

Data Citation Topic Group Report

Identifying barriers to wider convergence on practices and methods for data citation

**Editors:**  
Carlo Maria Zwölf (VAMDC), Peter Wittenburg (RDA Europe),

Zsuzsanna Szeredi (Vision & Values)

**Co-authors:**

Carlo Maria Zwölf, Peter Wittenburg

**Revision history**

|  |  |
| --- | --- |
| 2019/07/19 | V 1.0 Initiating the document |
| 2019/08/27 | V 1.1 Adding element from discussion (and comments) with Peter Wittenburg |

**Index**

1-Introduction 3

2-General concepts on data-citation 3

2.1-Preliminars: discussion on the linking between data-publication and data-citation 3

2.2-A neutral definition of Data-Citation 3

3-Some issues related with data citation 4

3.1-On the dual role of data citation 4

3.2-Attribution stacking 5

3.3-FAIR principles and attribution 5

3.4-Network vs. Networkless citations 5

3.4.1- Human oriented or machine oriented citation? Another aspect of the network/network-less issue. 6

4-Conclusions 6

# 1-Introduction

Citation is a key element in the production of new knowledge, since it enhances trust (the new results are based on proven/solid bases and a scientist does not need to prove again a used result), makes the process described by the cited work reproducible and gives credits to the author of the cited intellectual product.

According to the FAIR principles, most of the data should be re-used in derived works: the role of Citation is crucial in open-data-driven science. However, the data citation landscape is fast evolving and highly fragmented (*creolization phase*). The *Group of European Data Experts* (*GEDE*) established the *Data-Citation Topic Group* (*DCTG*) with the aim to replicate in data-citation what previously achieved in the case of Persistent Identifiers (*PID*): writing a document for fostering convergence and clarity about data-citation.

If the approach followed in *DCTG* is similar to the one adopted for *PID*, the output from *DCTG* is slightly different: unlike in *PID* case, it is still too early for a wider consensus and convergence on data-citation. In this report we will try to identify the remaining hurdles for consensus forming and convergence of practices.

We hope that this report may be a useful basis for further work and initiatives willing to remove the identified barriers.

# 2-General concepts on data-citation

## 2.1-Preliminars: discussion on the linking between data-publication and data-citation

The *data-publication* concept is closely linked to the *data-citation*: different understanding of the first concept may alter the interpretation of the second one.

*Some communities (e.g. Crossref) consider that the concept of “publication” implies a set of minimal required quality including curation, peer-reviewing and rich-metadata backing. The difference between “citation” and “reference” artificially imply the quality: a citation point to quality “publication” while a reference points to a content with no guarantee of quality.*

We consider that this distinction is not relevant, since it raises more questions than it answers: if we consider that “publication” implies a certain level of quality, what is the minimal required standard for being considered a “publication”? There is no agreement on the required degree of curation and this may differ from one scientific community to the other. For releasing this blocking point, in this work we focus on data-citation aspects being agnostic of what people consider as “publication”: once a datum is published (i.e. publically available whatever the curation level is), how to cite this datum?

## 2.2-A neutral definition of Data-Citation

Let us try to provide a definition, as neutral as possible, of what we call data-citation:

* ***Data-Citation****: The practice of providing a reference pointing to some data*

Even in this simple definition we did not succeed in finding a global consensus: e.g. the community of atomic and molecular data producer call data-citation the practice of pointing not directly to the data, but to the bibliographic references attached to the data (i.e. the papers and publications used for producing the data). We rather suggest calling this procedure *data-source-referencing* for avoiding ambiguities.

When introducing our Data-Citation definition we defined it “neutral” in the sense that:

* We are agnostic about what is consider a data-publication (cf. par. 2.1)
* We attached no particular purpose to the citation practice. Some definitions indeed include the goals of citation: *A data-citation is a reference to data for the purpose of credit attribution and facilitation of access to the data* [Out of Cite, out of Mind]. This is a first overview of the dual role of data-citation, which will be developed in sec. 3.2.
* It does not depend on the citing entity. This may be a human or a machine. The definition then applies also for automatic workflows, removing any eventual subtlety overloaded by people between referencing and citing.

## 2.3-Data-Citation and open science.

It is important to underline that the data-citation application field is broader than *Open-Science:* an author may cite in his/her work a closed data-set, with the dual goal of giving credits to the authors of the cited data-set or documenting the knowledge production workflow. Only the readers authorized to access the cited data-set may follow the worklfow. This is exactly the same in classic scientific publications: an author may cite a paper published in a commercial journal. The citation is public and allows any reader to reach the referenced publication. But if the reader does not pay for accessing the review, the paper is not accessible. Data-Citation and Open-Science are two disjoint problems, since the same data-citation issues occur in Open or Closed Science frameworks. However, both this forms of science rely on citations.

# 3-Some issues related with data citation

In this section we give an overview of some issues we identified as blocking for forming a wider consensus around data-citation. This list is not exhaustive.

## 3.1-On the dual role of data citation

As we mentioned in sec. 1, data-citation is a key element for reproducibility (by giving access to the data used for producing a given result) and for giving credits to the data producers.

The citation granularity required for these two mechanisms is not the same. To illustrate this, let us consider a usual data-driven knowledge production workflow: a scientist extracts some data *D* from a repository. By reusing this data the scientist produces new results and write a new paper citing *D*. Let assume that *D* has multiple authors (*A1, A2, A3*), each one having contributed separately to disjoints sub-parts *D1, D2, D3* of *D*. If the scientist uses only *D1* and *D2* for producing his/her results, then:

* Citing *D* authored by *A1, A2, A3* is a good citation in term of workflow reproducibility.
* Citing *D* is not a good citation in terms of giving bibliographic credits to the authors of the reused work. The author *A3* will be acknowledged while his/her work is not reused.

In the “credit-oriented perspective” the scientist, at the very end of his/her workflow, should refine the granularity of *D* for citing only and exactly what he/he really used. Does the repository from which *D* has been extracted allow this refinement? Or should the scientist perform it manually? In any case, this supplementary task hampers the good data-citation reflex.

We may also note a difference, between the “reproducibility-oriented perspective” and the “credit-oriented perspective”, in terms of access and availability of the cited data: in the first case long-lasting access and availability of the data is a crucial requirement. This is not the case in the latter case, where a crucial aspect is the identification of the authors.

## 3.2-Attribution stacking

This issue is strictly linked with what described in sec. 3.1. Authors must pay greatest attention in citing all the re-used data. Automatic citation tools (RDA compliant query stores, Scholix) are heading in this direction. However automatic citation tools, together with automatic workflows and nano-publications tools, may multiply the amount of citation, amplifying the data stacking effect.

The attribution stacking may also multiply the consequences of the wrong attribution phenomenon described in sec. 3.1.

## 3.3-FAIR principles and attribution

This issue is the other side of the coin of what described in sec. 3.2 (where some authors may be unduly or over-acknowledge by citations). The current issue arise when some authors do not receive proper (or expected) acknowledgement for his/her re-used work. The FAIR principles may amplify this issue, since they encourage massive data reuse.

Let us consider the following example: let *A1* be the author of a data-set *D1*. Assume that:

* *D1* is essential to another author *A2* to build an intellectual work *W* and
* *W* is massively cited (also through automatic tools), while *D1* is cited only once in *W*.

In this kind of configurations some data practitioners consider it not right that only *A2* (and not *A1*) receives credits when W is cited.

*It is worth noting that the described mechanism is the actual accepted standard for classic paper citation. The community perception is not the same while transposing the “paper practice” to data. We highlight this community feeling without trying to understand its origin or the underlying mechanisms.*

The previously cited automatic citation and bibliometric tools may help in addressing this issue, but there is a danger of slipping into the “Attribution stacking” issue described in sec. 3.2.

How far back shall we go in the citation-tree and at which point should be stopped the attribution of bibliometric credits?

## 3.4-Network vs. Networkless citations

While discussing all the technical aspects linked with data-driven science, a question is usually underlying:

*Do the data-science infrastructures and related technical solutions rely on a permanent Internet network access, or should they have a certain standalone behaviour/capability? (i.e. retain the option of working without permanent connection).*

This question may seem trivial or out-dated nowadays. However some industries, for security and confidential reasons, have part of their IT infrastructure separated from the Internet.

In the data-citation context this underlying question takes the particular form

*Does the data-citation mechanism rely on a network connection for operating and working?*

* If **YES**, the data-citation goes back to put the resolvable PID of the resource to cite. All the useful information (authors, curators, licenses, the data-set itself, etc…) may be recovered by simply resolving the PID.
* If **NO**, the data-citation should embed a set of limited metadata (typically authors, production date, repository, title, identifier…). This kind of citation may be said “semantically encoded” since some information is “coded” into the citation itself and the user have to understand the used code in order to consume the information.

In the case of “semantically encoded” citations, there is no agreement of standard about the metadata to include into the citation text. The existing best practices depend on disciplines specificities.

The discussions trying to define standards for encoding the metadata in the citation text are usually cut short by the argument “use PIDs and resolve them, no need to encode since you may resolve it”. A clear example of this was the discussion at the RDA plenary 10 (https://www.rd-alliance.org/data-citation-metadata-elements-rda-10th-plenary-bof-meeting): a convergence was not possible since people were working together on data-citation issues, but having given opposite answers to the “network/networkless” implied question.

### 3.4.1- Human oriented or machine oriented citation? Another aspect of the network/network-less issue.

By nature, a human is more a reader than a web-resolver and the data-citation always impacts humans: a human would like to understand something about the data-citation he/she finds in a scientific paper without having to resolve the data and parsing the landing page. These human based considerations go into the direction of a networkless approach. Should we work on two different mechanisms and procedures, one for humans (networkless oriented) and the other for machines (network oriented). In this case, how to coordinate the two mechanisms?

# 4-Conclusions

The GEDE Data-Citation Topic Group analysed the data-citation procedures, documents and best practices coming from different scientific communities and widely participated to Data-Citation discussion during RDA plenaries and events.

We reached the conclusion that conditions are still not fulfilled for a wider convergence on this topic. Lot of confusion remains about data-citation. We oriented our work for identifying the main confusion causes. This confusion starts because data citation plays a dual role –credit attribution and reproducibility, cf. sec 3.1-

The different interpretations of the role of data-citation inevitably lead to different technical and social consequences, which contribute to increase the confusion as described through sec. 3.

GEDE aim that having stated the confusion sources may help the data-community in building a wider consensus on practices and methods for data-citation.