GROUPE INTERPÔLE

GUIDE

Good practices for data citation

v1.0

SOMMAIRE

EDITORS

Initials	Name	Organization / Pole
EDO	Emilie Deschamps-Ostanciaux	Form@ter
SG	Sylvie Galle	Théia
FA	François André	Aeris
SH	Steven Hosford	Cnes
FM	Frédéric Merceur	Odatis

HISTORY OF REVISIONS OF THE DOCUMENT

Date	Modifications	Author
08/06/16	Document creation	FA
10/06/16	Addition to the initial document / Integration of discussion elements 08/06/16	FA, EDO, SG, SH, FM
29/05/17	Taking into account remarks on the previous document. Especially: - Added FAQ - Added paragraph on advantages / disadvantages of significant fragments	FA
31/05/17	Finalization at the Interpoles workshop	FA,SG,EDO
01/06/17	Passage 0.2.5 -> 1.0	FA

REFERENCE DOCUMENTS

Document name	Reference
CEOS Persistent Identifier Best Practices_v1.0.pdf http://www.aeris-data.fr/rest/redirect/doi/CEOS Persistent Identifier Best Practices_v1.0.pdf	[CEOS]
OGC OpenSearch Geo and Time Extensions http:// portal.opengeospatial.org/files/?artifact_id=56866	[OSGEOT]
ANDS DOI FAQ http://www.ands.org.au/working-with-data/citation-and-identifiers/data-citation/doi-faqs	[ANDSFAQ]
ODATIS - DOI de données - FAQ http://...	[ODATISFAQ]

1. DOCUMENT BOUNDARIES

1.1. OBJECTIVE

The objective of this document is to:

1. recall certain definitions around DOI
2. define a number of good practices around the quotation of digital objects. Once these practices have been defined and validated,
   - a gap analysis could be carried out within the different pôles or data centers
   - the implementation of each of them within the pôles could be monitored
3. answer some common questions through a FAQ.

1.2. WHAT ARE DIGITAL OBJECTS ?

DOIs are, literally, identifiers of digital objects. What are these digital objects? They can be of various natures. The most frequent cases are:

- A data set (or a data collection)
- An algorithm
- A website (of a project, a program or a mission)
- An article online
- ...

In this document, **only data sets and collections will be discussed**.

1.3. WHO IS THIS DOCUMENT FOR ?

The production of DOI is related to the cataloging and long-term conservation of data. Thus, a large part of the recommendations is addressed to data managers (in observation services, observatories or data centers).

However, some of the elements contained in this document, which provide a neutral view of the DOI, may be of interest to both the data producer and the representatives of the supervisory entities.

2. DATA SET CITATION

2.1. OBJECTIVES ET CONSTRAINTS

2.1.1. Objectives

The primary purpose of data quotation is to allow **concise, accurate and reliable referencing over time of data**. This quotation can be carried out for example in a scientific article but also on a website or in the body of an electronic message.

A second objective is to make it possible to establish, via the citation mechanism, an estimate of the importance of data sets in the community through the establishment **of bibliometric measurements**. Such a need can also be extended to the project / program / mission.

2.1.2. Constraints

Several constraints must be taken into account in order to achieve these objectives :

- **Non-sustainability of URLs**
  Unlike scientific papers that have a fairly long life, URLs are usually only valid for a few years. For example, a URL may become invalid as a result of a domain name change.
  So using URLs is not a suitable solution for quoting and setting up a redirection mechanism in which the name of the redirection is fixed and the redirect address is variable.

- **Lifecycle of the dataset and the related metadata**
  A dataset evolves after it has been cited: the data can be corrected, added, the format modified ... Similarly the list of authors of the dataset can also evolve in time In the case of a scalable data set. Thus, the citation mechanism must be able to provide flexible means for adapting to these different situations.

- **Conciseness**
  The citation must be able to appear in a document as a printed article to which strong space constraints are attached. So the quotation must remain short and not include elements that could affect its use such as printing on several lines because of the length, the presence of spaces or non-alphabetical characters.

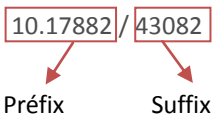
2.2. DOI

2.2.1. Structure

The DOI (Digital Object Identifier) is a widely used technology to allow the referencing of digital objects. A DOI is composed of the following elements:

Element	Role
Prefix	String identifying the entity that produced the DOI
Suffix	String identifying a dataset within the entity indicated in the prefix.
Metadata	The metadata necessary for understanding the dataset must be provided at the time of DOI production. Some of these metadata are mandatory:: ▪ Creator ▪ Title ▪ Publisher ▪ Publication year
Url of landing page	The landing page (LP) is a web page that will provide more information about the dataset and retrieve it (subject to possible restrictions).

Example DOI :



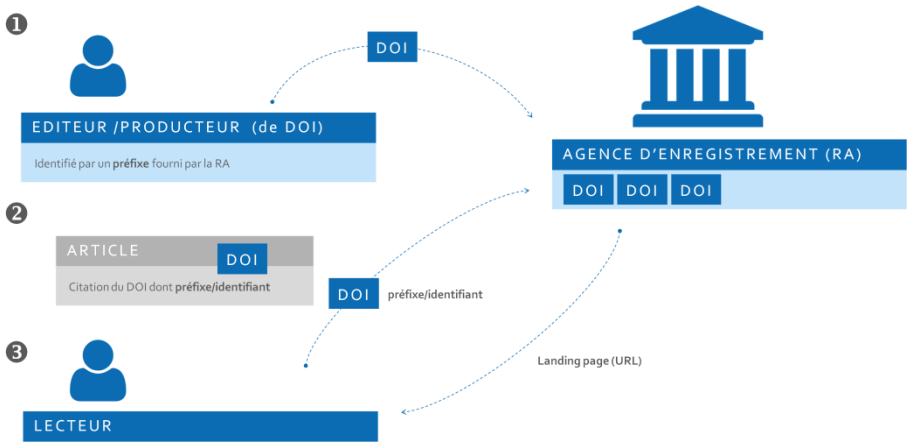
2.2.2. How does it work

DOIs are registered in a specialized agency (Registration Agency or R.A.). These agencies propose resolution mechanisms that allow a user to redirect to the landing page of a DOI.

Thus the 3 steps of using a DOI are :

1. Creation and registration in an R.A.
2. Quotation in an article (paper or other)

3. Resolution - via the R.A. - to find the landing page.



Once created, DOI information - except the suffix - can be modified (see above), including the URL of its landing page.

2.2.1. The notion of fragment identifiers (or fragments)

The notion of fragment identifiers is a standard, among others, by the W3C (https://www.w3.org/TR/media-frags/) aimed at pointing to a sub-part of a resource. To do this, the main url is concatenated as a string of complementary characters beginning with the sign #. In the remainder of this document, both *fragments* and *fragments indicator* are used to indicate fragment identifiers.

This standard is supported by DOI recommendations (https://www.doi.org/doi_handbook/5_Applications.html#5.8) when resolving the DOI if the fragment is transmitted to the LP.

So if we have the DOI *10.17882 / 43082* having as landing page *http: //mySite/myLP* the resolution of DOI *10.17882 / 43082 # myFragment* will be *http: //mySite/myLP # myFragment*.

It is important to note that using fragments only one DOI is published and updated.

2.2.2. Datacite

The international organization specialized in recording DOI for datasets is DataCite, with correspondents on national level . In France DataCite is respresented by INIST.

Each scientific institution can subscribe a contract at the INIST for 180 € / year which allows :

- Obtain one or more prefix(s).
- Issue an infinite number of DOI.

2.2.3. Alternatives

A number of alternatives to the DOI exist. Examples include the ARKs, Handle or PURLS mechanisms which also pass through the mediation of a recording agency.

2.2.4. Limits

One of the strong commitments when producing a DOI is its sustainability of the relevance of its landing page. Thus, a DOI can not be deleted (see FAQ).
In this context, the use of a DOI to provide a reliable and modifiable link to non-subtainable documents is a misuse of the main objective which will lead to obtaining a large number of empty LPs. In this case, it is better to use URL redirection services that are offered by certain websites or that can be set up locally.

2.3. UUID

The creation of a universally unique identifier (UUID) is a fairly simple operation and relies on the production of sufficiently long random strings (typically 16 to 32 hexadecimal characters). For reading purposes these strings are often separated by dashes.

A certain number of standards have been specified such as ISO / IEC 9834-8: 2008.

The advantage of these mechanisms is that they allow easy identifier generation, without having to worry about a local context. Moreover, the uniqueness of this identifier will be retained in another context. For example, one can think of the identifier of an author which must be able to be used in several publications.

By its triplet protocol (DOI) / prefix / suffix, the DOI also constitutes a UUID whose use is dedicated to the diffusion and the publication.

3. LIST OF GOOD PRACTICES

This section proposes a number of good practices around setting up an DOI. For each practice a reminder of the problematic is provided as well as some elements of explanation of the proposed choices. In addition, a recommendation level is also indicated.

3.1. GOOD PRACTICES RELATED TO THE SUFFIX

3.1.1. Neutrality of the suffix [Recommended]

The official recommendations do not indicate any constraints on the formalism of the suffix. Some DOIs do include references to the name of a project, an observation site etc. ... However, these elements do not necessarily have a significant durability and can be in contradiction with the DOI.

DOI-SUFF-01 [Recommended]: The suffix must not be meaningful..

This recommendation is very frequently repeated in the various recommendation documents such as [CEOS] or [ANDSFAQ].

3.1.2. Concision of the suffix

A number of identifier mechanisms provide identifiers based on 16- to 32-character UUIDs. For example OrcId (orcid.org) can provide identifiers.
In the case of the DOI type mechanism, the use of such a mechanism can be both counterproductive by producing DOIs that are unnecessarily too long (see above), since the prefix already guarantees a universality.

Thus the suffix must be as short as possible in the context of its prefix. Thus it is possible to consider simply increasing suffixes (1,2,3 ..).

DOI-SUFF-02 [Recommended]: The universality of the DOI must be expressed by both the prefix and the suffix (and not by the suffix alone)

This recommendation includes REC_03 in [CEOS]. This recommendation can also be compared to class naming in java in which part of the package name corresponds to the organization performing the development in order to guarantee the universal uniqueness of the class name.

3.1.3. Using fragments to indicate variants of a dataset

A dataset can exist in the form of several variants:

- As a subset of the larger dataset (temporal restriction, geographic restriction, ...)[1].
- As a version of a data set (in case of reprocessing)
- As a snapshot of a dataset at a specific time (for an article for example)

A large number of DOI issues concern how to set up a DOI mechanism to answer these different questions. A response is provided by the DOI standard, which makes it possible to establish links between DOIs (e.g.*: isPartOf, isPreviousVersionOf*, ...). These solutions seem complex to set up in a homogeneous way and to maintain.

Instead, we recommend the use of fragments (see above) to designate all the variants associated with a single DOI. With the use of fragments, the LP can take advantage of this complementary information to display more relevant services such as a more precise download link. This way the static LP becomes dynamic. **Thus, the fragment adds information about the data**.

a) Advantages

- With fragments, only one DOI must be set up to describe all quotations of a dataset.
- Fragments provide an initial answer to the question of fine grained quotation of data (see RDA recommendations: https://rd-alliance.org/system/files/documents/RDA-DC-Recommendations_151020.pdf).).

b) Disadvantages

The use of fragments adds a new meaningful element. **It will therefore be necessary to maintain the meaning provided by a fragment,** when the production of DOI was minted using these fragments:

- If the landing page is hosted on another system, it must support the interpretation of the fragments already issued.
- In case of evolution of the mechanism of the fragments (for example of the name of the parameters), a downward compatibility with the fragments already issued.
- Finally, the LP must notify the incomprehension of a transmitted fragment.

DOI-SUFF-03 [Recommended]: the variants of a dataset must be indicated via fragments and must not result in the production of new DOIs

3.1.1. Neutrality and concision of the fragment

It is important that the presence of the fragment does not contravene the above-mentioned requirements, in particular that of concision.

Thus, the content of the fragment must be an identifier having the smallest possible size. This implies that the actual meaning of the fragment must be stored in the information system producing it and be accessible by the LP. Except trivial cases, this corresponds to the notion of query-store defined in the recommendations RDA. The implementation of this mechanism implies, as indicated above, maintenance of the content of the requests.

DOI-SUFF-04 [Recommended]: Fragments must be concise and not meaningful. Their real meaning must be kept and maintained within query stores.

Note:
The non-significance of the fragments entails the obligation to go through the data portal to obtain the fragment. This may seem to be antagonistic with the current development of REST data extraction interfaces. Indeed in such mechanisms the extraction parameters are passed to the HTTP request via a succession of keys / values attached to the url. Thus, we can have a request of the type:

[1] More generally, this case corresponds to the data obtained by application, at a given moment, of a query composed of different search criteria (time, geographical zone, physical parameter, program name, ...) on one or more datasets.

http://my.extraction. service/datasetIdentifier?param1=value1&param2&value2...

With such mechanisms the user can freely act on the various parameters

Using the same syntax for fragments of DOI - prefix: suffix # param1 = value1 & param2 & value2 ... - brings the same freedom to the user who can modify / adapt the quotation himself.

If the number of parameters remains small, the citation can be considered to be sufficiently concise. In this case the use of explicit fragments has the following advantages:

- No need to set up a query-store.
- Homogeneity between the REST interface and the citation
- Freedom for the user.

However extraction has a temporary aspect whereas the DOI has a perennial vocation. The disadvantages are as follows:

- Bottom-up compatibility is more complex to implement because it is difficult to know which parameters and values are used by users, since the analysis of a query-store is not possible. Thus, it must be carried out systematically on all parameters / values.
- To prevent this maintenance task from becoming too cumbersome, care should be taken to ensure that the names of keys and values are appropriately chosen (for example, using names from dictionaries adopted by the community)
- Don't forget that one of the parameters must be the date of the request.

3.2. GOOD PRACTICES RELATED TO THE LANDING PAGE

3.2.1. Displaying quotation text

The landing page should recall the quote text in order to allow its quick incorporation into an article. The structure of the citation proposed by DataCite is as follows:

Creator PublicationYear: Title. Publisher. Identifier

Notes:

- Several styles can be applied to this structure (APA for example). Datacite offers a formatting service: http://crosscite.org/citeproc/
- The quotation text must take into account any fragments passed in the URL of the LP.

DOI-LP-01 [Recommended]: Displaying the citation text on the LP

3.2.2. Interpretation of the fragment

The LP must be able to transcribe a possible fragment in a form that the user can understand. It should be able to mention cases such as:

- a non-comprehensible fragment
- if data is no longer available in the version when the fragment was generated.
- ...

DOI-LP-02 [Recommended]: Interpretation of a possible fragment by the LP

3.2.3. A metadata sheet as landing page

In the case of a DOI for a dataset, it is appropriate to make the landing page a record of a metadata catalog that displays more complete information than the metadata of the DOI. This allows the implementation of DOI-SUFF-01, the information previously carried by the suffix appearing in the metadata sheet.

In the remainder of the document, we call MLP (metadata landing page) such a metadata sheet.

The MLP must contain at least the following information:

- Title
- Abstract
- Contact person for the data
- Link to the data
- Links on data formats (tools, ...)

The link to access the data can take several forms more or less precisely, in order of preference:

- Download link: link to retrieve the dataset or a subset of it. The person consulting the MLP is greatly guided.
- Search link: link to access a search / selection / retrieval interface. The person consulting the MLP must perform the manual extraction.
- Command link: link to a data order form. The extraction is carried out by an operator of the data center and then made available to the applicant

The following information is also strongly recommended:

- Temporal coverage
- Spatial coverage
- Data genealogy (links to other datasets, history of the different versions of the dataset)
- Conservation means implemented around the data
- Available extraction formats

These problems of the landing page content are in fact **problems related to the implementation of a catalog** (**granularity**, content, link between the pages).

DOI-LP-03 : The landing page of a DOI in a dataset must be a structured metadata sheet

Note: It may be useful to standardize the internal identifier of the MLP and the DOI. Under DOI-SUFF-02, candidates for this identifier could, for example, be of the form *prefix_suffix*.

3.2.4. Link MLP / DOI

All DOI metadata must be retrieved in an automated way from the MLP information.

This is valid at the time of publication of the DOI but also when **one of the metadata is modified** (example: adding a new author).

Note: Even if it is automatable, the production of a DOI does not have to be automatic.

DOI-LP-04 : All DOI metadata must be retrieved automatically from the metadata sheet used as the landing page. DOI metadata must be automatically updated if the content of the MLP is updated.

3.3. GOOD PRACTICES RELATED TO THE DATA DISTRIBUTOR

3.3.1. Presence of quotation information in data files

DOI-LP-01 indicates the need to present the text of the quotation in the MLP. This text, including the DOI, should also, if possible, be attached to the distributed data files.

For example, it can be put in the headers of NetCDF files or into a ReadMe.txt file included in the data archive.

DOI-DIFF-01 : The data file must contain a reminder of the quotation elements. At least its DOI.

3.4. GOOD PRACTICES RELATED TO THE DOI PRODUCER

3.4.1. The production of a DOI must be subject to a concerted decision

As soon as a DOI is produced and therefore can be quoted in an article, it is necessary that a quality approach - including in particular treatments aimed at its perpetuation - be implemented around the data concerned. In fact, the decision to produce a DOI for a dataset must be the subject of a concerted decision between the various actors of the data: scientific teams, technical teams, ...

DOI-PROD-01 [Recommended]: The production of a DOI is a decision

3.4.2. The DOI producer should be responsible for the data conservation

As noted above, the registration of a DOI is a significant commitment. It is therefore not advisable to issue a DOI for data that would not be controlled and therefore for which it would not be possible to guarantee the durability.

DOI-PROD-02 [Recommended]: Producer of DOI is responsible for data conservation

3.4.3. Validity of landing pages

It is necessary for a DOI producer to put implement a mechanism that will regularly - daily, for example - verify that the different landing pages are always active and do not correspond to dead links.

This verification mechanism should automatically raise alerts in the IS surveillance system.

DOI-PROD-03 : The producer must implement a mechanism to verify the status of LPs.

3.5. GOOD PRACTICES RELATED TO BIBLIOMETRIC ASPECTS

3.5.1. DOI bibliometric information

The DOI is a UUID whose main role is quotation within publications. It is therefore important to set up a number of bibliometric services around a DOI in order to know the number of times it is quoted. Indirectly, it may also be useful to know whether a DOI is currently not cited.

This service has been set up in some information systems but is not being proposed currently by the main actors in the field, but some work is being carried out in collaboration with the publishers. One example is https://dliservice.research-infrastructures.eu. These third-party mecanisms should be the future solution.

DOI-BIB-01 : Bibliometric information on the use of a DOI should be able to be calculated natively via future third-party mechanisms.

It is possible that these future mechanisms make it possible to distinguish between the different variants of a DOI (fragments). It would be interesting, however, if bibliometric services allow the use of wildcard characters (such as "*").

3.5.2. Specific attribution of authors

DOIs manage the notion of an author. To illustrate this point, it can be emphasized that if an OrcId identifier is present, the contributor's Orcid profile is automatically updated.

Thus, the DOI mechanism opens up the possibility of calculating precise metrics for the producers of data sets themselves, such as the h-index.

Again, services are not yet available. However, a number of assumptions can be made:

- Such services will be proposed by the main players in the field
- These services will not take into account the fragments.

The consequence of the last point is that the calculation of the h-index will only take into account the information available at the time of the calculation within Datacite which means the last list of authors.

Several cases arise:

a) The dataset is not scalable

In this case the author list is frozen. The aforementioned solutions may apply.

b) The dataset is scalable

b1) No need for precise allocation

If the producer of the data set is not interested in precise attribution of the authors, the best practice is to have the list of authors sent to Datacite (and thus that of the MLP) contain all the authors who intervened on the dataset without deletion.

b2) Need for precise allocation

In the case of a scalable dataset, the data producer may want the attribution to relate only to the authors concerned with the downloaded data.

As mentioned above, **in this case the mechanism of the fragments will be ineffective**.

Several solutions are possible:

- Publish a DOI for each product file
- Publish a DOI at each author change

These solutions have the advantage of their simplicity. However, they have the disadvantage of producing a large number of DOIs, most of which will not be used.

A solution, closer to the use of the fragments, may be the following :

1. **Add for the authors a period of activity within the MLP**.
2. When data is downloaded, instead of quoting with the DOI of the dataset (myDoi#myFragment), the system must provide a **quotation that converts the DOI fragment into a static DOI** (for example myDoi__myFragment). This DOI will have the list of authors actually concerned with the data.
3. Publish this static DOI on Datacite (either systematically or via a request mechanism from the person who downloaded the data). This DOI will be compatible with the calculation attribution mechanism of Datacite.

DOI-BIB-02 : The attribution information of the authors must be able to be calculated natively via future third-party mechanisms.
If precise allocation is not required, the mechanism of the fragments must be used.
Otherwise, the mechanism of the fragments must be adapted for the dataset in order to publish static DOIs corresponding to the dataset actually used in publications.

4. FAQ

4.1. WHAT DATA TO QUOTE ?

According to [ANDSFAQ]
A DOI can be applied on totally public or available data with restricted access (e.g. opening to members of a community). In these cases the LP is public and contains either links allowing the retrieval of the data or contact information of the community mediator.

For private data, the use of DOI's external mechanism seems somewhat contradictory and other mechanisms need to be used.

4.1. WHAT METADATA OF A DOI MAY BE MODIFIED ?

According to [ODATISFAQ]
With the exception of the suffix, all metadata can be modified.

4.2. CAN A DOI BE DELETED ?

According to [ODATISFAQ]
It is impossible to delete a DOI. Contractually, an organization that assigns a DOI is committed to keeping the DOI LP online. If a dataset becomes inaccessible, the LP should mention to its visitors the reasons for the deletion of the access to the data.

4.1. CAN DOI BE TRANSFERED ?

It is possible to transfer DOI's responsibility from one entity to another when the second entity is registered with DataCite (and is therefore the repository of a prefix different from the first).

In this case, no change is made in the name of the DOI: the prefix and the suffix remain the same. The transfer of responsibility is purely administrative. However, in general, such a transfer is followed by a change in the URL of the LP.

4.1. WHAT TO DO WITH ISSUED DOI THAT DO NOT COMPLY WITH THE RECOMMENDATIONS OF THIS DOCUMENT ?

Some of the recommendations in this document can be implemented gradually if a DOI is already issued such as those concerning the content of the LP.

On the other hand, some cannot be applied a posteriori. Especially those concerning the suffix. In this case, however, it may be worthwhile to perform bibliographic searches to estimate whether the DOI is actually used to measure the consequences of producing a second suffix on the same dataset (i.e. the same LP). In this case, it will be necessary to keep the first DOI active and to ensure that the information of the two DOIs will keep the same information over time.

4.2. CAN TWO DOI POINT TO THE SAME DATASET ?

Two cases arise:

1) The DOIs have different prefixes

In this case, the answer must be clearly no. Indeed, only the center responsible for conservation must publish the DOI. The introduction of DOI on data managed by another center constitutes a risk of a dead LP in the medium or long term.

2) The DOIs have the same prefix
In this case the *ideal* answer is *no*. However, this case is possible:

- In the case of the issuing a DOI, which is more in accordance with the recommendations as indicated above.

- It is also possible that a fragment mechanism can generate different DOIs for the same data set. However, efforts should be made to prevent this from occurring.

4.3. How to apply these practices to quote several datasets ?

The exact citation of data from the same data producer may relate to several datasets (and thus several DOIs having the same prefix). Thus, it can become long and therefore difficult to quote.

In this case, a good solution is to put in place a metadata sheet describing the extraction and containing all the DOIs in the links section. A DOI that matches with this page will make quotation.

4.4. Should there be a DOI by treatment level ?

Some datasets have several levels of processing (raw data, geophysical parameters, ...). Normally, cataloging should distinguish these processing levels in several separate metadata records, each with its DOI.

Note:
If the records have a genealogy section, they could use the DOIs of the parent datasets.

4.5. Where to quote a DOI of a dataset in a publication ?

According to [ODATISFAQ]

Some journals are beginning to impose specific guidelines. For example, the journal Science requires that the data be cited in *the Acknowledgments* section. The journal PLOS ONE has set up a *Data Availability* section in the heading of its articles.

4.6. Where to find the attributed DOI by DataCite ?

According to [ODATISFAQ]

The metadata of DOI data can be accessed through the site https://search.datacite.org/.