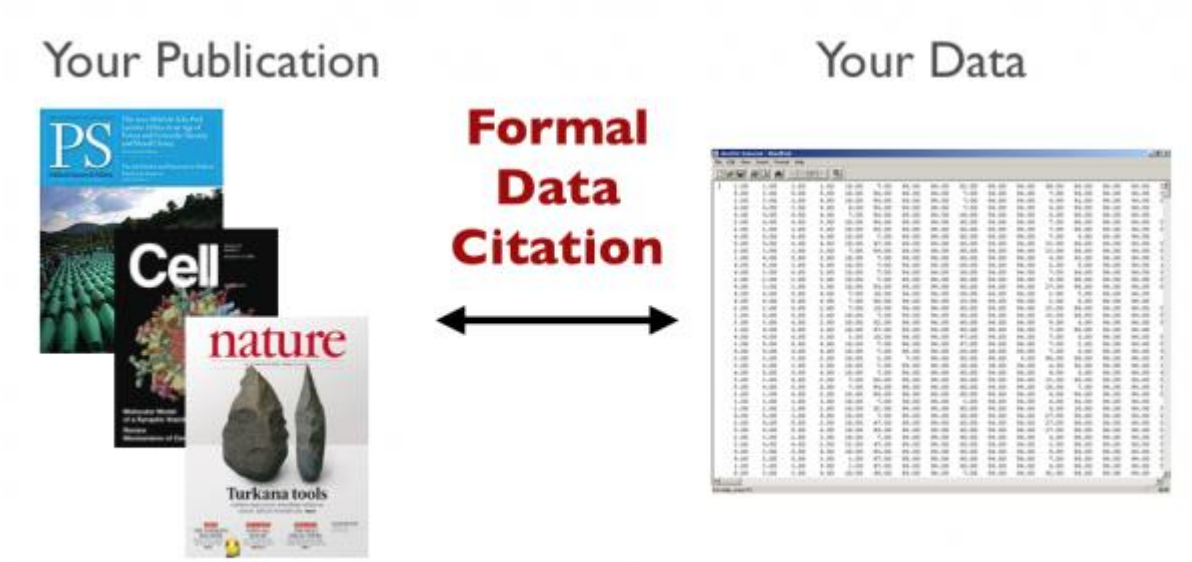
Data Citation



Dataverse standardizes the citation of datasets to make it easier for researchers to publish their data and get credit as well as **recognition** for their work. When you create a dataset in Dataverse, the citation is generated and presented automatically. As an open source framework and research data repository, Dataverse is committed to helping researchers, journals, and organizations make scientific data accessible, reusable, and open (when possible), which includes implementing community accepted standards for data publication (Altman & Crosas 2013). For nearly 20 years, members of IQSS and its Data Science team, who work on Dataverse, have played an active part in the the work to standardize data citation (King 1995, Altman & King 2007, Altman & Crosas 2013). Illustrated in the figure below, is an example of how the data citation is formulated in Dataverse, using the Joint Declaration of Data Citation Principles (2014) : a synthesis of all previously existing principles and initiatives on data citation..

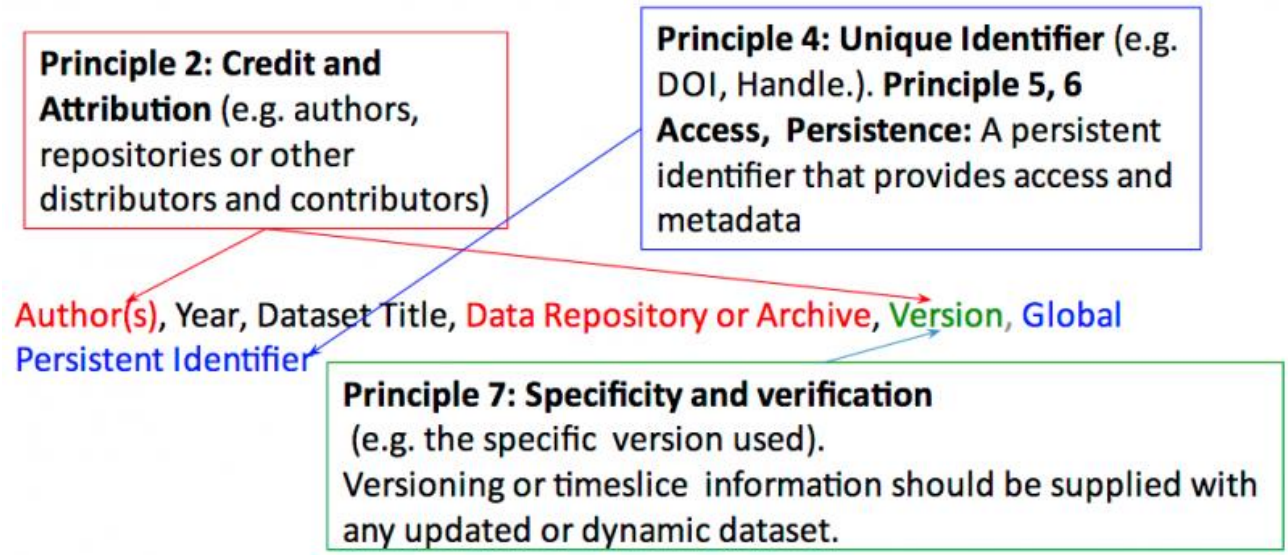


Fig. 1 Example of a Data Citation based on the the Joint Declaration of Data Citation Principles (2014).

In addition to getting recognition with a citation you can also make your particular Dataverse recognizable by setting up your own branding in Dataverse Theme + Widgets.

References

1. King, Gary. 1995. Replication, Replication. PS: Political Science and Politics 28: 443–499. Link.
2. Altman, Micah, and Gary King. 2007 . "A proposed standard for the scholarly citation of quantitative data." *D-lib Magazine* 13.3/4. Link.
3. Altman, Micah, and Mercè Crosas. 2013. "The Evolution of Data Citation: From Principles to Implementation." *IASSIST Quarterly* 2013;37. Link.

Data Citation Standard

The citation standard defined here offers proper recognition to authors as well as permanent identification through the use of global, persistent identifiers in place of URLs, which can change frequently. Use of universal numerical fingerprints (UNFs) guarantees to the scholarly community that future researchers will be able to verify that data retrieved is identical to that used in a publication decades earlier, even if it has changed storage media, operating systems, hardware, and statistical program format.

Following are two authentic examples of replication data citations:

From International Studies Quarterly, King and Zeng, 2006, p. 209:

Gary King; Langche Zeng, 2006, "Replication data for: When Can History be Our Guide? The Pitfalls of Counterfactual Inference", Harvard Dataverse, V2, http://hdl.handle.net/1902.1/DXRXCFAWPK UNF:3:DaYlT6QSX9r0D50ye+tXpA==

From Political Analysis, Hanmer, Banks, and White, 2013:

Hanmer, Michael J.; Banks, Antoine J., White, Ismail K., 2013, "Replication data for: Experiments to Reduce the Over-reporting of Voting: A Pipeline to the Truth", Harvard Dataverse, V1, http://dx.doi.org/10.7910/DVN/22893 UNF:5:eJOVAjDU0E0jzSQ2bRCg9g==

This citation has seven components. Five are human readable: the author(s), title, year, data repository (or distributor), and version number. Two components are machine-readable:

1. Of the machine-readable components to these citations, the unique global identifier begins with either "hdl" (this refers to the international HANDLE.NET system) or "doi" (this refers to a Digital Object Identifier (DOI) system). This identifier is designed to persist even if URLs–or the web itself–are replaced with something else. When the citation appears online, the identifier is hot-linked to the URL that references the identifier, which works in browsers available today. In print, the URL is also included in the citation.
2. The universal numerical fingerprint begins with "UNF". Four features make the UNF especially useful: The UNF algorithm's cryptographic technology ensures that the alphanumeric identifier will change when any portion of the data set changes. Not only does this assure future researchers that they can use the same data set referenced in a years-old journal article, it enables the data set's owner to track each iteration of the owner's research. When an original data set is updated or incorporated into a new, related data set, the algorithm generates a unique UNF each time. The UNF is determined by the content of the data, not the format in which it is stored. For example, you create a data set in SPSS, Stata or R, and five years later, you need to look at your data set again, but the data was converted to the next big thing (NBT). You can use NBT, recompute the UNF, and verify for certain that the data set you're downloading is

the same one you created originally. That is, the UNF will not change. Knowing only the UNF, journal editors can be confident that they are referencing a specific data set that never can be changed, even if they do not have permission to see the data. In a sense, the UNF is the ultimate summary statistic. The UNF's noninvertible, cryptographic properties guarantee that acquiring the UNF of a data set conveys no information about the content of the data. Authors can take advantage of this property to distribute the full citation of a data set–including the UNF–even if the data is proprietary or highly confidential, all without the risk of disclosure.

For information on how to implement the Universal Numerical Fingerprint (UNF), see:

- Technical UNF documentation in our Developers Guide
- Dr. Micah Altman's paper "A Fingerprint Method for the Verification of Scientific Data".

Learn more:

1. Micah Altman and Gary King. (2007). "A Proposed Standard for the Scholarly Citation of Quantitative Data," D-Lib Magazine, Vol. 13, No. 3/4 (March). Link.
2. Paul E. Uhlir, R., Board on Research Data, Information, Policy, Global Affairs, & National Research Council. (2012). For attribution – developing data attribution and citation practices and standards: Summary of an international workshop. The National Academies Press. Link