An Introduction to the Joint Principles for Data Citation

by Micah Altman, Christine Borgman, Mercè Crosas and Maryann Martone

EDITOR'S SUMMARY

While the conventions of bibliographic citation have been long established, the sole focus is on reference to other scholarly works. Access to the data serving as the basis for scholarly work has been limited. Data citation extends important access to material that has been largely unavailable for sharing, verification and reuse. The *Joint Declaration of Data Citation* Principles, finalized in February 2014, is a formal statement pulling together practices used in the research and publishing arenas and in common use. The declaration encompasses eight principles that stress the importance and legitimacy of data, the need to give scholarly credit to contributors and the importance of data as evidence. Cited data should have unique and persistent identifiers, and the citation should facilitate human and machine access and support verification and interoperability. The principles are expected to have broad impact on research, recognition and scholarly publishing.

KEYWORDS

data sets
data curation
access to resources
source materials
citation analysis

NOTE: This article summarizes and extends a longer report published as [1]. Contributors are listed in alphabetical order. We describe contributions to the paper using a standard taxonomy described in [2]. Micah Altman and Mercè Crosas were the lead authors, taking equal responsibility for revisions and authoring the first draft of the manuscript from which this is derived. All authors contributed to the conception of the Force 11 principles discussed, to the methodology, to the project administration and to the writing through critical review and commentary.

Data citation is rapidly emerging as a key practice supporting data access, sharing and reuse, as well as sound and reproducible scholarship. Consensus data citation principles, articulated through the *Joint Declaration of Data Citation Principles* [3], represent an advance in the state of the practice and a new consensus on citation.

Lowering the barrier to research data discovery and use, coupled with an increased ability to link data with publications, could enable new forms of scholarly publishing, promote interdisciplinary research, strengthen the linkage between policy and science and lower the costs of replicating and extending previous research. For this reason, the submission requirements for *Science* – one of the most cited, read and respected journals in the sciences – stipulate that "all data necessary to understand, assess and extend the conclusions of the manuscript must be available to any reader of *Science*" and that "*citations to unpublished data* [emphasis added] and personal communications cannot be used to support claims in a published paper" [4]. Too often, however, this proscription and others like it have been honored only in the breach. Few research articles provide access to the data on which they are based, nor specific citations to data on which the findings rely, nor protocols, algorithms, code or other technology necessary to reproduce, reuse or extend results.

The practice of bibliographic citation to supporting materials was formalized in scholarly publishing more than a century ago. In this tradition, a "bibliographic citation" refers to a formal, structured reference to another scholarly work. In most fields, citations are made in the body of the work. Full references typically appear at the end of the main text, providing more detailed bibliographic information for each work referenced. Following the establishment of the first scientific digital data archives in the late 1960s, bibliographic standards for data were developed and refined over the next decades but never widely used in practice.

The theory and practice of data citation have advanced considerably over the last five years, and these parallel efforts led to concern for a unified approach.

Micah Altman is director of research and head/scientist, program on information science for the MIT Libraries. He can be reached at escience<at>mit.edu. | Christine Borgman is professor and Presidential Chair in information studies at UCLA. She can be reached at borgman<at>gseis.ucla.edu. |Mercè Crosas is director of data science at the Institute for Quantitative Social Science at Harvard University. She can be reached at mcrosas<at>iq.harvard.edu. | Maryann Martone is co-director of the National Center for Microscopy and Imaging Research at the UCSD. She can be reached at maryann<at>ncmir.ucsd.edu.

The joint principles on data citation represent a new phase of activity that focuses on principled integration with the scholarly research and publishing ecosystem and a broad consensus on data citation practices. What has emerged in the publishing and research communities is agreement that citation is needed to support attribution and verification, recognition that citations must support both human and machine clients, maturity of robust persistent identifiers, and the desire to integrate data citation in standardized ways within publications, catalogs, tool chains and larger systems of attribution.

In the summer of 2013, the Data Citation Synthesis Group was formed to unify various parallel recommendations. Meeting weekly from July to November of 2013, the group thoroughly deconstructed previous data citation principles and produced a synthesis set that included the input of more than 25 organizations. The group also met in September 2013 as part of the Research Data Alliance conference, in two half days of public workshop. As a result, in November 2013, the proposed *Joint Declaration of Data Citation Principles* [3] was released to the public for open comment and then finalized at the end of February 2014. The joint principles reference and synthesize principles and recommendations from earlier work: a report by the CODATA/ICTSI Task Force on Data Citation [5] and a workshop held by the National Research Council [6] as well as standards previously proposed by Altman & King [7] and Ball & Duke [8].

The scope of the principles is solely to provide data citation recommendations, not to include detailed specifications for implementation or to focus on technologies or tools or research data repositories. The principles should extend to all disciplines and all types of data. As will be seen below, the *Joint Declaration of Data Citation Principles* reflects various efforts and presents a broad convergence on eight core principles:

1. **Importance.** Data should be considered legitimate, citable products of research. Data citations should be accorded the same importance in the scholarly record as citations of other research objects, such as publications.
2. **Credit and Attribution.** Data citations should facilitate giving scholarly credit and normative and legal attribution to all contributors to the data, recognizing that a single style or mechanism of attribution may not be applicable to all data.
3. **Evidence.** In scholarly literature, whenever and wherever a claim relies upon data, the corresponding data should be cited.
4. **Unique Identification.** A data citation should include a persistent method for identification that is machine actionable, globally unique and widely used by a community.
5. **Access.** Data citations should facilitate access to the data themselves and to such associated metadata, documentation, code and other materials as are necessary for both humans and machines to make informed use of the referenced data.
6. **Persistence.** Unique identifiers, and metadata describing the data and its disposition, should persist – even beyond the lifespan of the data they describe.
7. **Specificity and Verifiability.** Data citations should facilitate identification of, access to and verification of the specific data that support a claim. Citations or citation metadata should include information about provenance and fixity sufficient to facilitate verifying that the specific time slice, version and/or granular portion of data retrieved subsequently is the same as was originally cited.

CONTENTS

TOP OF ARTICLE

< PREVIOUS PAGE

NEXT PAGE >

8. **Interoperability and flexibility.** Data citation methods should be sufficiently flexible to accommodate the variant practices among communities but should not differ so much that they compromise interoperability of data citation practices across communities.

The Force 11 website [9] currently hosts the data citation principles and includes examples, detailed documentation and references to the standards and reports they incorporate.

Less than a month after the principles were finalized, they were endorsed officially by 30 organizations (now 85, as of late 2014), including many major publishers and data archives.

Several data repositories and systems are already compliant, or close to being compliant, with these principles (for example, Dataverse, DataDryad, DataCite). We anticipate that the impact of the unified, widely disseminated *Joint Declaration of Data Citation Principles* will be substantial and will change current publication workflows, create new data citation technologies, define new metrics for scholarly impact and recognition and, more importantly, provide persistent access to the data supporting scientific results to validate and extend previous scientific work. The principles will facilitate interoperability across existing and new implementations [10] and will help guide enhancements and new versions of the current implementations. ■

Resources Mentioned in the Article

[1] Altman, M. & Crosas, M. (2013). The evolution of data citation: From principles to implementation. *IASSIST Quarterly*, 63. Retrieved from www.iassistdata.org/downloads/iqvol371_4_altman.pdf

[2] Allen, L., Brand, A., Scott, J., Altman, M., & Hlava, M. (2014). Credit where credit is due. *Nature*, 508, 312-313. doi:10.1038/508312a

[3] Data Citation Synthesis Working Group. (February 2014). *Joint Declaration of Data Citation Principles – Final*. Force 11. Retrieved from www.force11.org/datacitation

[4] Science. (2015). General information for authors. Retrieved from www.sciencemag.org/site/feature/contribinfo/prep/gen_info.xhtml#dataavail

[5] CODATA/ITSCI Task Force on Data Citation. (2013). Out of cite, out of mind: The current state of practice, policy and technology for data citation. *Data Science Journal*, 12, 1-75. doi:10.2481/dsj.OSOM13-043

[6] Uhlir, P. (Ed.). (2012). *For attribution – developing data attribution and citation practices and standards.* Washington D.C.: The National Academies Press.

[7] Altman, M., & King, G. (2007). A proposed standard for the scholarly citation of quantitative data. *D-lib Magazine, 13*(3/4). doi:10.1045/march2007-altman

[8] Ball, A., & Duke, M. (2012, 21 June). Data citation and linking. *DCC Briefing Papers*. Edinburgh: Digital Curation Centre. Retrieved from www.dcc.ac.uk/resources/briefing-papers/introduction-curation/data-citation-and-linking

[9] Force 11: www.force11.org

[10] Starr, J., Castro, E., Crosas, M., Dumontier, M., Downs, R.R., Duerr, R., Haak, L.L., . . . Clark, T. (2014, forthcoming). Achieving human and machine accessibility of cited data in scholarly publications. *PeerJ*.