A generalist perspective on data citation

Alex Ball
University of Bath (and ex-DCC)

ESIP Summer Meeting
17 July 2019
Data Citation and Linking
By Alex Ball and Monica Duke, UKOLN, University of Bath

Introduction
On the surface, citing datasets is a trivially easy thing to do. Style manuals such as the Publication Manual of the American Psychological Association and the Oxford Manual of Style have provided sample citations for datasets since at least the early 2000s. The process of making datasets citable, however, is rather more difficult. In consequence of this and other factors, a culture of citing datasets has been slow to develop. Nevertheless, it is vital that researchers cite the datasets they use, if datasets are to be regarded as legitimate academic outputs in their own right.

Short-term Benefits and Long-term Value
There are several short-term benefits to making datasets citable, citing them in practice, and linking datasets to papers that make use of the data.

- If the authors of a scientific publication properly cite the data that underlies it, it is much easier for the reader to locate that data. This in turn makes it easier for the reader to validate and build on the publication’s findings.

- Data citations ensure that data contributors receive proper credit when their work is reused by other researchers.

- If a dataset links back to the paper that describes its collection, a reader coming to the dataset direct can use that link to put it in context and understand the methodology used.

- If a dataset links to other papers that make use of it, these links can be used by the contributors and data publishers to demonstrate the impact of the data. Potential reusers might use these links to discover critiques of the data or to provide inspiration for how to use them.

Once a culture of data citation has been established, several other benefits are likely to become apparent.

- The publishing infrastructure that makes the data citable will also help to ensure they are available for reference and reuse long into the future.

- There will be less danger of rival researchers ‘stealing’ results from those who publish their data openly, as failure to give due credit would amount to plagiarism and thus be punishable.

- Services built around data citation will make it easier for researchers to discover relevant datasets.

- Data citations could be used to measure the impact of both individual datasets and their contributors.

- Researchers could gain professional recognition and rewards for published data in the same way as for more traditional publications.

Taking these points together, there would likely be an increase in the quantity and quality of data published, with all the benefits this implies for the transparency and rate of scientific research.
WHAT MAKES A GOOD CITATION?

Criteria for a good citation

- Location
- Attribution
- Identification
  - Which one do you mean?
  - Is this the right one?
- Resilience
  - against typographical error
  - against changing circumstances
- Relevance
  - Is it worth my time looking this one up?
Citation styles

Four data citation styles: which elements do they use?

Altman and King (2007): Dataverse

Lawrence et al. (2011): BADC

Green (2010): OECD

Starr and Gastl (2011): DataCite
Citation styles

**Author (Creator)**

Altman and King (2007): Dataverse
- Sidney Verba.

Lawrence et al. (2011): BADC
- Iwi, A. and B. N. Lawrence

Green (2010): OECD
- OECD

Starr and Gastl (2011): DataCite
- Irino, T; Tada, R
Citation styles

Publication date

Altman and King (2007): Dataverse
- NORC [Producer];

Lawrence et al. (2011): BADC

Green (2010): OECD
- OECD (2009),

Starr and Gastl (2011): DataCite
- Irino, T; Tada, R (2009):
Citation styles

Altman and King (2007): Dataverse

Title

Altman and King (2007): Dataverse


Lawrence et al. (2011): BADC

Title

Lawrence et al. (2011): BADC


Green (2010): OECD

Title

Green (2010): OECD

OECD (2009), “Key short-term indicators”, Main Economic Indicators

Starr and Gastl (2011): DataCite

Title

Starr and Gastl (2011): DataCite

Irino, T; Tada, R (2009): Chemical and mineral compositions of sediments from ODP Site 127-797.
Citation styles

Altman and King (2007): Dataverse

Lawrence et al. (2011): BADC

Green (2010): OECD
  › OECD (2009), “Key short-term indicators”, Main Economic Indicators

Starr and Gastl (2011): DataCite
  › Irino, T; Tada, R (2009): Chemical and mineral compositions of sediments from ODP Site 127-797. V.2.
Citation styles

Feature

Altman and King (2007): Dataverse

NORC [Producer];

Lawrence et al. (2011): BADC


Green (2010): OECD

OECD (2009), “Key short-term indicators”, Main Economic Indicators

Starr and Gastl (2011): DataCite

Citation styles

Resource type

Altman and King (2007): Dataverse

Sidney Verba. 1998. “U.S. and Russian Social and Political Participation Data,” NORC [Producer]; data set [Type (DC)]

Lawrence et al. (2011): BADC


Green (2010): OECD

OECD (2009), “Key short-term indicators”, Main Economic Indicators (database).

Starr and Gastl (2011): DataCite

Citation styles

Altman and King (2007): Dataverse


Lawrence et al. (2011): BADC


Green (2010): OECD

OECD (2009), “Key short-term indicators”, Main Economic Indicators (database).

Starr and Gastl (2011): DataCite

Citation styles

Identifier

Altman and King (2007): Dataverse

NORC [Producer]; data set [Type (DC)] ICPSR [Distributor].

Lawrence et al. (2011): BADC


Green (2010): OECD


Starr and Gastl (2011): DataCite

Geological Institute, University of Tokyo. Dataset. doi:10.1594/PANGAEA.726855.
Citation styles

Location

Altman and King (2007): Dataverse

  NORC [Producer]; data set [Type (DC)] ICPSR [Distributor].

Lawrence et al. (2011): BADC

  [Available from http://badc.nerc.ac.uk/data/coapec500yr].

Green (2010): OECD

  http://dx.doi.org/10.1787/data-00039-en

Starr and Gastl (2011): DataCite

  http://dx.doi.org/10.1594/PANGAEA.726855
Citation styles

Altman and King (2007): Dataverse


Lawrence et al. (2011): BADC


Green (2010): OECD


Starr and Gastl (2011): DataCite

Citation styles

**Unique Numeric Fingerprint**

Altman and King (2007): Dataverse

  UNF:3:ZNQRI14053UZq389x0Bffg?== NORC [Producer]; data set [Type (DC)] ICPSR [Distributor].

Lawrence et al. (2011): BADC

  [Available from http://badc.nerc.ac.uk/data/coapec500yr].

Green (2010): OECD


Starr and Gastl (2011): DataCite

  Geological Institute, University of Tokyo. Dataset. doi:10.1594/PANGAEA.726855.
  http://dx.doi.org/10.1594/PANGAEA.726855
Key citation elements

› Author
› Publication date
› Title
› Publisher
› Location/Identifier
Criteria for a good citation

- Location – **Location**
- Attribution – **Author**
- Identification – **Identifier**
- Resilience – **Publisher**
- Relevance – **Title, Date**
Note on Locators vs. Identifiers

From *Data Citation Guidelines for Earth Science Data*:

“Confusingly, a Digital Object Identifier is a locator.”
Note on Locators vs. Identifiers

From *Data Citation Guidelines for Earth Science Data*:

“Confusingly, a Digital Object Identifier is a locator.

> 10.15125/BATH-00069
Note on Locators vs. Identifiers

From *Data Citation Guidelines for Earth Science Data*:

> Confusingly, a Digital Object Identifier is a locator.

> 10.15125/BATH-00069
> https://doi.org/10.15125/BATH-00069
Note on Locators vs. Identifiers

From *Data Citation Guidelines for Earth Science Data*:

"Confusingly, a Digital Object Identifier is a locator.

- 10.15125/BATH-00069
- http://dx.doi.org/10.15125/BATH-00069

identifier
locator?
Note on Locators vs. Identifiers

From *Data Citation Guidelines for Earth Science Data*:

“Confusingly, a Digital Object Identifier is a locator.”

- 10.15125/BATH-00069
- https://hdl.handle.net/10.15125/BATH-00069
Note on Locators vs. Identifiers

From *Data Citation Guidelines for Earth Science Data*:

“Confusingly, a Digital Object Identifier is a locator.

- 10.15125/BATH-00069
- https://hdl.handle.net/10.15125/BATH-00069

curl -LH "Accept: text/x-bibliography; style=apa"
https://doi.org/10.15125/BATH-00069
Note on Locators vs. Identifiers

From *Data Citation Guidelines for Earth Science Data*:

Confusingly, a Digital Object Identifier is a locator.

- 10.15125/BATH-00069
- https://hdl.handle.net/10.15125/BATH-00069

`curl -LH "Accept: application/vnd.citationstyles.csl+json" https://doi.org/10.15125/BATH-00069`
Note on Locators vs. Identifiers

From *Data Citation Guidelines for Earth Science Data*:

“Confusingly, a Digital Object Identifier is a locator.”

- 10.15125/BATH-00069
- https://hdl.handle.net/10.15125/BATH-00069

```
curl -LH "Accept: application/vnd.datacite.datacite+xml" https://doi.org/10.15125/BATH-00069
```

**identifier**

**locator?**

**full metadata**
Note on Locators vs. Identifiers

From *Data Citation Guidelines for Earth Science Data*:

“Confusingly, a Digital Object Identifier is a locator.

- 10.15125/BATH-00069
- https://hdl.handle.net/10.15125/BATH-00069

```
curl -LH "Accept: application/vnd.datacite.datacite+xml"
https://doi.org/10.15125/BATH-00069
```

From *DOI Handbook*:

“A DOI name is an identifier (not a location) of an entity on digital networks.”
Note on Locators vs. Identifiers

From *Data Citation Guidelines for Earth Science Data*:

“There is no guarantee that the object at the [DOI’s] registered location will remain unchanged.”
Note on Locators vs. Identifiers

From *Data Citation Guidelines for Earth Science Data*:

“There is no guarantee that the object at the [DOI’s] registered location will remain unchanged.”

From UK Data Service DOI landing pages:

“A new Digital Object Identifier (DOI) is assigned to the data collection each time there is a major change to data, documentation or metadata.”

From the DOI Allocation Agreement between BL and U of Bath:

“Once a DOI Name has been registered [. . . ], the Registrant acknowledges and agrees that the DOI Name and the Referent to which that DOI Name relates shall not be changed.”
Note on Locators vs. Identifiers

From *Data Citation Guidelines for Earth Science Data*:

“Multiple versions or representations of the same data set may be located through one DOI.”

“Two identical data sets may be provided and managed by two different repositories and have different DOIs.”
Note on Locators vs. Identifiers

From *Data Citation Guidelines for Earth Science Data*:

“Multiple versions or representations of the same data set may be located through one DOI.

“Two identical data sets may be provided and managed by two different repositories and have different DOIs.

From *DOI Handbook* (emphasis mine):

“Each DOI name shall specify **one and only one referent** in the DOI system. While a referent *can* be specified by more than one DOI name, it is **recommended** that each referent has **only one DOI** name.
Zenodo model for DOIs and versioning

FRBR
Work

FRBR
Expressions

See
Tillet (2004)

From http://help.zenodo.org/#versioning:

You should normally always use the DOI for the specific version of your record in citations.
Zenodo model for DOIs and versioning

From http://help.zenodo.org/#versioning:

“You should normally always use the DOI for the **specific version** of your record in citations.”
Dynamic datasets

1. Differentiate versions by access date rather than ID
   A
   B
   C

2. Take time slices
   A
   B
   C

3. Take snapshots
   A
   B
   C
Dynamic datasets

4. Apply version control and log queries (Rauber et al. 2015)
Thank you for your attention
Bibliography (1)


Rauber, A. et al. (2015), *Data Citation of Evolving Data: Recommendations of the Working Group on Data Citation (WGDC)*, (Research Data Alliance, 20 Oct.). doi: 10.15497/RDA00016.
