# Perspectives on Data Citation:
# Building Data Citations for Discovery

---

Hailey Mooney
Data Services and Reference Librarian
Michigan State University Libraries
mooneyh@msu.edu

Mark Newton
Digital Collections Librarian
Purdue University Libraries
newton@purdue.edu



IASSIST Annual Conference – June 1, 2011

# Data citations in context

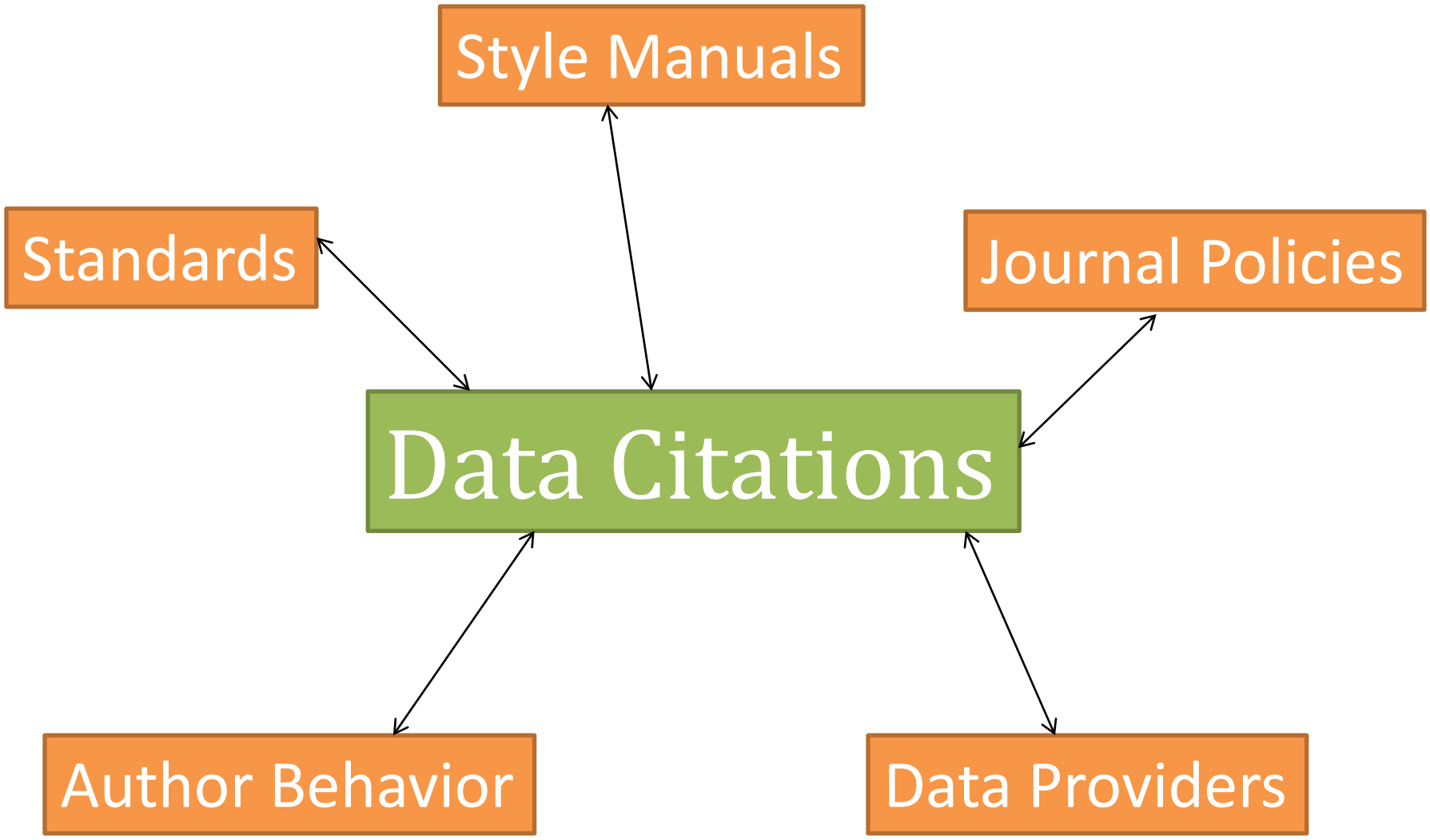
- Citations for data enable reuse, sharing, publishing

## Recommendations to Editors of Scientific Journals…

"Giving appropriate credit to data collectors should serve to encourage others to share data as a matter of good scientific practice. Criticism of the original data collection should be factual, temperate, and made in light of reasonable standards of data collection.

*Recommendation 12. Journals should require full credit and appropriate citations to original data collections in reports based on secondary analysis*."

Fienberg, S.E., Martin, M.E., & Straff, M.L. (1985). *Sharing research data*. Washington, DC: National Academy Press. p. 31.


Style Manuals
Standards
Journal Policies
Data Citations
Author Behavior
Data Providers

Our study compares the elements of recommended data citations with actual citations in published articles drawn. It seeks to help us understand what guidance is offered to authors who want to cite data and how authors actually compose these citations.

# Guiding Questions

- What are the key elements needed in a data citation and what constitutes an adequate data citation?
- How do the sources of citation guidance (publications & providers) instruct authors?
- How do actual citations in journal articles measure against instruction and best practice?

# Data Citation Adequacy Index (DCAI)

1. Assess Existing Studies and Standards
   - DataCite
   - OECD
   - Sieber & Trumbo
   - Altman & King
   - Dodd
   - Citation Standards
2. Select Citation Elements for DCAI
3. Weight Elements to Determine Scoring

# Definitions of data for the purpose of citation

| DataCite (2011) | OECD (2009) | Altman & King (2007) | Sieber & Trumbo (1995) | Dodd (1979) | Mooney & Newton (2011) |
|---|---|---|---|---|---|
| **"Scientific research data on the Internet"** "Please note that in this document, the resource that is being described can be of any kind, but it is typically a dataset. We use the term 'dataset' in its broadest sense. We mean by it to include not only numerical data, but any other research data outputs." | **OECD data** "Taking these requirements into account, OECD is proposing to implement a metadata standard for publishing datasets, collections of datasets and individual tables." | **Quantitative data** "…no special restrictions on what constitutes a quantitative data set, a definition may be useful: A quantitative data set represents a systematic compilation of measurements intended to be machine readable. The measurements may be the result of scientific research or information produced by governments or others for any purpose, so long as it is systematically organized and described." | **Research data** Analysis of citations to General Social Survey data (machine readable numeric dataset) | **Social science numeric data** "Social science numeric data files make up a substantial body of information known in the generic sense as machine readable data files (MRDF)…" | **Digital research data** Any primary source in electronic format that is subject to (secondary) analysis. |

# Developing the DCAI

| | DataCite (2011) | OECD (2009) | Altman & King (2007) | Sieber & Trumbo (1995) | Dodd (1979) | ANSI/NISO Z39.29 (2005) | ISO 690-2010 | Mooney & Newton (2011) |
|---|---|---|---|---|---|---|---|---|
| Author | x | x | x | x | x | x | x | 2 |
| Title | x | x | x | x | x | x | x | 2 |
| Date | x | x | x | x | x | x | x | 2 |
| Publisher | x | | x | x | x | x | x | 2 |
| Location | | | | x | x | x | x | |
| Funder | | | | x | x | | | |
| Material designator | x | x | x | x | x | x | x | 1 |
| Notes | | | | x | x | | | |
| Edition | x | | x | | x | | | |
| URL | | x | x | | | x | x | 2 |
| DOI (persistent ID) | x | x | x | | | x | x | 1 |
| Accessed date | | x | | | | x | x | |
| Parent/series | | x | | | x | | x | |
| Study/Accession number | | | | | | | | |

# Drawing the Sample

(data OR dataset) <in> Abstract AND ("data bank" OR repository OR archive OR study OR studies OR empirical OR research OR obtain* OR retriev* OR use* OR analy*) <in> Abstract AND Feature Article <in> ARTICLE_TYPE AND Date: between 2010 and 2010 AND Limited to: PEER_REVIEWED

# Coding the Sample

"We analyzed data from the National Hospital Ambulatory Medical Care Survey (NHAMCS), an annual, national probability sample survey of hospital EDs conducted by the National Center for Health Statistics."

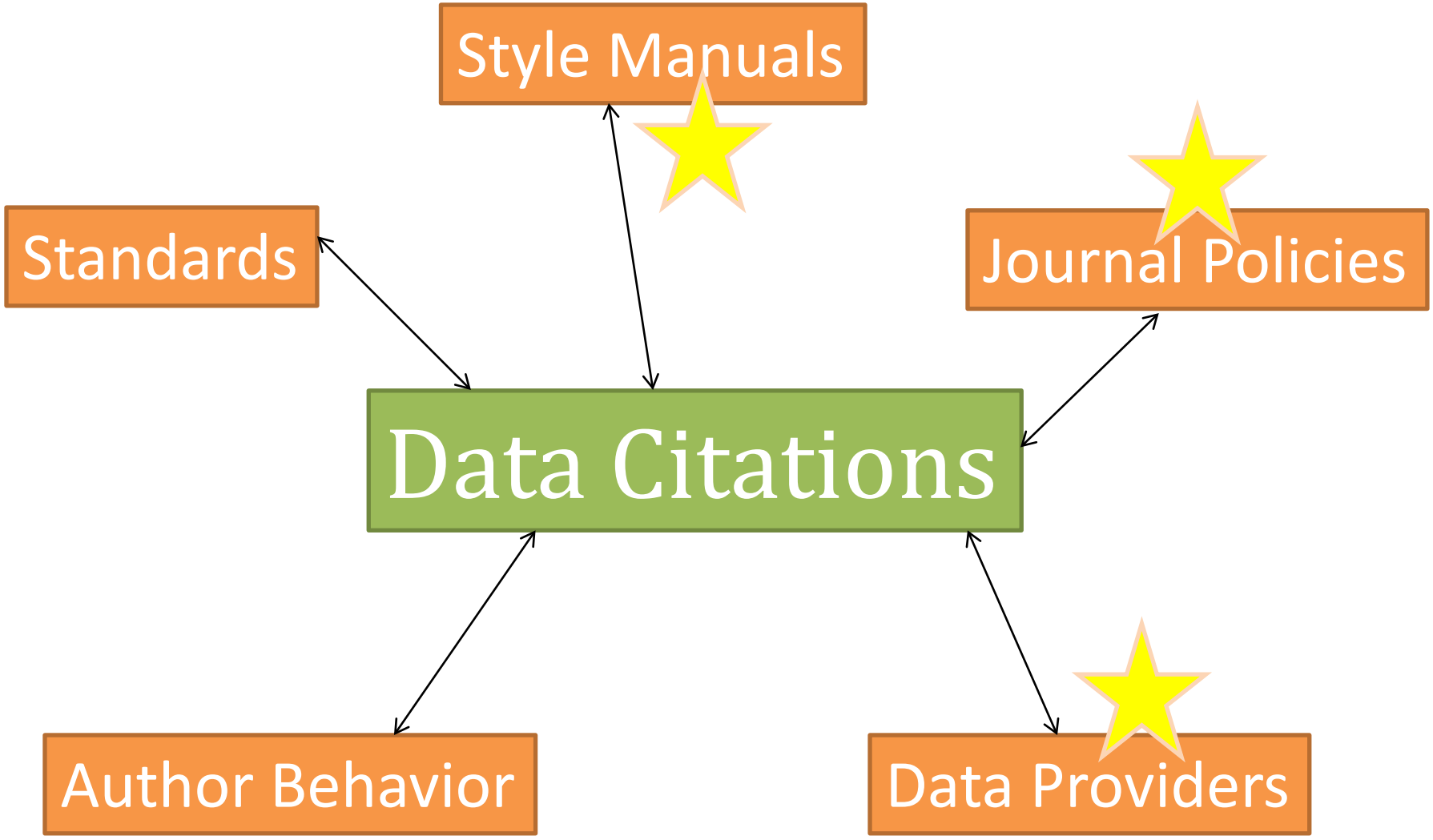
| | | |
|---|---|---|
| Author | Text | 1 × 2 |
| Title | Text | 1 × 2 |
| Date | None | 0 × 2 |
| Publisher | Text | 1 × 2 |
| Location | None | 0 × 2 |
| Persistent ID | None | 0 × 1 |
| Material Designator | None | 0 × 1 |

# Counts of Citation Elements by Location

Primary Citation Location and Location Code

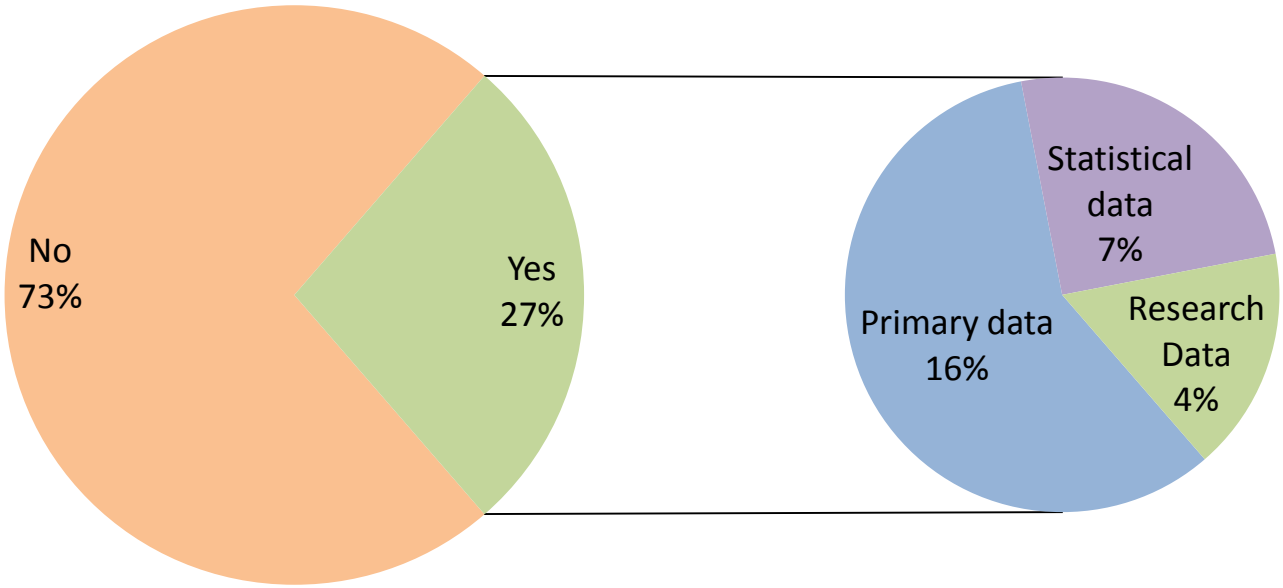
| Element and Code | [0] None | [1] Text | [2] Notes | [3] References |
|---|---|---|---|---|
| Author [2] | 40 (61.5) | 14 (21.5) | 0 (--.-) | 11 (16.9) |
| Title [2] | 6 (09.2) | 45 (69.2) | 3 (04.6) | 11 (16.9) |
| Date [2] | 54 (83.1) | 0 (--.-) | 0 (--.-) | 11 (16.9) |
| Publisher [2] | 38 (58.5) | 15 (23.1) | 3 (04.6) | 9 (13.8) |
| Material Designator [1] | 61 (93.9) | 1 (01.5) | 0 (--.-) | 3 (04.6) |
| Electronic Retrieval Location [2] | 53 (81.5) | 8 (12.3) | 2 (03.1) | 2 (03.1) |
| Persistent Identifier [1] | 65 (100.0) | 0 (--.-) | 0 (--.-) | 0 (--.-) |

# Our study: Part 2 – Instructions to authors



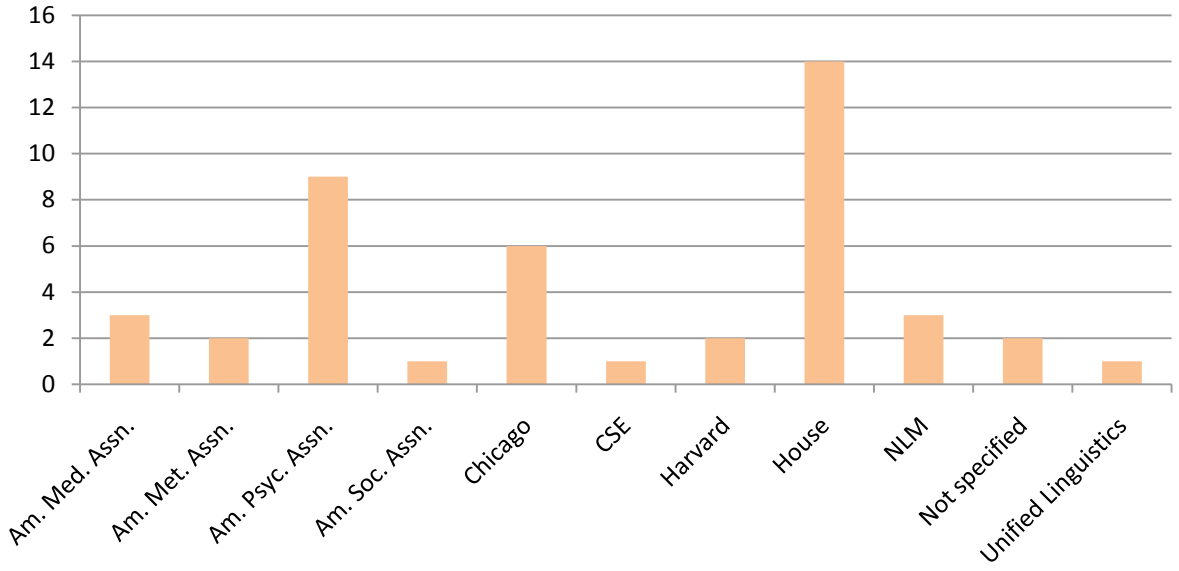

Literary and Linguistic Computing
Environmental Health Perspectives
Journal of the American Geriatrics Society
The Journal of Child Psychology and Psychiatry
The Wilson Journal of Ornithology
Journal of the American Dietetic Association
Annals of the Association of American Geographers
Journal for the Scientific Study of Religion
History
The Economic Journal
The Manchester School
Linguistics
The Lancet
Cell
Science
Challenge
The Sociological Quarterly
Journal of Community Health
Sociology of Religion
Nature
The American Journal of Human Genetics
PNAS
Journal of the Atmospheric Sciences
Presidential Studies Quarterly
Land Economics
Journal of Anthropological Archaeology
Journal of Family Issues
The Journals of Gerontology: Series B
Journal of Climate
Near Eastern Archeaology
Social Science & Medicine
Mexican Studies
Demography
Journal of the American Medical Association
American Journal of Public Health
Economic Development and Cultural Change
Journal of Marriage and Family
English Studies
Health and Social Work
Journal of Family
Journal of Communication
Journal of the American Planning Association
Journal of Criminal Justice
Medicine & Science in Sports & Exercise
Integrative and Comparative Biology

**Data citation addressed in journal author instructions document (n=44)**



## Style Manuals used by Journals

**(n=44)**




Citing Medicine: The NLM Style Guide for authors, editors, and publishers

American Sociological Association Style Manual

Publication Manual of the American Psychological Association

house style

American Meteorological Association Author Reference and Citation Guide

Uniform Requirements for Manuscripts submitted to Biomedical Journals

Unified style sheet for linguistics

not specified

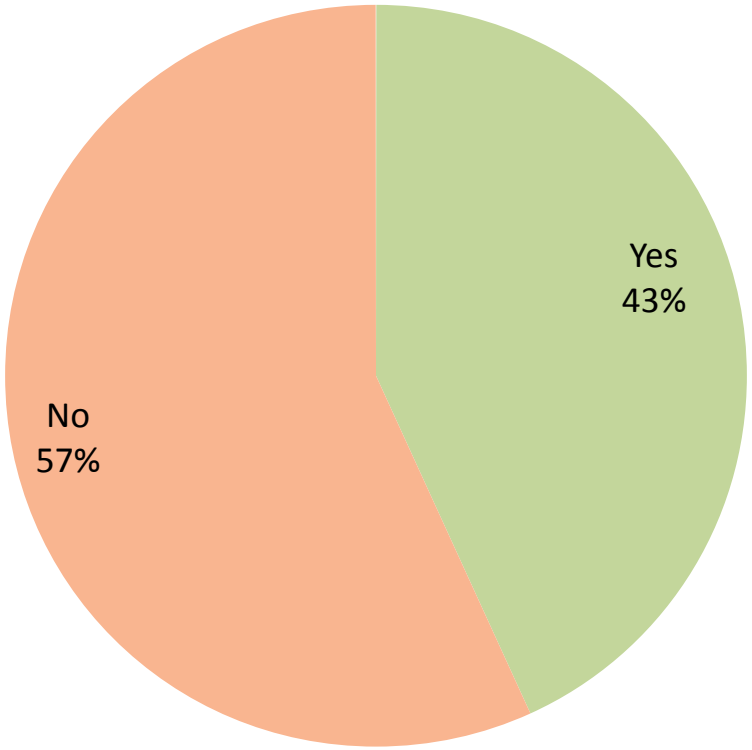
Chicago Manual of Style

Harvard

American Medical Association Manual of Style

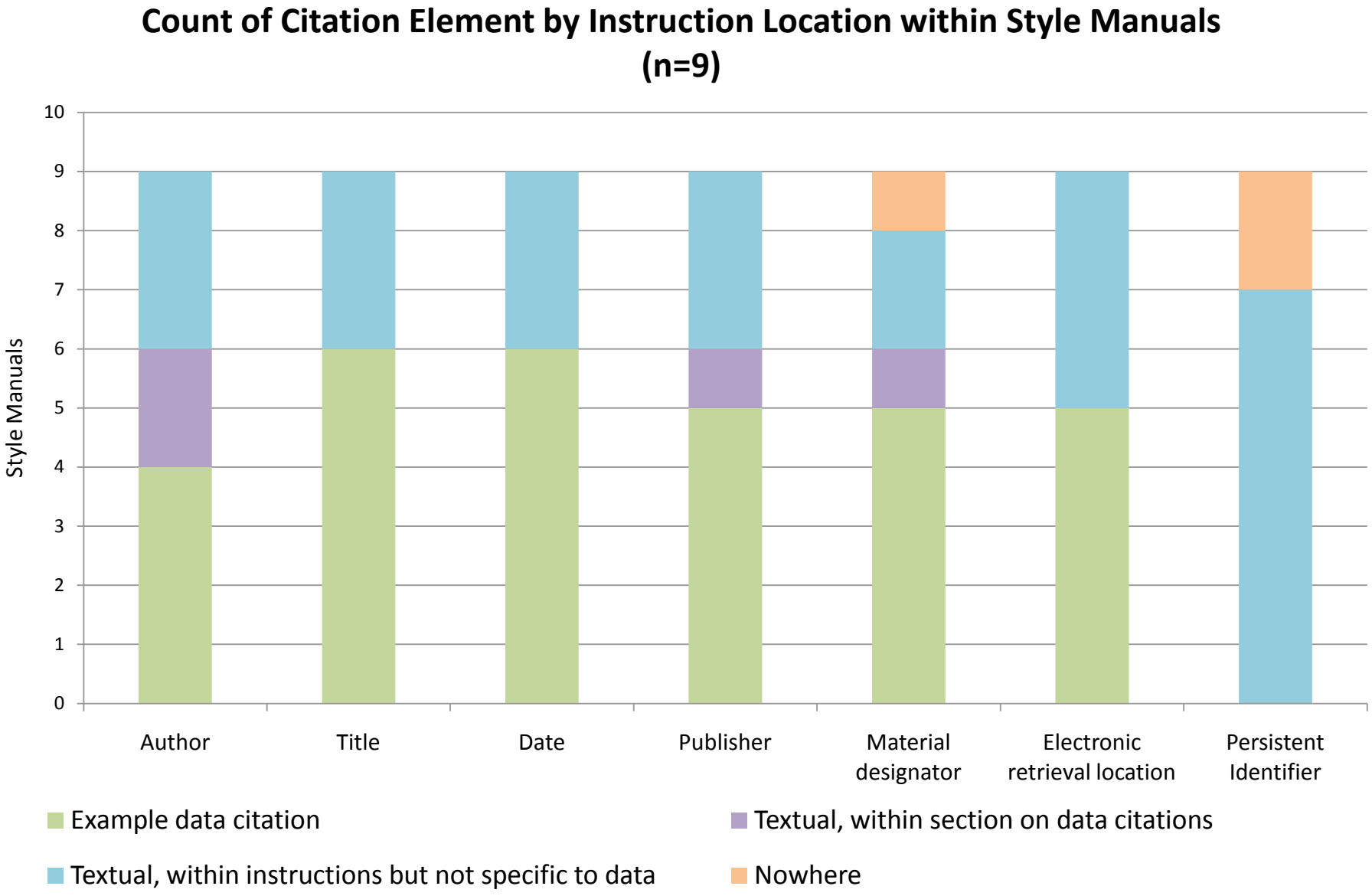
Scientific style and format: the CSE manual for authors, editors, and publishers

**Data citation addressed in Style Manual used by journals (n=44)**



# Data in Style Guides

| Style Manual | Data Type | Example Data Citation |
|---|---|---|
| Am. Med. Assn. | **Databases** | PDQ: NCI's Comprehensive Cancer Database. Bethesda, MD: National Cancer Institute; 1996. http://www.cancer.gov/cancerinfo/pdq/cancerdatabase. Updated December 18, 2001. Accessed April 29, 2004 |
| Am. Met. Assn. | **Digital media/NSIDC data** | Jackson, T. J., and M. H. Cosh, 2003: SMEX02 watershed soil moisture data, Walnut Creek, Iowa. National Snow and Ice Data Center, Boulder, CO, digital media. [Available online at http://nsidc.org/data/nsidc-0143.html.] |
| Am. Psyc. Assn. | **Data Sets** | Pew Hispanic Center. (2004). Changing channels and crisscrossing cultures: A survey of Latinos on the news media [Data file and code book]. Retrieved from http://pewhispanic.org/datasets/ |
| Am. Soc. Assn. | **Machine Readable Data Files** | American Institute of Public Opinion. 1976. Gallup Public Opinion Poll #965 [MRDF]. Princeton, NJ: American Institute of Public Opinion [producer]. New Haven, CT: Roper Public Opinion Research Center, Yale University [distributor]. |
| Chicago | None | None |
| CSE | **Databases on the Internet** | IMGT/HLA Sequence Database [Internet]. Release 2.9.0 Cambridge (England): European Bioinformatics Institute. 2003 - [update 2005 Jun 1; cited 2005 Jun 22]. Available from http://www.ebi.ac.uk/imgt/hla/ |
| Harvard | None | None |
| House | None | None |
| NLM | **Part of a Database on the Internet** | Entrez Genome [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information. [date unknown] - . Haloarcula marismortui ATCC 43049 plasmid pNG200, complete sequence; [cited 2007 Feb 27]. Available from: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=genome&cmd=Retrieve&dopt=Overview&list_uids=18013 |
| Not specified | None | None |
| Unified Linguistics | None | None |


Count of Citation Element by Instruction Location within Style Manuals (n=9)
Style Manuals
10
9
8
7
6
5
4
3
2
1
0
Author
Title
Date
Publisher
Material designator
Electronic retrieval location
Persistent Identifier
Example data citation
Textual, within section on data citations
Textual, within instructions but not specific to data
Nowhere

# Data Providers

# National Elevation Dataset

Q: How should the NED be cited?

**A:** To cite the NED in a publication, please use the following **literature references**:

Gesch, D.B., 2007, The National Elevation Dataset, in Maune, D., ed., Digital Elevation Model Technologies and Applications: The DEM Users Manual, 2nd Edition: Bethesda, Maryland, American Society for Photogrammetry and Remote Sensing, p. 99-118.

Gesch, D., Oimoen, M., Greenlee, S., Nelson, C., Steuck, M., and Tyler, D., 2002, The National Elevation Dataset: Photogrammetric Engineering and Remote Sensing, v. 68, no. 1, p. 5-11.

http://ned.usgs.gov/Ned/faq.asp#CITED

# European Centre for Medium-Range Weather Forecasts
# 40-yr Re-Analysis (ERA-40) datasets

**Citation:**

Users of the ECMWF data sets are requested to reference the source of the data in any publication, e.g. "***ECMWF ERA-40 data used in this study/project have been provided by ECMWF/have been obtained from the ECMWF Data Server***".

http://www.ecmwf.int/products/data/archive/index.html#citation

# North American Breeding Bird Survey

**Please cite this Page as:**

Sauer, J. R., J. E. Hines, and J. Fallon. 2007. *The North American Breeding Bird Survey, Results and Analysis 1966 - 2006. Version 10.13.2007.* *USGS Patuxent Wildlife Research Center*, *Laurel, MD*

**URL is to main page for USGS Patuxent Wildlife Research Center, NOT the page this survey appears on.**

http://www.mbr-pwrc.usgs.gov/bbs/bbs2006.html

# DataCite

- Recommended minimum format

  Creator (Publication Year): Title. Publisher. Identifier

- With optional properties

  Creator (Publication Year): Title. Version. Publisher. Resource Type. Identifier

# Citation Issue: Publisher

## DataCite

- Definition of publisher
  - Holder of the data (e.g., archive) OR institution which submitted the work
- Example:
  Irino, T; Tada, R (2009): Chemical and mineral compositions of sediments from ODP Site 127-797. Geological Institute, University of Tokyo. doi:10.1594/PANGAEA.726855.
  - Dataset is held/archived by PANGAEA, but the publisher field shows the home institution of the authors

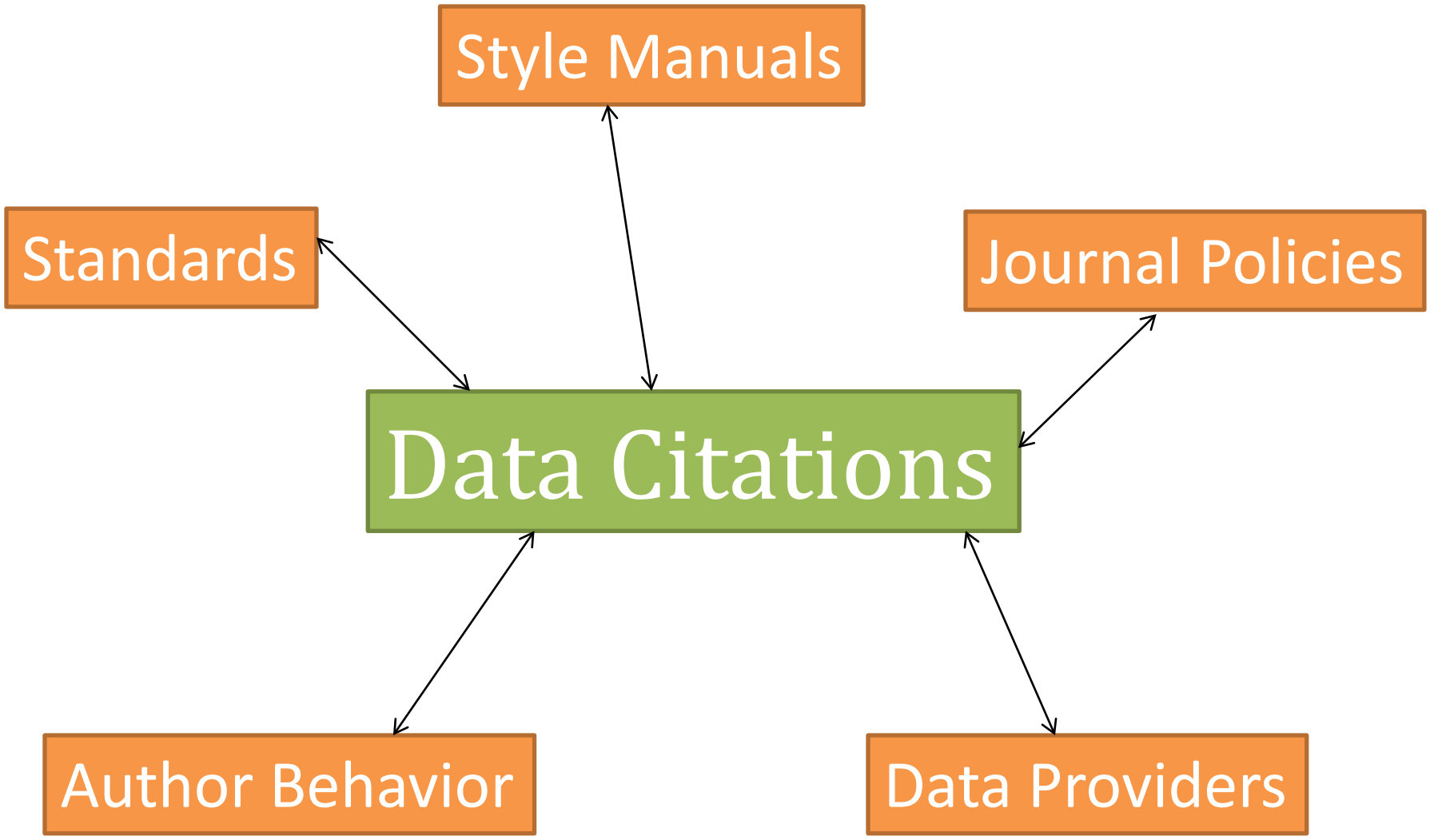
## Dodd

- Publisher -> Imprint
  - Producer: person/organization responsible for the actual encoding of the data into machine-readable form
  - Distributor: person/organization authorized to execute a machine-readable copy of the file
- Example
  Milberger, S. (2002). Evaluation of violence against women with physical disabilities in Michigan, 2000-2001 (ICPSR version) [data file]. Detroit, MI: Wayne State University [producer]. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor]. doi:10.3886/ICPSR03414

# Citation Issue: Publisher vs. Electronic Retrieval Location/Persistent Identifier

APA 6th

- Publisher replaced with URL/DOI
  - Omit publisher and replace with URL/DOI for electronic documents
  - Example:
    Pew Hispanic Center. (2004). *Changing channels and crisscrossing cultures: A survey of Latinos on the news media* [Data file and code book]. Retrieved from http://pewhispanic.org/datasets


Style Manuals
Standards
Journal Policies
Data Citations
Author Behavior
Data Providers

# Status-quo

- Lack of consistent data citations across the spectrum

# Solutions

- Increased recognition of data as a citable object
- Minimum recommended format for data citation widely recognized and adapted to many types of data

# Bibliography


- 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
- American Psychological Association. (2010). Publication manual of the American Psychological Association (6th ed.). Washington, DC: Author.
- DataCite Metadata Working Group. (January 2011). DataCite metadata scheme for the publication and citation of research data. (Version 2.0). doi:10.5438/0001
- DataCite Metadata Working Group. (March 2011) DataCite metadata scheme for the publication and citation of research data. (Version 2.1). doi:10.5438/0003
- Dodd, S. A. (1979). Bibliographic references for numeric social science data files: Suggested guidelines. *Journal of the American Society for Information Science, 30*(2), 77-82. doi: 10.1002/asi.4630300203
- Fienberg, S.E., Martin, M.E., & Straff, M.L. (1985). *Sharing research data*. Washington, DC: National Academy Press.
- Green, T. (2009). We need publishing standards for datasets and data tables *OECD Publishing White Paper*. Paris: OECD Publishing. doi:10.1787/603233448430
- International Organization for Standardization (2010). *ISO 690:2010. Information and documentation: Guidelines for bibliographic references and citations to information resources*. Geneva, Switzerland: Author.
- Mooney, H. (2011). Citing data sources in the social sciences: do authors do it? *Learned Publishing*, 24(2): 99-108. doi:10.1087/20110204
- National Information Standards Organization, & American National Standards Institute. (2005). *Bibliographic references*. Bethesda, MD: NISO Press. http://www.niso.org/kst/reports/standards?step=2&gid=None&project_key:ustring:iso-8859-1=87775a75d6ea19921a41d75b2fb012b0d6339b3a
- Newton, M., Mooney, H., & Witt, M. (December 2010). [Poster Session] *A description of data citation instructions in style guides*. International Digital Curation Conference. Chicago, Illinois. http://docs.lib.purdue.edu/lib_research/121/
- Sieber, J. E., & Trumbo, B. E. (1995). (Not) giving credit where credit is due: Citation of data sets. *Science and Engineering Ethics, 1*(1), 11-20. doi: 10.1007/BF02628694
- Starr, J., &Gastl, A. (2011). isCitedBy: A metadata scheme for DataCite. D-Lib Magazine, 17(1/2). doi:10.1045/january2011-starr