher, depends of the DOI curation (bibtex generated by ADS is more rich than those provided by DataCite). We will describe the mandatory items that cover the known citations and we will complete the metadata with added relevant information.

Note: DataCite provides different format that allows to check different citation output.

```
curl -LH "Accept: application/x-bibtex" https://doi.org/10.5270/esa-1ugzkg7
```

Full Bibtex template : the following example map bibtex with DataCite schema.

```
@dataset{{localref},
        author = {{authors},
        title = "{title}",
        year = {year},
        month = {month},
        eid = {usualName},
        url = {url},
        keywords = {keywords},
        publisher = {publisherName},
        copyright = {rights},
        DOI = {DOI}
}
```

Linking bibtex with DataCite schema:

Equivalent in APA style:

```
{authors} ({year}). {title} [Data set]. {publisher}. {DOI}
```

VizieR example: (catalogue J/MNRAS/320/451)

```
Beers, T. C., Rossi, S., O'Donoghue, D., Kilkenny, D., Stobie, R. S.,
Koen, C., & Wilhelm, R. (2006). A-G star metallicity [Data set].
Centre de Donnees Strasbourg (CDS). https://doi.org/10.26093/CDS/VIZIER.73200451
```

Bibtex	DataCite	Relevance
authors	Authors	MUST, see ??
title	Title	MUST, see ??
year	Date:creation	RECOMMENDED, see ??]
month	Date: creation	
eid	alternateIdentifier	MAY, see ??
url		SHOULD, use https://doi.org/{DOI}
keywords	Sujects	SHOULD, Please, see ??
publisher	Publisher	MUST
copyright	Rights	RECOMMENDED, see ??
DOI		MUST

Table 2: linking bibtex and DataCite schema

4.2.2 Evolving datasets

see DataCite evolving dataset guidance[16]

We distinguish different ways for Citing evolving datasets. The method depends on the data Center implementation.

1. Cite a snapshot of the Dataset. In this approach, the Data Center make snapshot an adopts a versioning mechanism.

2. Cite the dataset as to be an evolving Dataset. Example in APA style:

   ```
   {authors} ({year}). {title} [evolving Data set]. {publisher}.
   {DOI}. Accessed {date_of_access}
   ```

   In the example date_of_access could be the 'upodate' date of the DOI record (supposing that the dataset and its DOI recors are well synchronised).

3. Cite a sub part of a DataSet resulting of a query. The solution to these problems can become complex if we take into account the reproducibility. A simple solution which does not take reproducibility into account consists to cite the access protocol (for example scs, TAP), optionaly completed with query details.

   Example in APA style: (note that "'protocol"' is independent of the DOI record!

[16] https://datacite-metadata-schema.readthedocs.io/en/4.6/guidance/dynamic-datasets/

```
{authors} ({year}). {title} [evolving Data set]. {publisher}.
{DOI}. Accesses {date_of_access}, via {protocol}
```

A more advanced option, that take into account the reproducibility has been adopted by the VAMDC Query Store wher both: query and result are hosted with a DOI (giving details are not in the scope of this note).

Note see also the acknowledgment proposed in DataOrgin[17] Appendix Citation, Template.

4.2.3 Collections

a voir s'il faut reprendre la section Collection dans une section precedente

un des interets de la collection est de rassembles les statstiques de ses ressource individuels. Lorsque les ressources indivuduelles ont elle meme avoir un DOI, il faudra envisager une relation (voir section ...)

4.2.4 Service

We encourage Data publisher to provide a way to cite or acknowledge services used by authors in articles. The recommendation is in the Publisher discretion and should be visible in the landing page of the Service (see **??**).

It is rather a common practice to acknowledge than cite a service in an article. For both, the Service DOI which includes the URL resolver has to be consider.

4.3 Landing pages

Landing page is a human readable WEB page attached to the DOI (eg: ....) Datacite provides a documentation about good usage [18].

> **Note** The landing page is primarily dedicated for Human. The DOI metadata should be visible in the web page. In particular, the DOI, title, authors, licenses should be highlighted.

Additional information are often added. In particular links to access the data, but also any other information, specific to the data center or included in other workflows provided by the Data Center (bibtex, schema.org [19], DCAT [20], etc)

[17] https://www.ivoa.net/documents/DataOrigin/
[18] https://support.datacite.org/docs/landing-pages
[19] https://schema.org/
[20] https://www.w3.org/TR/vocab-dcat-3/

Landing page for service of Knowledgebase The landing page of a service or a Knowledgebase is generally its web portal. It includes both service access and information (eg: contact, about, etc.)

Landing page for collection This page contains information that is common to every datasets in the collection. Specificity of each dataset are not shown on this page. The content depends of the nature of the collection: for instance the Organization, an abstract, the date of creation or the materials used to generate the individuals datasets.

The page should provide the way to query individual datasets.

4.3.1 Generate a machine-readable landing page

Search engines harvest landing page for indexation and expect some metadata serialized with a standard semantic. For instance "schema.org" or DCAT serialized in JSON-LD are needed by Google Datasets [21].

The FAIRness of the landing page can be evaluated with validators:

- FAIRchecker [22]
- Google Search central [23]

4.3.2 Pathologies

- https://doi.org/10.17189/1519607 (NASA/PDS): there is a title, authors, DOI, summary... but there is no direct link to the data. The "Search/Access Data" link goes to a search interface, where the user should search again with some parameter (not specified, to be guessed)
- https://doi.org/10.48322/rgf7-3h67 (HPDE): there is everything about the citation, title, doi, summary, content... but the formatting is very close to the SPASE XML record. It's ok for user fluent with SPASE, not so much for outsiders.
- https://doi.org/10.57780/esa-3xcjd4w (ESA): All required/recommended information is present, but the granuality (a single DOI for the full experience, not split with version ro processing levels...). And the data access goes to a search interface, where the user should search again for the products.

[21] https://datasetsearch.research.google.com/
[22] https://fair-checker.france-bioinformatique.fr/
[23] https://developers.google.com/search/docs/appearance/structured-data

4.4 Linked Data

4.4.1 Connect resources together

Linked Data allows to connect resources together, it helps discovery and allows to generate a map of interconnected resources. Linked Data are expressed with DataCite:**RelatedIdentifier** in DataCite schema. They have been the subject of a supplementary section in the DataCite guidance.

We recommend to add at least one relation (for instance to link a dataset with a reference article).

DataCite provides a list of controlled vocabulary that allows to specify the relationship between resources. These relations may use DOI or any other sustainable identifier (eg: bibcode).

> **Note** The relatsionship vocabulary is precise. Too often, providers give their own interpretation of the terms. If there's any doubt about the meaning of the vocabulary, it's better to put no relations than a misunderstood relationship.

See relation vocabulary and their definitions in the DataCite Schema[24].

> **Note** Privilege DOI linking to any relations

DOI has to be privileged than any other identifier or URI. DOI facilitate the harvesting/merging jobs. For instance the EOSC portal selects the relations type to display with and privileges resources having DOI.

4.4.2 Link resources within a DataCenter

DataCenter may provides resources in different granularity. For instance Chandra provides catalogue of observations. Each observation is a Product that has his own DOI. Then the products are linked to the catalogue which has a DOI too.

Particular interest for Collections In a case of Collection, DataCite provides terms to link a resource to its collection. DataCite:**isPartOf** seems to be appropriate to link an individual Dataset to a Collection, however, the verb used for linking depends of the Data Center architecture.

[24] https://datacite-metadata-schema.readthedocs.io/_/downloads/en/4.5/pdf/, section 12

This section has been added moved from "best Practices for DOI workflows" by Gilles

Include related identifiers linking the collection DOI to internal and if possible external identifiers for the dataset(s). This will make the DOIs broadly more useful and interoperable. Other systems will be able to use the alternate identifiers to correctly link their records to yours.

provide example or a link

4.4.3 Link data from outside

There are many reasons to link a resource with external products, we give just few examples:

- link the datasets that have been used to create a product
- link other copy DataCite:**isIdenticalTo**
- for a derived product, link the original data with a qualifed terms such as DataCite:**isVariantFormOf**, DataCite:**isDerivedFrom**, ...
- cite a resource DataCite:**Cites**
- etc.

add Example

4.4.4 Linking Dataset to a publication

DataCites provides terms to link Datasets with articles. For instance DataCite:**Cites**, DataCite:**isSupplementTo**, DataCite:**isDescriebdedBy**, DataCite:**References**, etc. Choosing the good term is specfific to the Datasets.

Note the DataCite:**Cites** relations which makes an automated citation (the article is cited when the DOI of the dataset is created). Other terms like DataCite:**References** generates also statistics in DataCite.

Query DataCite statistics , see "referenceCount", "citationCount"

```
curl "https://api.datacite.org/dois/{DOI}"
```

Query Crossref statistics

```
curl "https://doi.crossref.org/search/doi?doi={DOI}&format=info&pid=mail@address
```

Example: link a dataset with its reference article (vizier DOI extract from https://doi.org/10.26093/cds/vizier.22640008).
The above example provides a relations using a bibcode.

```
  {
    "relationType": "IsSupplementTo",
    "relatedIdentifier": "2023ApJS..264....8H",
    "relatedIdentifierType": "bibcode"
  }
]
```

4.4.5 Linking Dataset with its original resource

We distinguish a mirror copy and resource derived from an orginal resource. Both can have their own DOI. We encourage derived resource as wel as copy mirror to link the original resource using DataCite relations.

There are lots of reasons why a DataCenter that provides a derived resource want to provide a DOI. Often, the derived product is the result of a curation that provides added values.

Before creating a DOI, we encourage the DataCenter to tell their DOI plan to the original archive. Some metadata of the original resources should be copied in the derived product, and other will be dedicated to the derived product.

- Authors: the full authors (creators and contributors) list should be replicated in the derived metadata. The list can be completed with contributors involved in the derived product curation.
- Dates: the creation and the modification dates are the dates of the DataCenter that provide the derived product
- Abstract: it is not needed to duplicate the original abstract, but the added curation should be explained
- relations: add link to the original product. You can Cite (relation type DataCite:**Cites**) the original resource that you can complete with a relation explaining the relationship between the original and the derived product.

Please : refer to DataCite:**relationType** vocabulary and do not extrapolate the semantic.

give Example

```
  {
    "relationType": "IsVariantFormOf",
    "relatedIdentifier": "10.5270/esa-qa4lep3",
    "relatedIdentifierType": "DOI"
  },
  {
    "relationType": "Cites",
    "relatedIdentifier": "10.5270/esa-qa4lep3",
    "relatedIdentifierType": "DOI"
  }
]
```

4.4.6 Link other networks

Link an other network (see **??**) is not a capability provided by DataCite. However, networks using a specific identifier can be exploited to link the DOI record.
DataCite:**alternateIdentifier** is a term to specify an other identifier of the same resource. For instance the Usual Name used by the Data Center or the IVOID of the Virtual Observatory.

> **Note** When possible, we suggest to add the IVOID of the resource in DOI using DataCite:**alternateIdentifier** and to add the DOI in the VO Registry records using altIdentifier.

4.4.7 Linked Data for Services and knowledgebase

It is again the role of the Data Center to choose relevant links. However, we suggest to limit links with a reference article.

In the other sens, we haven't found any equivalent in DataCite of the servedBy term of the VO registry to link datasets with services.

the last paragraph is dedicated for service only and not knowledgebase: what about isSourceOf or isCom-

4.5 Versioning and evolving datasets

Distinguish dataset versioning (the Data Center responsibility) and metadata versioning (automated versioning done by DataCite at each metadata update)

A good practice consists to follow a versioning mechanism for each data update. DataCite (version 4.5) recommends to create a new DOI for each major version and to stipulate the version number with DataCite:**Version**. It is recommended to adopt a version number folowing the pattern:

```
major_version.minor_version[.patch_version]
```

Different mechanisms exist:

- Zenodo [25] method: makes a DOI collection of versions. Each version has its own DOI which are linked (related identifier)with the DOI collection.
- Make 1 DOI for version, link the DOI version together using related identifier DataCite:**isNewVersionOf**, DataCite:**isPreviousVersionOf**

4.5.1 Evolving datasets

Versioning is well adapted for data subject to planified update, such as survey releases. Versioning implying a DOI per version is preferable for reproducibility, However, there are datasets that evolves regularly and for which versioning is inapropriate. For instance logs of observations evolves regularly.

For those type of datasets, we suggest to add in the DOI DataCite:**Title** or in DataCite:**Description** the evolving nature of the datasets and to modify the "Update" date at each modification.

We recommend also to specify the evolving nature in the DataCite:**ResourceType** (see examples in **??**).

4.6 Tombstone page

DOI requires a mechanism that provide a sustainable URL. The URL itself can change (eg: the domain used by the data center evolved). In an article, a citation with a DOI allows to link the landing page. The information as well as the datasets access are relevant for the article content. So, in principle, the datasets having DOI should be sustainable.

If for any unfortunate reason, the datasets is no more available, provide a *tombstone page* explaining the reasons and a link to a copy of the archive if it exists.

Note Tombstone page must not be used for versioning.

[25] https://zenodo.org/

5 Metadata list

This section has been added by Gilles (from sections)

We are focusing on a sub-list of particularly interesting terms that should be given a particular attention in curation.

5.1 Choosing a type for resources

The Resource type is specified with DataCite:**ResourceTypeGeneral** which use a controlled vocabulary, completed with a DataCite:**ResourceType** which is a free text.

Examples: Choosing the good Resource type is leaded to the Data publisher. the following list is indicative only and does not constitute a rule of good practice:

	ResourceTypeGeneral	**ResourceType**
table	Dataset	Dataset
spectrum	Dataset	Spectrum
image	Image	Image
logs of observations	Dataset	Evolving Dataset
Collection	Collection	Collection
Service VO	Service	IVOA Service
service Web	Service	Web Service
knowledgebase	Dataset	Evolving Dataset

Table 3

Example:

```
        "resourceType": "evolving Dataset",
        "resourceTypeGeneral": "Dataset"
},
"titles": [
{
        "lang": "en",
        "title": "The Gemini Observation Log (evolving Dataset)"
}
]
```

5.2 Authors

> **Note** for datasets. The good usage consists to add all authors and not only the first author.

All authors are indeed asked by journals for citations. We encourage also to add the authors ORCID with their affiliations. When possible inform the RoR.

Example of a unique author.


```
[
  {
    "name": "Ochsenbein, Francois",
    "nameType": "Personal",
    "givenName": "Francois",
    "familyName": "Ochsenbein",
    "affiliation": [
      {
        "name": "Observatoire astronomique de Strasbourg",
        "schemeUri": "https://ror.org",
        "affiliationIdentifier": "https://ror.org/04xsj2p07",
        "affiliationIdentifierScheme": "ROR"
      }
    ],
    "nameIdentifiers": [
      {
        "schemeUri": "https://orcid.org/",
        "nameIdentifier": "0000-0003-4667-015X",
        "nameIdentifierScheme": "ORCID"
      }
    ]
  }
]
```


5.2.1 Creators and contributors

DataCite distinguishes Creators and Contributors. The authors are conceptually the same that the Datacite term Creators.

The DataCite:**creator** can be a DataCite:**Person** or an DataCite:**Organization**. Contributors are persons or organizations that contributed to the development of the resources (sic DataCite Schema 4.5). A contributor has a role (eg: Editor, Supervisor, etc.) taken in a controlled vocabulary (see DataCite:**contributorType** term and the complete list in DataCite Schema).

In a way DataCite:**Contributors** is an alternative of DataCite:**Creators**. For instance, Zenodo provides a way to add a role to creators in its upload form. However, DOI created by Zeonodo used only Creators term.

The usage of Contributor in citation has not been evaluated today.

5.2.2 Authors for Collection, Services, Knowledgebases

Staaf implied in this type of resource evolves. We recommend to specify the Organization (and the RoR if it exists) rather than contributors. Adding human (for instance the main Creator of the service) in the leaves to the Data publisher discretion.

5.3 Keywords

We encourage the usage of recognized keywords such as UAT[26], IVOA-UAT[27] or any keywords driven with a Web semantic.

The term in the DataCite schema is DataCite:**subject**.

Example:

```
[
  {
    "subject": "Sky surveys",
    "valueUri": "https://astrothesaurus.org/uat/1464",
    "schemeUri": "http://astrothesaurus.org",
    "subjectScheme": "UAT"
  },
  {
    "subject": "Earth (planet)",
    "valueUri": "https://astrothesaurus.org/uat/439",
    "schemeUri": "http://astrothesaurus.org",
    "subjectScheme": "UAT"
  }
]
```

[26] https://astrothesaurus.org/
[27] https://www.ivoa.net/documents/uat-as-upstream/

5.4 Dates

DataCite provides a list of terms to qualify dates such as DataCite:**Created**, DataCite:**Updated**, DataCite:**Validated**, etc.

The dates involve the publication in the data center only. For example, a dataset provided by a Data Center A may have been created months or even years ago and already published in a reference article or another Data Center B. In any case, the date of DataCite:**Creation** is the date of creation of A.

Note: dates have not to be used as versioning.

Dates for Services, Knowledgebase and Collection By nature, these type of resources may evolve. The frequency with which content is updated may make inappropriate to use the update date. For changes in architecture or data model, versions should be used.

5.5 Title and description

Titles is an important metadata, it describes the resource and is often exploited by search engine (such as ADS or VO-registry) in a text search process.

Assigning a title is specific to each dataset. It is a short and unique sentence that contains the most relevant aspects of the dataset. For datasets derived or attached to a reference article, it is better to create a new description that describes the dataset.

Example : Reference article: ApJ (Draper Z.H, 2000), [28]
"Disk-loss and disk-renewal phases in classical Be stars. II. Contrasting with stable and variable disks"

Dataset Title : VizieR, https://doi.org/10.26093/cds/vizier.17860120
"Spectropolarimetric survey of classical Be stars"

Making a good title is of course specific to the dataset. For instance, we can name the object or type studied, the facility used, the release version, or the measurement methods if it is spectropscopy or photometry, etc.

Description completes the title. Generaly, the description is not exploited by search process. However, like title, description describes the data and not its reference, even when the data comes from an original resource.

Datacite provides a description qualifier (Abstract, Methods, ...) which depends of the resource.

[28] https://doi.org/10.1088/0004-637X/786/2/120

Description for data derived from external resource When the data derived from an external resource is not a simple copy, it is recommended to adapt the original description. For example, in the case of data attached to a published article, rather than reproducing the abstract of the reference article (which may also be subject to the same license as the article), the description may focused on the dataset content with added information such as those useful to produce the datasets.

An other good practice consists to add in the description a reference to the data origin.

Example : VizieR DOI description example -

"VizieR online Data Catalogue associated with article published in journal Monthly Notices of the Royal Astronomical Society with title ...

5.6 Licenses

License is one of the metadata required by FAIR principles[29]

Institutes as well as countries encourage the usage of licenses for the datasets. For instance:

- NASA encourages CC0 licence (see SPD-41)
- French government imposes LO/OL or CC-by licenses (see Etalab)

Note Use machine readable licences

It is hightly recommended to use machine-readable licences using the official term or the URL link:

- SPDX licenses https://spdx.org/licenses/
- Creative Common https://creativecommons.org

Datacite metadata (see "Rights" in the DOI metadata schema)

```
"rightsList": [
  {
    "rights": "Creative Commons Attribution 4.0 International",
    "rightsUri": "https://spdx.org/licenses/CC-BY-4.0.html",
    "schemeUri": "https://spdx.org/licenses/",
    "rightsIdentifier": "CC-BY-4.0",
    "rightsIdentifierScheme": "SPDX"
  }
]
```

[29] https://www.go-fair.org/fair-principles/

5.7 Identifier for project

This section has been added by Gilles (not in sections) At the time of writing, RAID (Research Activity Identifier) is new in the open DATA scene. We do not have feedbacks to provide any best practice. However, an agreement between the RAID maintenance organisation (Australian Research Data Commons[30]) and DataCite makes possible to generate such identifiers from DOIs (using DataCite:**resourceTypeGeneral** "Project")

6 Recommendation checklist

This section has been added by Gilles (from sections)

1. Select the type of your Resource and adopt usage specific to the chosen category.
2. **Landing page** (see **??**)
   - Check if DOI metadata are displayed in the landing page
   - Complete the landing page with added information such as links to datasets and any other metadata used in other dissemination workflows
   - Check if the landing-page is machine-readable
3. **Provide curated metadata**, in particular:
   - List all authors with ORCID and affiliation when they are known (see **??**)
   - Add machine readable licence (see **??**)
   - Add the publication date in the Data Center (see **??**)
   - Choose title matching with the Content ans the type of the resource (see **??**)
4. Check if **metadata allows to generate citations** (see **??**)
5. **Link your Datasets** with resources (see **??**)
   - Do not misinterpret the meaning of a semantic. The relations semantic is precise, in case of doubt it is better to put nothing. The IVOA DCP working group can also help you[31]
   - Link the DOI with resources which are used to generate the dataset. For instance datasets attached to an article, or when they come from an original archive.

[30] AustralianResearchDataCommons
[31] mailto:datacp@ivoa.net

- add Alternate identifiers. In particular add IVOID if it exists (see **??**)

6. Ensure the maintenance of the DOI
   Ensure that all your dissemination workflows are consistent, in particular, with IVOA registry and with "schema.org" (see **??**)

7. Use a **versioning mechanism** each time your datasets evolved
   distinguish dataset versioning (the Data Center responsability) and metadata versioning (automated versioning done by DataCite at each update) (see **??**)

8. Ensure the sustainability of the landing pages.
   If for any unfortunate reason, the datasets is no more available, provide a Tombstone page explaining the reasons (see **??**)

7 Appendixes