

## PRACTICE PAPER

# Bringing Citations and Usage Metrics Together to Make Data Count

Helena Cousijn<sup>1</sup>, Patricia Feeney<sup>2</sup>, Daniella Lowenberg<sup>3</sup>, Eleonora Presani<sup>4</sup> and Natasha Simons<sup>5</sup>

<sup>1</sup> DataCite, DE

<sup>2</sup> Crossref, US

<sup>3</sup> California Digital Library, University of California Office of the President, US

<sup>4</sup> Elsevier, US

<sup>5</sup> Australian Research Data Commons, AU

Corresponding author: Helena Cousijn ([helena.cousijn@datacite.org](mailto:helena.cousijn@datacite.org))

Over the last years, many organizations have been working on infrastructure to facilitate sharing and reuse of research data. This means that researchers now have ways of making their data available, but not necessarily incentives to do so. Several Research Data Alliance (RDA) working groups have been working on ways to start measuring activities around research data to provide input for new Data Level Metrics (DLMs). These DLMs are a critical step towards providing researchers with credit for their work. In this paper, we describe the outcomes of the work of the Scholarly Link Exchange (Scholix) working group and the Data Usage Metrics working group. The Scholix working group developed a framework that allows organizations to expose and discover links between articles and datasets, thereby providing an indication of data citations. The Data Usage Metrics group works on a standard for the measurement and display of Data Usage Metrics. Here we explain how publishers and data repositories can contribute to and benefit from these initiatives. Together, these contributions feed into several hubs that enable data repositories to start displaying DLMs. Once these DLMs are available, researchers are in a better position to make their data count and be rewarded for their work.

**Keywords:** research data; data metrics; data citation; Research Data Alliance; Scholix

## 1. Introduction

The importance and value of sharing data is well known and increasingly accepted by the scientific community (Piwowar & Vision 2013); the benefits too great to ignore. Research can be better validated and understood by fellow researchers. Existing research can be reproduced and expanded. Researchers who want to build on published research can reuse existing data to arrive at new conclusions (Bierer et al. 2017). In addition, linking scholarly literature and data leads to increased visibility, discovery and retrieval of both literature and data, facilitating reuse, reproducibility and transparency. In a digital world where data can be more easily shared and documented, scholarly literature and its underpinning data are increasingly seen as inseparable.

At the same time, while the importance of data sharing is accepted, there are essential questions that still require an answer (Borgman 2012). For example, why should authors go through the effort of documenting and publishing datasets, if their career depends on the publication of articles (Mongeon et al. 2017)? How can funding bodies and other assessment boards include data in the evaluation of projects and people when there is no recognized metric or method to measure the quality and impact of the published data? How can the community create the necessary infrastructure for publication and evaluation of data, if there is no standard for metadata and basic attribution information around data? Several RDA projects are underway to provide answers to these questions by creating a framework to measure data reuse in a standardized fashion.

Finding the right way to measure the impact of shared data is crucial if research data is to be included as one of the scholarly outputs used for research evaluation. The current meritocratic system in academia relies heavily on the publication of scientific results in recognized academic journals, supported by an international editorial board and peer review system. The most commonly used metric to measure the impact of a publication is counting the number of times it receives a citation from other publications that are also peer reviewed and published in recognized journals. This currently offers a base to support quality assessments for research projects, career advancement, and funding opportunities (Cantu-Ortiz 2017).

The temptation to use the same metrics for data, and measure citations of datasets in articles, is certainly strong. However, the interaction and impact of research data is more complex than that. The very definition of what a citation for data is, is fuzzier than the equivalent for articles. At the time of writing, community practices around data citations have not fully evolved: it will take time and discipline to reach a unified and standard citation method for data (Silvello 2018). In addition, there are different ways to interact with a dataset. As described in Kratz and Strasser (2015), the value that a researcher gets from data is not given by simply opening its description page. For articles, reading it online or downloading it offers practically the same value to the reader, but a dataset would need to be downloaded to be fully consumed. While citations remain the most popular metric to measure the impact of any scholarly output within the academic community (Kratz and Strasser 2015), measuring the impact of data encompasses more dimensions. Citations will therefore need to be accompanied by other metrics, which could include data usage statistics and social media mentions.

In this paper, we describe how the outputs of two RDA working groups (WGs), the Scholix WG and the Data Usage Metrics WG, can be used to assess data reuse and make data usage statistics and citations available. We will first outline how data repositories and publishers can expose article-data links using Scholix approaches and data usage metrics following the new code of practice for research data. We will then explain how they can consume this information to make DLMs available and help researchers get credit for their work.

## 2. Data Citation

### 2.1. Scholix: aggregating article-data links to count data citations

The goal of the Scholix WG was to establish a high-level framework for exchanging article-data links. It aimed to enable an open information ecosystem to understand systematically what data underpins literature and what literature references data (Burton et al. 2017a).

While there are clear benefits to literature and data linking, in practice these links are difficult to find or share. The main reason for this is that there is no universal way of exchanging link information between organizations and systems which hold this information. Instead, there are different bilateral agreements and technical frameworks for exchanging link information between the different partners and systems that hold this information.

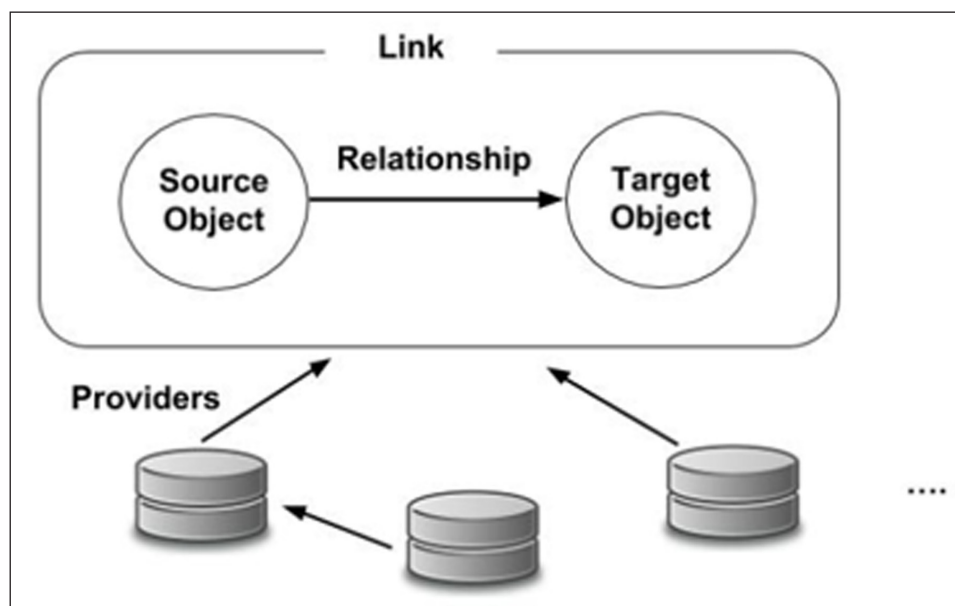
The Scholix WG addressed this problem. Its goal was to improve the links between scholarly literature and research data as well as between datasets, thereby making it easier to discover, interpret, and reuse scholarly information. The Scholix initiative offers:

- a universal, global framework that enables information about the links to be exchanged
- technical guidelines that specify how the interoperability framework works
- a common conceptual model, an information model, and open exchange protocols
- a community that discusses, develops and applies these specifications

Within the Scholix framework:

- Data repositories, journals, and others provide information about the links between literature and data that they hold to community 'hubs' such as OpenAIRE, Crossref and DataCite (with Crossref and DataCite working on a shared infrastructure). This supports and respects existing community-specific practices and the existing means of exchanging this information.
- The community 'hubs' – which are natural places to collect and exchange information about the links between literature and data – commit to a common information model for exchanging the links that they hold and an agreed open exchange method enables this to occur.

The conceptual model (**Figure 1**) is about the link between two objects, such as a journal article and the underpinning data. Rather than describing in detail the properties of each of the two objects, the conceptual model focuses on the relationship between the objects. It also enables a record of who asserted the link and who made the link available.



**Figure 1:** Scholix information model. Providers contribute links by sharing information about the source object (article or dataset), target object (article or dataset) and the nature and direction of the relationship.

The Scholix metadata schema (Burton et al. 2017b) includes a range of relationship types that may be applied to describe the relationship between an article and a dataset, for example ‘isReferencedBy’. This allows for the identification of data citations and therefore provides input for the development of citation metrics for data.

## 2.2. Contributing data citations: publishers

As mentioned in the previous section, within the Scholix framework organizations contribute information through community hubs. The majority of scholarly publishers work with non-profit organization Crossref to share metadata about publications. These metadata records include comprehensive information about the items being registered, and increasingly include links to related scholarly artifacts such as data, software, protocols, and reviews. When data citations are included in Crossref metadata records they are made available to the wider community.

As can be seen in **Figure 2**, Crossref provides two paths to registering data citations: references and relations. Relations are a way to associate related digital objects with each other through metadata. A publisher can register metadata with Crossref explicitly linking a dataset to a journal article. References are formal citations (such as would be provided in a bibliography) and are a type of relation but are provided separately within Crossref metadata.

Crossref members should deposit data citations as references if:

- the data citation includes a DataCite DOI
- they include data citations in their reference lists (recommended)

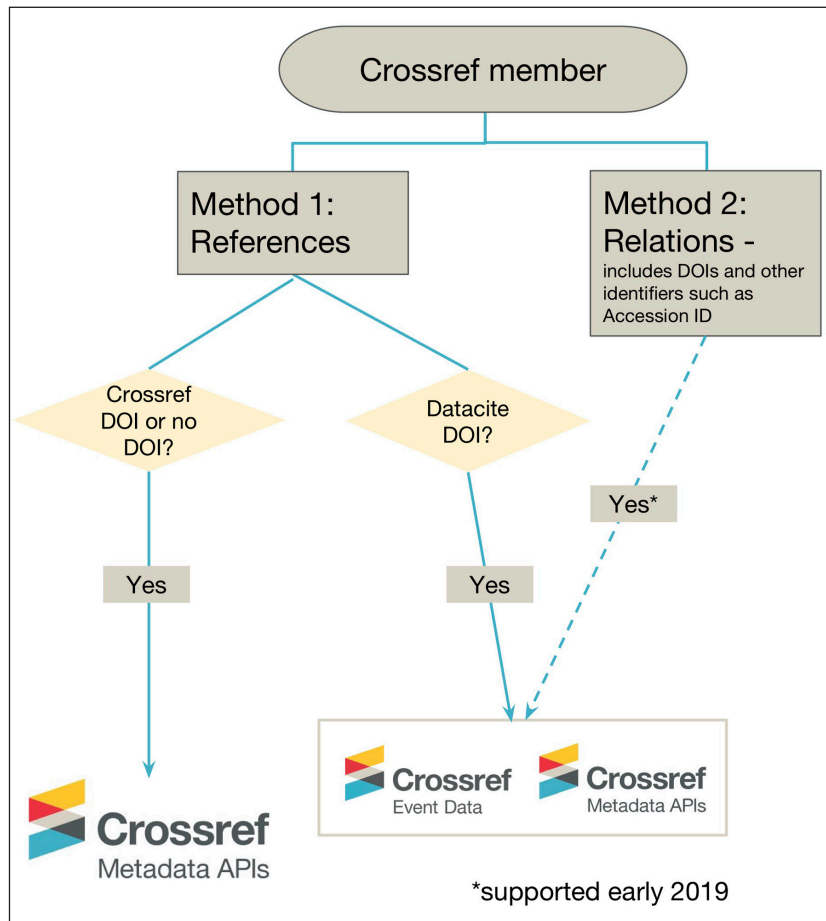
Crossref members should deposit data citations as relations if:

- they want to capture specific relation types (e.g. isSupplementedBy) beyond ‘references’
- they are not able to supply data citations as references

In 2019 Crossref will be expanding citation support to allow publishers to explicitly identify data citations in line with the data citation roadmap for scientific publishers (Cousijn et al. 2018). This will allow for deposition of data citations with all types of persistent identifiers as references.

## 2.3. Contributing data citations: data repositories

Many data repositories actively curate and keep track of which articles are using the datasets they host. This is valuable information that is currently not always available to other organizations in the data community. For data repositories that use DataCite DOIs, the DOIs and accompanying metadata are registered with



**Figure 2:** Depositing Data Citations with Crossref. Publishers can deposit data citations following two different methods: references or relations.

DataCite. Therefore, information about any journal publications related to a dataset can be included in the metadata records that are sent to DataCite. This additional information should follow the DataCite metadata schema which is aligned with the Scholix metadata schema (Burton et al. 2017b).

Dataset metadata can be enriched with links between literature (related resources) and data by including the related identifier. The DataCite Metadata Schema supports the `relatedIdentifier` property which is used to generally connect an object to other related resources, in this case datasets to articles. The identifiers used by this property must be globally unique. The `relatedIdentifierType` indicates the type of identifier, e.g. DOI. `relationType` describes the relationship of the resource being registered with a DOI and the related resource. To describe the relationship between datasets and articles, `IsReferencedBy`, `References`, `IsSupplementTo` or `IsSupplementedBy` are usually appropriate, depending on the actual relationship. In most cases, a dataset `IsSupplementTo` an article or a dataset `IsReferencedBy` an article. The metadata schema also includes “`resourceTypeGeneral`” to describe the resource type of the related resource, e.g. “Text” for a journal article with Crossref DOI.

When these elements are added to the metadata that is registered with DataCite, the information about the links will automatically become openly available.

#### **2.4. Contributing data citations: institutional repositories**

For data centers that do not assign DataCite DOIs to datasets, OpenAIRE is currently the best place to deposit article-data links. Institutional repositories can export metadata descriptions of their datasets with links to articles as Dublin Core records or as Scholix records and register with OpenAIRE’s Scholexplorer Service (Burton et al. 2017c) as a data source. Scholexplorer will bulk collect metadata records from the repository APIs; Scholexplorer is compatible with the OAI-PMH protocol or REST search APIs that allow collection of all records with a paging system (collecting by means of several calls) and with “last date of indexing” (incremental approach). Scholexplorer will then enrich its graph of article-dataset links with the ones collected from the repository, de-duplicate when necessary, and expose all links as Scholix records via APIs on

behalf of the registered repository. All links exported by OpenAIRE carry provenance information about the data sources that provided the links (more than one source may have provided the same link), to ensure visibility of the contributing repositories and provide a degree of trust to the consuming services. OpenAIRE asks the database to display the Scholix logo on their website and indicate that it is harvested by Scholexplorer.

### 3. Data Usage Metrics

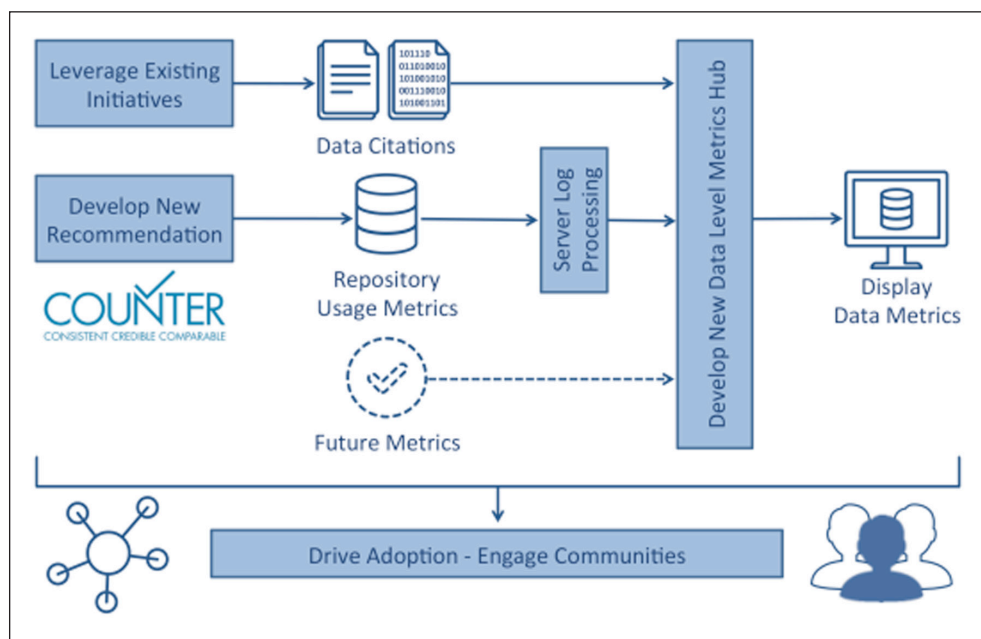
#### 3.1. Standards for data usage metrics

Following the Scholix initiative and the related work of the RDA Data Citation WG, it was clear that there are broader metrics for data that the community needs to address. With the Scholix working group focusing on the relationships between articles and datasets and the Data Citation Working Group addressing challenges related to dynamic data citation (Rauber et al. 2016), there was a need for a working group to define usage for data. The Data Usage Metrics WG started in 2018 and focuses on metrics that reflect usage of research data. The group is working to build a comprehensive list of use cases that covers the spectrum of types of 'usage metrics' that may apply to research data, build a recommendation for community guidance on what types of usage metrics should be applied at the data and repositories levels, and drive adoption of usage metrics across the research landscape. Specifically, the working group is aimed at outlining the barriers to adoption of data-level-metrics standards and current implementations of usage metrics across the data repository landscape. These conversations, surveys, and findings will aid in defining recommendations for types of data and associated metrics that repositories should be considering. The group works closely with the Make Data Count project and leverages the COUNTER code of practice for research data (below).

#### 3.2 Contributing data usage metrics

Data are different and more complex than articles in regard to counting usage. Data are more than just a PDF, they can be versioned, they have multiple components, and they are accessed and used in a variety of ways by humans and machines. Because of this, a standard recommendation for data usage metrics was necessary (Fenner et al. 2018). The Make Data Count project has been working on developing a standard and corresponding infrastructure and COUNTER formally endorsed this as the official COUNTER recommendation for data. The WG is using this Code of Practice as a starting point for further recommendations.

This first release of the *Code of Practice for Research Data* specifically targets research data usage. The recommendations are aligned as much as possible with the *COUNTER Code of Practice Release 5* for the major categories of e-resources (journals, databases, books, reference works, and multimedia databases)



**Figure 3:** Framework of the Make Data Count project. Repositories process log files against the new Code of Practice and these processed files feed into the same hub as the article-data links collected following the Scholix framework. All this information is made openly available to the community so organizations can develop and display DLMS.

and mainly concern views and downloads – called investigations and requests in the Code of Practice. Many definitions, processing rules and reporting recommendations apply to research data in the same way as they apply to other resources. The *Code of Practice for Research Data* enables the reporting of usage statistics by different data repositories following common best practices, and thus is an essential step towards realizing usage statistics as a metric available to the community to better understand how publicly available datasets are being reused.

As a first example, utilizing the COUNTER standard, developers at California Digital Library developed an open source Python tool for log processing (Make-Data-Count, 2018) and all processes are documented in a “how-to” format so that the tool can be reused by other data repositories.

The resulting reports are sent to DataCite, where the reports are processed and usage counts for individual DOIs are made available (see below). **Figure 3** shows how usage statistics and citations feed into the same hub to be shared with the community.

#### 4. Consuming data usage statistics and citations

The citations and usage statistics contributed by data repositories and publishers are made openly available to the community via APIs. Crossref and DataCite developed Event Data, a shared underlying infrastructure that holds (among other things) all citations that are contributed as part of article and dataset metadata. Crossref and DataCite each have their own API through which they make these citations available.

Services such as Scholexplorer retrieve data citations from the Crossref Event Data service using this Scholix API endpoint: <http://api.eventdata.crossref.org/v1/events/scholix>.

Scholexplorer combines this information with the citations that are provided to OpenAIRE.

Views and downloads processed against the COUNTER Code of Practice are sent to DataCite and any repository or research data service can consume usage statistics for a given dataset DOI from an Event Data Query API provided by DataCite (<https://support.datacite.org/docs/eventdata-guide>). The API combines citations and other events into one API call.

All these APIs can be used by data repositories and publishers wanting to consume and display links between articles and datasets.

#### 5. Conclusions

Measuring data (re)use and the development of DLMs are crucial if data is to become a first-class research output. Both the Scholix and Data Usage Metrics WGs are making significant contributions in this area by developing clear guidance on how to collect and share data usage statistics and article-data links. Whereas the Scholix WG has reached the end of two very successful 18 month working group terms, the Data Usage Metrics only just started and will continue the work on DLMs and the adoption thereof.

In this paper, we described how data repositories and publishers can contribute to and participate in these initiatives. Work is still ongoing, but organizations can already contribute information through existing community hubs and consume usage statistics and citations through the different APIs available. OpenAIRE was the first to enable this through the Scholexplorer API and Crossref and DataCite both developed EventData APIs, with the DataCite API making both usage statistics and citations available.

The openness of the systems developed offers an infrastructure for collaboration using accepted standards. Community organizations, publishers, data repositories, and service providers can rely on common guidelines and standards to share (re)use information they collect about datasets. The most important next step is for as many organizations as possible to standardize usage counts and contribute usage and citations to the open infrastructure hubs. Once the community has comparable and transparent numbers for usage and citation of data, this enables further work with the bibliometrics community to properly evaluate what metrics are meaningful for research data.

Defining data metrics is by no means trivial, as community practices around data citation still need to crystallize (Silvello et al. 2018). The work by the Scholix WG underlines this. In most cases, article-data links contributed by publishers will provide what is traditionally seen as a citation: formal mention of another research output in an article. When article-data links are contributed by repositories, it is less clear what constitutes a data citation. The relationship types provide a first indication of this and are currently being used to count citations, but more work is needed to develop real citation metrics for data.

When these data-level metrics become available, research data can finally become part of quality assessments for research projects, career advancement, and funding opportunities and researchers can be rewarded for sharing data.

## Acknowledgements

We would like to thank the co-chairs of both working groups: Adrian Burton, Martin Fenner, Wouter Haak, Paolo Manghi (Scholix), Ian Bruno, and Dave Vieglais (Data Usage Metrics).

## Funding Information

The Make Data Count project is funded by the Alfred P. Sloan Foundation: <https://sloan.org/grant-detail/8020>. This paper was supported by the RDA Europe 4.0 project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 777388.

## Competing Interests

The authors have no competing interests to declare.


## References

- Bierer, B, Crosas, M and Pierce, H.** 2017. Data authorship as an incentive to data sharing. *N Engl J Med*, 376. DOI: <https://doi.org/10.1056/NEJMs1616595>
- Borgman, C.** 2012. The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology*, 63: 1059–1078. DOI: <https://doi.org/10.1002/asi.22634>
- Burton, A, et al.** 2017a. The Scholix Framework for Interoperability in Data-Literature Information Exchange. *D-Lib Magazine*, 23(1/2). DOI: <https://doi.org/10.1045/january2017-burton>
- Burton, A, et al.** 2017b. Scholix Metadata Schema for Exchange of Scholarly Communication Links. *Zenodo*. DOI: <https://doi.org/10.5281/zenodo.1120265>
- Burton, A, et al.** 2017c. The data-literature interlinking service: Towards a common infrastructure for sharing data-article links. *Program*, 51(1): 75–100. DOI: <https://doi.org/10.1108/PROG-06-2016-0048>
- Cantu-Ortiz, F.** 2017. *Research Analytics: Boosting University Productivity and Competitiveness Through Scientometrics*. Boca Raton: CRC Press. DOI: <https://doi.org/10.1201/9781315155890>
- Cousijn, H, et al.** 2018. A data citation roadmap for scientific publishers. *Scientific Data*, 5. DOI: <https://doi.org/10.1038/sdata.2018.259>
- Fenner, M, et al.** 2018. Code of practice for research data usage metrics release. *PeerJ Preprints*, 6(e26505v1). DOI: <https://doi.org/10.7287/peerj.preprints.26505v1>
- Kratz, J and Strasser, C.** 2015. Making data count. *Scientific Data*, 2. DOI: <https://doi.org/10.1038/sdata.2015.39>
- Make-Data-Count.** 2018. *Implementing the COUNTER Code of Practice for Research Data in Repositories*. Github. Available at: <https://github.com/CDLUC3/Make-Data-Count/blob/master/getting-started.md> [Last accessed 30 August 2018].
- Mongeon, P, et al.** 2017. Incorporating data sharing to the reward system of science. *Aslib Journal of Information Management*, 69: 545–556. DOI: <https://doi.org/10.1108/AJIM-01-2017-0024>
- Piwowar, H and Vision, T.** 2013. Data reuse and the open data citation advantage. *PeerJ*, 1(e175). DOI: <https://doi.org/10.7717/peerj.175>
- Rauber, A, et al.** 2016. Identification of Reproducible Subsets for Data Citation, Sharing and Re-Use. *TCDL Bulletin*, 12(1).
- Silvello, G.** 2018. Theory and practice of data citation. *JASIST*, 69(1): 6–20. DOI: <https://doi.org/10.1002/asi.23917>

**How to cite this article:** Cousijn, H, Feeney, P, Lowenberg, D, Presani, E and Simons, N. 2019. Bringing Citations and Usage Metrics Together to Make Data Count. *Data Science Journal*, 18: 9, pp. 1–7. DOI: <https://doi.org/10.5334/dsj-2019-009>

**Submitted:** 31 August 2018    **Accepted:** 14 February 2019    **Published:** 01 March 2019

**Copyright:** © 2019 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/4.0/>.

 *Data Science Journal* is a peer-reviewed open access journal published by Ubiquity Press.

**OPEN ACCESS** 