science editing

pISSN 2288-8063
eISSN 2288-7474

Sci Ed 2019;6(2):166-170
https://doi.org/10.6087/kcse.165

Training Material

How publishers can work with Crossref on data citation

Rachael Lammey

Crossref, Oxford, UK

Abstract

It aims to explain why data citation is important, how publishers and data repositories can do this and what use will be made of the information they provide. There are large benefits to be accrued from sharing research data such as guarantee of reproducibility and transparency. Consistent citation practice around data is essential to helping these benefits to be realized. Data citation metadata is being disseminated and used through its application programming interfaces and the Event Data application programming interface. Event Data extracts this information into a separate service, so data citations are pre-filtered from the Crossref metadata. There are two methods by which publishers can register data citation information with Crossref. The first method is to deposit data citations in the citation section of the metadata, i.e., the part containing the reference list of the article. The second method publishers can use to register data citations with Crossref is to use the relationships section of the metadata. There are a number of services already using Event Data to show information on data citation. To achieve the benefits of data citation, publishers or editors should have a data sharing and citation policy so that they share with their authors and readers.

Keywords

Crossref; DataCite; Digital object identifier; Data citation; Best practice

Introduction

To put it simply, data citation is done when a journal article references the data the research was built upon or references, in the same way that a paper would normally provide a reference list to other scholarly resources, such as other papers. Even though an increasing number of researchers are starting to share their data, this data is often not cited in the same way as other publications. Furthermore, it is not cited in a consistent way. This makes it hard to track how the data that underlies the research is actually being used if it's sitting in a separate repository or on another website that is not linked to the paper in any way.

Making it easier to see these links between different types of research outputs will provide many benefits to the community. It aids transparency—showing the underlying data that pro-

Received: April 11, 2019
Accepted: May 9, 2019

Correspondence to Rachael Lammey
rlammey@crossref.org

ORCID
Rachael Lammey
https://orcid.org/0000-0001-5800-1434

duced the research results so that you can verify it—and reproducibility—if you can get the data, it can make it easier to reproduce or reuse to replicate or build upon the research.

To incentivise the sharing of research data, researcher need credit for doing this. When they cite the data they use this forms the basis for a data credit system, i.e., a way to reward data sharing by looking at citation and other metrics to measure the impact of the data. Funders are keen to realise these benefits for the research they fund, and many are starting to bring in mandates around how data is published and shared.

How Publishers Can Get Started

There are a few steps that publishers can work through to get to the point where they can register data citations with Crossref. DataCite, Crossref and others have collaborated on a paper that sets out what publishers need to do in a number of steps [1]. These are as follows: 1) develop a data policy that includes data citation; 2) explain to authors how they should be citing data in the submissions they make to their publications; 3) update internal workflows, DTD (Document Type Definition) check and instructions to suppliers (for example proofreaders, typesetters) about how to handle these citations; and 4) include the citations in the metadata that you register with Crossref so that these can be disseminated and used.

Many publishers are already taking steps to implement data citation policies, so others can learn from them, and recent case studies by Springer Nature and Taylor & Francis may also be of assistance [2]. They also provide information on how this information can be communicated to authors via the journal Instructions for Authors and their online submission systems.

Copyeditors, typesetters and other parties who prepare content for publication can be briefed on how to format data citations and data availability statements, and finally, there are a number of methods to register this information with Crossref when publishers deposit identifiers and metadata upon or just before publication.

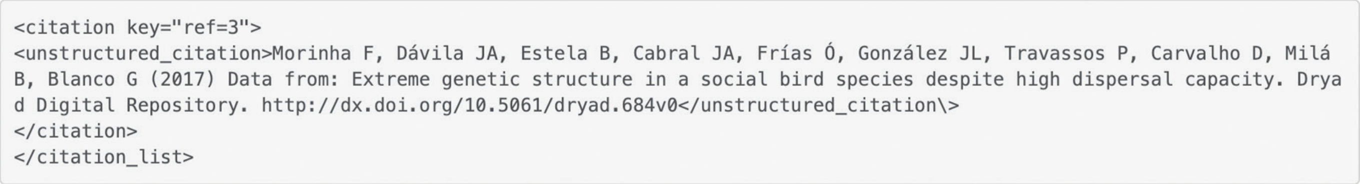
Registering Data Citation Information with Crossref

There are two methods by which publishers can register data citation information with Crossref. Both are supported and publishers are using a mixture of these based on what fits best with their publication workflows.

The first method is to deposit data citations in the citation section of the metadata, i.e., the part containing the reference list of the article. Publishers can deposit the full data or software citation as a unstructured reference, an example is shown in Fig. 1. It is recommended that publishers ask authors to cite the dataset or software based on community best practice, outlined in the Joint declaration of data citation principles [3], FORCE11 citation placement [4], and FORCE11 software citation principles [5].

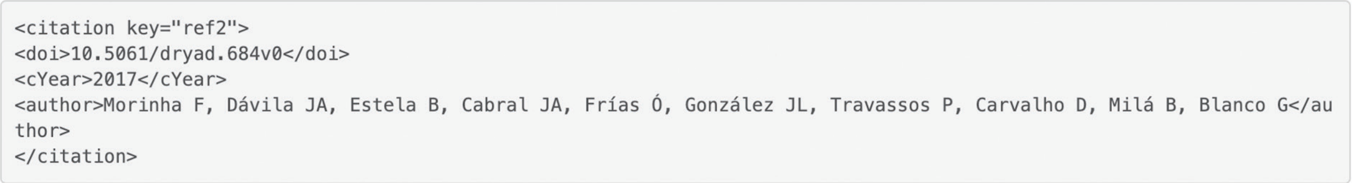
Otherwise, they can employ any reference tags [6] currently accepted by Crossref, shown in Fig. 2. At the very least, even just a reference containing a DataCite DOI is enough for Crossref to recognize that the author is citing data in the reference.

Adding data citations using this method is a good option for publishers who already deposit reference metadata with Crossref (this is optional but recommended) and make this metadata openly available via Crossref (also optional as explained in Crossref's reference distribution policy [7]). If publishers who make their references openly available via Cross-



```
<citation key="ref=3">
<unstructured_citation>Morinha F, Dávila JA, Estela B, Cabral JA, Frías Ó, González JL, Travassos P, Carvalho D, Milá
B, Blanco G (2017) Data from: Extreme genetic structure in a social bird species despite high dispersal capacity. Drya
d Digital Repository. http://dx.doi.org/10.5061/dryad.684v0</unstructured_citation\>
</citation>
</citation_list>
```

Fig. 1. Unstructured citation example of data citation in Crossref metadata.



```
<citation key="ref2">
<doi>10.5061/dryad.684v0</doi>
<cYear>2017</cYear>
<author>Morinha F, Dávila JA, Estela B, Cabral JA, Frías Ó, González JL, Travassos P, Carvalho D, Milá B, Blanco G</au
thor>
</citation>
```

Fig. 2. Citation example of data citation in Crossref metadata using Crossref reference tags.

ref deposit data citations in this way, they are automatically made available via Crossref's different metadata delivery methods [8]. If the citations contain DataCite DOIs, they will also be made available by a new service called Event Data [9], which is being co-developed by Crossref and DataCite.

The second method publishers can use to register data citations with Crossref is to use the relationships section [10] of the metadata shown in Fig. 3. As shown in Fig. 3, the relationships method lets publishers provide more specific information as to how the data being cited and the article relate to each other. If the article simply references a dataset, the relationship type 'references' can be used, or if the author wants to specify that the data was generated as part of the research, 'isSupplementedBy' can be used. A description of the data, an identifier for the data and the identifier type should all be provided, and this identifier type is not limited to the DOI, many other types of identifiers are accepted: PMID, PMCID, PURL, ARK, Handle, UUID, ECLI, and URI.

Publishers may want to use this method for depositing data citations if they prefer to provide this level of specificity as to how the data is related to the paper. This information can be used to support scientific validation and funding management, i.e., funders may use it to check that authors they fund are following their mandates regarding data availability. Publishers who do not make their references openly available via Crossref should also use this method.

For publishers interested in using the reference method of depositing this information, Crossref is working to implement the JATS4R recommendations [11] which will let publishers add more descriptive information like relationships between data and publications to their reference list metadata. These changes to the Crossref schema will be implemented in 2019.

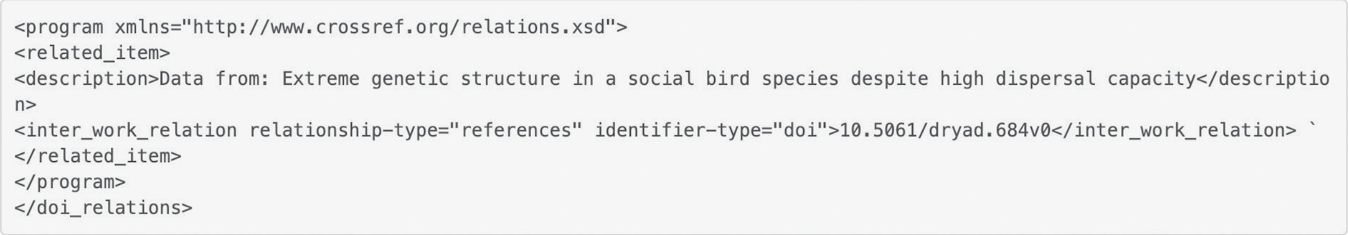
Data citation metadata can be registered with Crossref using the existing methods that members use to register content, e.g., it can be added to complete metadata deposits, or registered in reference-only deposits [12], which many publishers use to add reference data to their records. Additionally, relationships information can also be added to existing Crossref deposits using a resource deposit [13]—so that a publisher can patch this into their metadata if they want to add it retrospectively. Crossref's new Metatdata Manager tool (beta) [14] supports the deposit of both types of article/data links.

How Data Citation Metadata is Being Disseminated and Used

Asking authors for this information, collecting it and then registering it with Crossref in a standard way is a valuable thing for publishers to do. However, this information then needs to be disseminated effectively so that the research community can get maximum value from it. Crossref shares information on data/article links via a number of output methods: its application programming interfaces (APIs) and the Event Data API. Relationships metadata is not currently available via Event Data, but this will be added in 2019.

The Event Data service was jointly developed by Crossref and DataCite to capture references, mentions and other events around DOIs that are not provided via DOI metadata. It includes references of different DOI registration agencies, like data citations. Before Event Data, if anyone wanted to find links from articles to data, they would have to access the full set of Crossref metadata and extract this piece of information across all Crossref DOIs and metadata. The same goes for data to article links—the DataCite metadata would all need to be mined and this information extracted. This approach is quite manual and therefore does not scale well, so pubishers and repositories were disincentivised to provide this information as they could not see it getting used.

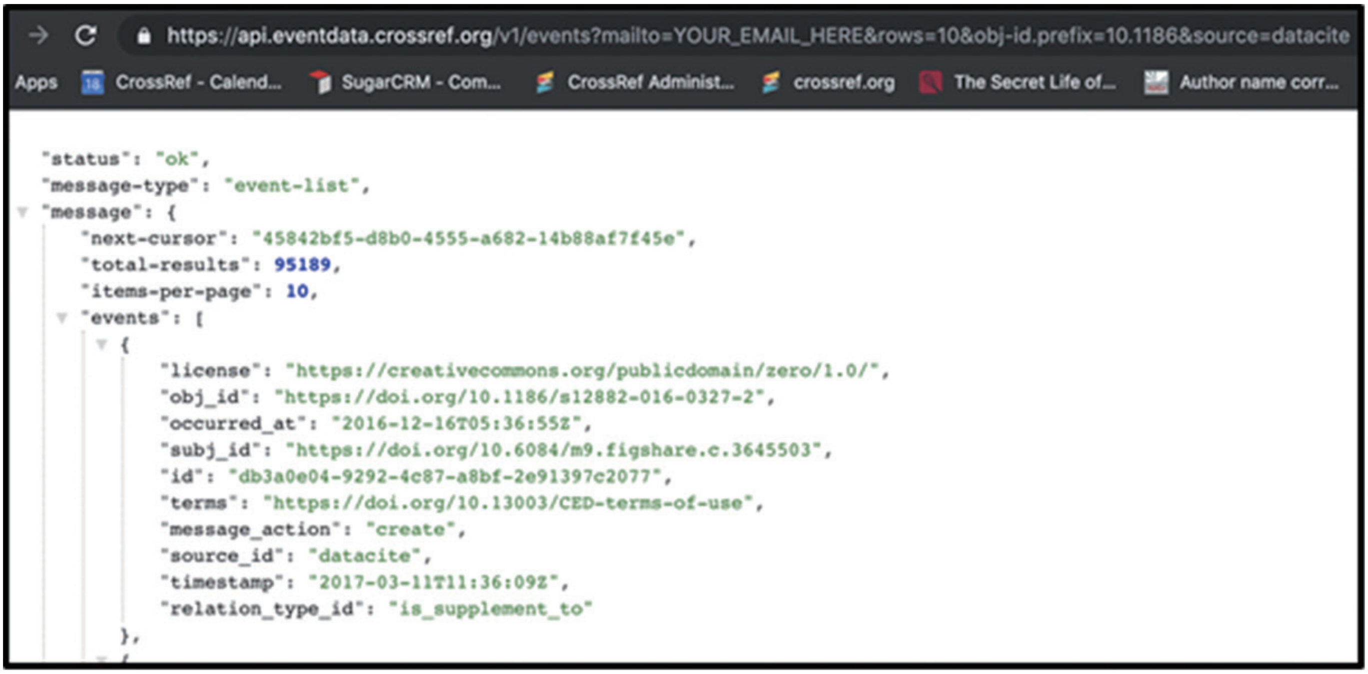
Event Data extracts this information into a separate service, so data citations are pre-filtered from the Crossref metadata (and the same at DataCite for data to article links). That way they can be easily found, filtered upon and integrated into tools and services. For example, a publisher interested in seeing the links to their publications from dataset metadata held by DataCite can query the Crossref Event Data API by their DOI prefix to ask for that information (Fig. 4). Both Crossref and DataCite's Event Data APIs are open and free to use, so anyone can access this information.



```
<program xmlns="http://www.crossref.org/relations.xsd">
<related_item>
<description>Data from: Extreme genetic structure in a social bird species despite high dispersal capacity</descriptio
n>
<inter_work_relation relationship-type="references" identifier-type="doi">10.5061/dryad.684v0</inter_work_relation> `
</related_item>
</program>
</doi_relations>
```

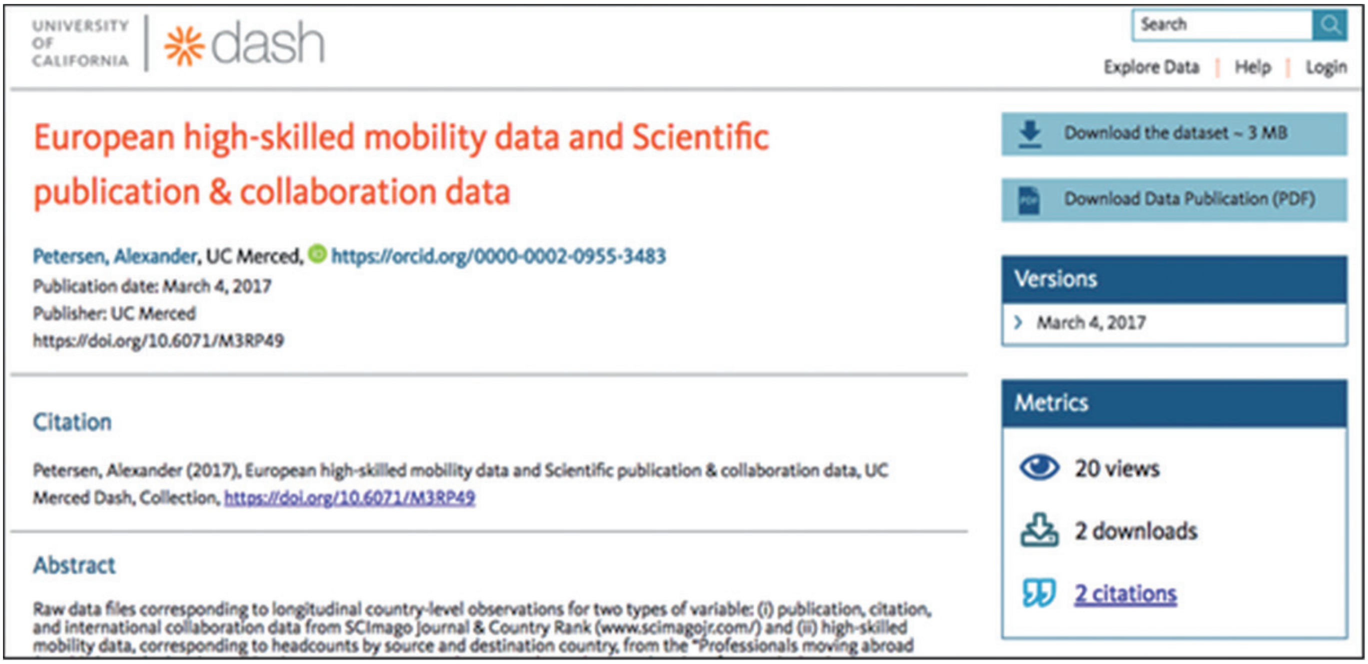
Fig. 3. Example of data citation in Crossref metadata using the relationships schema.

There are a number of services already using Event Data to show information on how data is being cited, so that researchers can easily see who is interested in the data that they have produced. Fig. 5 shows the University of California's Dash tool. Dash is a self-service tool for researchers to describe, upload, and share their research data. To the right of the bibliographic information on the dataset they are displaying the citations to the data that they have uncovered using Event Data. Clicking


https://api.eventdata.crossref.org/v1/events?mailto=YOUR_EMAIL_HERE&rows=10&obj-id.prefix=10.1186&source=datacite
Apps
CrossRef - Calend...
SugarCRM - Com...
CrossRef Administ...
crossref.org
The Secret Life of...
Author name corr...

```
 "status": "ok",
 "message-type": "event-list",
▼"message": {
    "next-cursor": "45842bf5-d8b0-4555-a682-14b88af7f45e",
    "total-results": 95189,
    "items-per-page": 10,
  ▼ "events": [
    ▼ {
          "license": "https://creativecommons.org/publicdomain/zero/1.0/",
          "obj_id": "https://doi.org/10.1186/s12882-016-0327-2",
          "occurred_at": "2016-12-16T05:36:55Z",
          "subj_id": "https://doi.org/10.6084/m9.figshare.c.3645503",
          "id": "db3a0e04-9292-4c87-a8bf-2e91397c2077",
          "terms": "https://doi.org/10.13003/CED-terms-of-use",
          "message_action": "create",
          "source_id": "datacite",
          "timestamp": "2017-03-11T11:36:09Z",
          "relation_type_id": "is_supplement_to"
      },
```

Fig. 4. Event Data application programming interface results (this sample query asks Event Data to show 10 events related to DataCite DOIs (links to datasets) for the PLOS (Public Library of Science) DOI prefix.


UNIVERSITY OF CALIFORNIA
dash
Search
Explore Data | Help | Login
Download the dataset ~ 3 MB
Download Data Publication (PDF)

European high-skilled mobility data and Scientific publication & collaboration data

Petersen, Alexander, UC Merced, https://orcid.org/0000-0002-0955-3483
Publication date: March 4, 2017
Publisher: UC Merced
https://doi.org/10.6071/M3RP49

Versions
› March 4, 2017

Metrics
20 views
2 downloads
2 citations

Citation

Petersen, Alexander (2017), European high-skilled mobility data and Scientific publication & collaboration data, UC Merced Dash, Collection, https://doi.org/10.6071/M3RP49

Abstract

Raw data files corresponding to longitudinal country-level observations for two types of variable: (i) publication, citation, and international collaboration data from SCImago Journal & Country Rank (www.scimagojr.com/) and (ii) high-skilled mobility data, corresponding to headcounts by source and destination country, from the "Professionals moving abroad

Fig. 5. Data citation display in the University of California Dash repository (https://dash.ucmerced.edu/stash/dataset/doi:10.6071/M3RP49).

on the hyperlink then brings up the specific citations to the dataset which is useful information for anyone looking at this.

Regarding these citation counts, the Make Data Count project is working (among other things) to address the significant social as well as technical barriers to widespread incorporation of data-level metrics in the research data management ecosystem through consultation, recommendation, new technical capability, and community outreach [15]. They are strong advocates around data citation practices. Information on data/article links can also be aggregated by search services and databases. DataCite is integrating information on data citation and usage into its own DataCite search so that anyone using it can see how often a dataset has been viewed, cited or downloaded. Another example is Scholexplorer, service that populates and provides access to a graph of links between dataset and literature objects and dataset and dataset objects [16]. Scholexplorer uses Event Data, along with other sources to populate the information it contains on published articles and datasets. The expectation is that use of this information will grow over time as more publishers and data repositories collect and register this information as part of their standard workflows.

Conclusion

Linking to the data related to a published article provides a wide range of benefits for the research community. To achieve these, publishers should have a policy around data citation that they share with their authors, they should aim to collect this information and register it with Crossref in the metadata they provide, either in the references or relationship section of the Crossref schema. Data repositories can do the same (via DataCite) to assert connections between data and published articles. This information can then be made available by Crossref and DataCite via their metadata and Event Data APIs. This lets them be integrated into a growing amount of tools and services and easily accessed and used by anyone engaged in research.

Conflict of Interest

No potential conflict of interest relevant to this article was reported.

References

1. Cousijn H, Kenall A, Ganley E, et al. A data citation roadmap for scientific publishers. Sci Data 2018;5:180259. https://doi.org/10.1038/sdata.2018.259
2. Jones L, Grant R, Hrynaszkiewicz I. Implementing publisher policies that inform, support and encourage authors to share data: two case studies. Insights 2019;32:11. https://doi.org/10.1629/uksg.463
3. Martone M, editor. Data Citation Synthesis Group: joint declaration of data citation principles. San Diego, CA: FORCE11; 2014. https://doi.org/10.25490/a97f-egyk
4. FORCE11. Example: placement of citation [Internet]. La Jolla, CA: FORCE11 [cited 2019 Apr 6]. Available from: https://www.force11.org/node/4771
5. Smith AM, Katz DS, Niemeyer KE; FORCE11 Software Citation Working Group. Software citation principles. PeerJ Comput Sci 2016;2:e86. https://doi.org/10.7717/peerj-cs.86
6. Crossref. Adding references to your metadata record [Internet]. Lynnfield, MA: Crossref [cited 2019 Apr 6]. Available from: https://support.crossref.org/hc/en-us/articles/215578403-Adding-references-to-your-metadata-record
7. Crossref. Reference distribution [Internet]. Lynnfield, MA: Crossref [cited 2019 Apr 6]. Available from: https://www.crossref.org/reference-distribution/
8. Crossref. Metadata delivery [Internet]. Lynnfield, MA: Crossref [cited 2019 Apr 6]. Available from: https://www.crossref.org/services/metadata-delivery/
9. Crossref. Event Data [Internet]. Lynnfield, MA: Crossref [cited 2019 Apr 6]. Available from: https://www.crossref.org/services/event-data/
10. Crossref. Relationships between DOIs and other objects [Internet]. Lynnfield, MA: Crossref [cited 2019 Apr 6]. Available from: https://support.crossref.org/hc/en-us/articles/214357426-Relationships-between-DOIs-and-other-objects
11. JATS for Reuse. Data citations [Internet]. JATS for Reuse [cited 2019 Apr 7]. Available from: http://jats4r.org/data-citations
12. Crossref. Adding references to your metadata record [Internet]. Lynnfield, MA: Crossref [cited 2019 Apr 10]. Available from: https://support.crossref.org/hc/en-us/articles/215578403-Adding-references-to-your-metadata-record
13. Crossref. Adding metadata to an existing record (resource deposits) [Internet]. Lynnfield, MA: Crossref [cited 2019 Apr 10]. Available from: https://support.crossref.org/hc/en-us/articles/214002366-Adding-metadata-to-an-existing-record-resource-deposits-
14. Crossref. Metadata Manager support documentation [Internet]. Lynnfield, MA: Crossref [cited 2019 Apr 10]. Available from: https://www.crossref.org/help/metadata-manager/
15. Make Data Count. About [Internet]. Make Data Count [cited 2019 Apr 10]. Available from: https://makedatacount.org/about/
16. Scholexplorer. About [Internet]. Scholexplorer [cited 2019 Apr 10]. Available from: http://scholexplorer.openaire.eu/index.html#/about