PLANT SYSTEMATICS WORLD

Edited by Vicki A. Funk

A LACK OF ATTRIBUTION: CLOSING THE CITATION GAP THROUGH A REFORM OF CITATION AND INDEXING PRACTICES

The last several decades have seen enormous changes in how science is conducted. Advances in computing power and the availability of massive data compilations have facilitated an increasing emphasis on integrative and broad-scale research that continues to produce fundamental scientific breakthroughs (Sidlauskas & al., 2009). However, these datasets and their resulting insights ultimately depend on primary data, which were laboriously collected and published by hundreds or even thousands of individual scientists over many decades. In fact, each of us has published studies using datasets compiled from hundreds of primary references. And community-wide efforts typically aggregate much larger volumes of data from the primary literature: for example, the Paleobiology Database (www.paleodb.org) includes data from over 42,000 primary sources and Fishbase (www.fishbase.org) includes data from more than 47,000 primary sources.

Given the size and scope of the resulting datasets, proper attribution has become a non-trivial challenge. Typically, database compilers are acknowledged or cited in the publications that use these compilations, but the primary data or publications that were aggregated in the database are not. Although primary sources are generally cited explicitly within databases by their compilers, subsequent database users typically do not reference these sources individually when publishing analyses of the database. And when they do, these references are generally relegated to supplementary online materials because of their sheer number. The lack of formal citation of the primary data sources is typically a result of such logistical obstacles and we do not mean to imply that authors of synthetic analyses are unappreciative of the enormous time, effort, and expertise required for the collection of primary data. In fact, for all of these reasons, many of our own previous studies fail to cite exhaustively the hundreds to thousands of primary sources from which the compiled data were derived (see, e.g., Kowalewski & al., 1998; Novack-Gottshall & Miller, 2003; Smith & al., 2004; Payne, 2005; Huntley & Kowalewski, 2007; Novack-Gottshall, 2007, 2008; Payne & Finnegan, 2007; Finnegan & al., 2008, 2011; Novack-Gottshall & Lanier, 2008; Payne & al., 2009; Villéger & al., 2011).

The magnitude of the citation gap is enormous. As an example, consider that the database of maximum body sizes of late Quaternary mammals (MOM; Smith & al., 2003) has been cited 125 times to date. The database itself was compiled from 283 primary sources. If all of the studies citing the MOM database used the entire database without referencing each of the source publications, this would constitute some 35,375 'missing' citations, most of them primary systematic studies. For larger compilations, each citation of the database may overlook tens of thousands of source studies. Given that these sources have generated hundreds of citations, this implies there may be millions of 'missing' citations in just the last decade.

The problem is pervasive. A simple search of Web of Science for the term 'meta-analysis' returns 45,957 references. We narrowed the search to 'Plant Science' and took the first ten for which we had full-text access. (We have deliberately refrained from citing the studies used here, as it is unfair to single out a small sample of studies chosen simply to illustrate a prevalent problem.) Of these ten studies, five cited all of the papers used in their regular reference list, four cited none, and one cited 12% (7 of 58). Of the five papers that cited few or none of the publications used in the meta-analysis within the regular reference list, only three included reference lists as supplementary material. In total, the ten papers used 2145 primary studies in their synthetic analyses. Of those publications, only 118 (5.5%) were cited, yielding a ratio of true to apparent citations of approximately 18:1 (Fig. 1).

The problem is even worse for analyses of large databases than for meta-analyses. We chose five recent publications using data from the Paleobiology Database to examine this issue quantitatively. Together, these five articles include 306 articles in their regular reference lists. Using supporting online data, we were able to directly identify the primary data sources for two of the studies. For the other three, we reconstructed the list of primary data sources by recreating searches of the Paleobiology Database using the search criteria enumerated in the studies. Together, the five studies of the Paleobiology Database used data from 3134 primary sources, of which only two (0.06%) were cited in the regular reference lists, yielding a ratio of true to apparent citations of approximately 1567:1 (Fig. 1).

The failure to formally cite the hundreds or thousands of primary data sources in broad-scale, synthetic analyses is a serious problem. It leaves numerous scientists unfairly uncredited, with detrimental long-term consequences for the health of foundational subdisciplines such as taxonomy, systematics, and natural history. Electronic search engines (e.g., PubMed, Web of Science, Scopus, Google Scholar) are increasingly used to provide quantitative assessment of the relative importance of journals, scientists, and

whole fields of inquiry (Abbott & al., 2010). Often, these statistics inform hiring, tenure, and funding decisions. Unfortunately, search engines cannot index citations to primary sources if they are not listed. Moreover, most do not index citations in supplementary online material, which is where primary sources used in analyses of large literature compilations are typically found when they are provided. This situation leads to a severe and pervasive undervaluation of the disciplines that disproportionately provide foundational data for science. The discrepancies between apparent and true impact negatively influences hiring and funding practices, to the detriment of many fields and of the scientific endeavor in general (McClain, 2011; Waegele & al., 2011).

To further highlight the effects of the citation gap, consider the following example. The mean citation rate in Web of Science over the past decade for a paper published in the *Journal of Paleontology* (a leading systematic journal in the geosciences) during the 1980s and 1990s is 0.49 citations per year. Of the 1894 articles published during those decades, 896 (47%) are included in the Paleobiology Database, which produced an average of 18 official papers per year between 2007 and 2011 and undoubtedly numerous studies where the authors did not apply for an official Paleobiology Database publication number. Even if only 10% of the official papers using the Paleobiology Database incorporated data from any given primary source, this represents more than a 2.6-fold difference between the apparent and true citation rate for the typical paper in the *Journal of Paleontology* (Fig. 1) and would have a substantial impact on the *h*-index and career citation statistics for an active systematist. The true gap may be much larger, depending upon the number of 'unofficial' publications using data from the Paleobiology Database. We examined two other prominent paleontological journals (*Journal of Vertebrate Paleontology* and *Palaeontology*) and the statistics are nearly identical. Yet, paleobiology is a comparatively small field; the differences between apparent and true citation rates have potentially even greater ramifications in larger data-rich disciplines such as ecology, genomics, and biogeography.

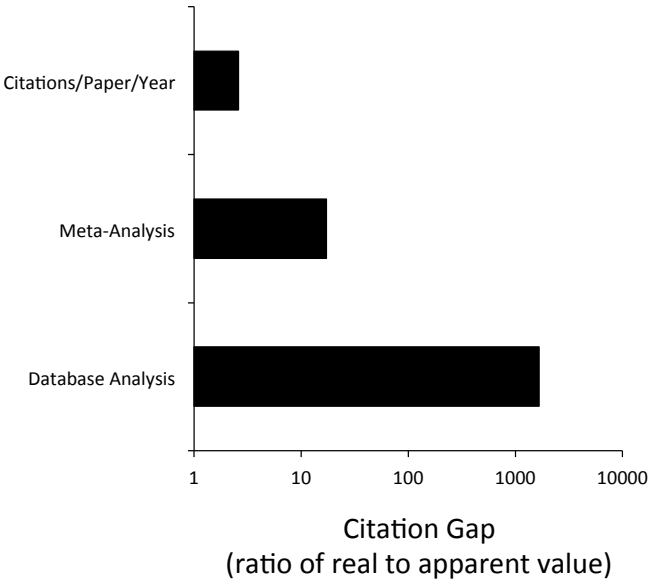

Fig. 1. Ratios between true and apparent citation rates for database analyses, meta-analyses, and taxonomic papers, illustrating the enormous undercitation of the taxonomic literature resulting from current citation and indexing practices. Database value based on a sample of recent studies using data from the Paleobiology Database. Meta-analysis value based on a sample of recent papers in the area of plant science. Citations rates for taxonomic papers based upon a study of papers published in the *Journal of Paleontology* between 1980 and 1999. See text for further details.

Despite its magnitude, the solution to this problem is easily implemented. Scientific journals and funding agencies must take a leadership role by: (1) requiring the use of explicit referencing for source data to ensure proper attribution, and (2) working with scientific index services to capture supplementary references in supporting online material. If supplementary reference lists are so long as to present logistical problems (e.g., many thousands of supporting references), alternative approaches may be required. However, we suspect that at present such cases are rare. Recent history indicates that solutions such as these are both feasible and desirable. In the area of data reporting, numerous journals in the biological and geological sciences (the fields with which we are most familiar) have adopted policies requiring the raw data for all studies to be archived online in permanent, publicly accessible sites such as Dryad (www.datadryad.org) and GenBank (www.ncbi.nlm.nih.gov/genbank/). These policies have been instrumental in making the results of scientific research more transparent and verifiable, while facilitating more efficient reanalysis of such data. With a similar motivation to make research results universally available, the United States National Institutes of Health have enacted a policy that all NIH-funded research papers must be archived at the publicly accessible site PubMed Central (http://www.ncbi.nlm.nih.gov/pmc/). Currently, over 1000 journals are participating directly with PubMed Central, which contains over 2.4 million archived articles. Thus, history shows that simple changes in policy from journals and governmental agencies can quickly produce dramatic and positive shifts in scientific culture, suggesting that the citation gap can also be closed quickly and efficiently.

The systematic archiving of data has greatly enhanced our understanding of pattern and process in a variety of complex and heterogeneous natural systems. However, ongoing growth and development of our fields depends not only upon advances in data storage and analysis but also on the ongoing collection of new, high quality, foundational data that refine the basic patterns and fill in gaps in the existing primary knowledge. Unfortunately, current citation and indexing practices do not accurately reflect the vital roles that many foundational disciplines play, and such unintended omissions are likely to result in the decline of the disciplines upon which the synoptic analyses depend. We urge that policies similar to those developed for the storage of data and the archiving of federally funded research studies be enacted by the National Science Foundation, other funding agencies, and major cross-disciplinary journals regarding citation of primary data sources. Unless such a solution is enacted, this problem will worsen as the use of synthetic databases grows and departments and funding agencies increase their reliance on citation rates and associated metrics (e.g., *h*-index) in hiring, promotion, and funding decisions.

Literature cited

Abbott, A., Cyranoski, D., Jones, N., Maher, B., Schiermeier, Q. & Van Noorden, R. 2010. Do metrics matter? *Nature* 465: 860–862.

Finnegan, S., McClain, C.R., Kosnik, M.A. & Payne, J.L. 2011. Escargots through time: An energetic comparison of marine gastropod assemblages before and after the Mesozoic Marine Revolution. *Paleobiology* 37: 252–269.

Finnegan, S., Payne, J.L. & Wang, S.C. 2008. The Red Queen revisited: Reevaluating the age-selectivity of Phanerozoic marine genus extinctions. *Paleobiology* 34: 318–341.

Huntley, J.W. & Kowalewski, M. 2007. Strong coupling of predation intensity and diversity in the Phanerozoic fossil record. *Proc. Natl. Acad. Sci. U.S.A.* 104: 15006–15010.

Kowalewski, M., Dulai, A. & Fürsich, F.T. 1998. A fossil record full of holes: The Phanerozoic history of drilling predation. *Geology* 26: 1091–1094.

McClain, C.R. 2011. Op-Ed: The mass extinction of scientists who study species. *Wired*, http://www.wired.com/wiredscience/2011/01/extinction-of-taxonomists/.

Novack-Gottshall, P.M. 2007. Using a theoretical ecospace to quantify the ecological diversity of Paleozoic and modern marine biotas. *Paleobiology* 33: 274–295.

Novack-Gottshall, P.M. 2008. Ecosystem-wide body size trends in Cambrian-Devonian marine invertebrate lineages. *Paleobiology* 34: 210–228.

Novack-Gottshall, P.M. & Miller, A.I. 2003. Comparative geographic and environmental diversity dynamics of gastropods and bivalves during the Ordovician Radiation. *Paleobiology* 29: 576–604.

Novack-Gottshall, P.M. & Lanier, M.A. 2008. Scale-dependence of Cope's rule in body size evolution of Paleozoic brachiopods. *Proc. Natl. Acad. Sci. U.S.A.* 105: 5430–5434.

Payne, J.L. 2005. Evolutionary dynamics of gastropod size across the end-Permian extinction and through the Triassic recovery interval. *Paleobiology* 31: 269–290.

Payne, J.L. & Finnegan, S. 2007. The effect of geographic range on extinction risk during background and mass extinction. *Proc. Natl. Acad. Sci. U.S.A.* 104: 10506–10511.

Payne, J.L., Boyer, A.G., Brown, J.H., Finnegan, S., Kowalewski, M., Krause, R.A., Lyons, S.K., McClain, C.R., McShea, D.W., Novack-Gottshall, P.M., Smith, F.A., Stempien, J.A. & Wang, S.C. 2009. Two-phase increase in the maximum size of life on Earth over 3.5 billion years reflects biological innovation and environmental opportunity. *Proc. Natl. Acad. Sci. U.S.A.* 106: 24–27.

Sidlauskas, B., Ganapathy, G., Hzakani-Covo, E., Jenkins, K.P., Lapp, H., McCall, L.W., Price, S., Scherle, R., Spaeth, P.A. & Kidd, D.M. 2009. Linking big: The continuing promise of evolutionary synthesis. Evolution 64: 871–880.

Smith, F.A., Brown, J.H., Haskell, J.P., Alroy, J., Charnov, E.L., Dayan, T., Enquist, B.J., Ernest, S.K.M., Hadly, E.A., Jablonski, D., Jones, K.E., Kaufman, D.M., Lyons, S.K., Marquet, P., Maurer, B.A., Niklas, K., Porter, W., Roy, K., Tiffney B. & Willig, M.R. 2004. Similarity of mammalian body size across the taxonomic hierarchy and across space and time. *Amer. Naturalist* 163: 672–691.

Smith, F.A., Lyons, S.K., Ernest, S.K.M., Jones, K.E., Kaufman, D.M., Dayan, T., Marquet, P.A., Brown, J.H. & Haskell, J.P. 2003. Body mass of late Quaternary mammals. *Ecology* 84: 3402.

Villéger, S., Novack-Gottshall, P.M. & Mouillot, D. 2011. The multidimensionality of the niche reveals functional turnover in benthic marine biotas across geological time. *Ecol. Lett.* 14: 561–568.

Waegele, H., Klussmann-Kolb, A., Kuhlmann, M., Haszprunar, G., Lindberg, D., Koch, A. & Waegele, J.W. 2011. The taxonomist—an endangered race. A practical proposal for its survival. *Frontiers Zool.* 8: 25.

Jonathan L. Payne,[1] Felisa A. Smith,[2] Michal Kowalewski,[3] Richard A. Krause, Jr.,[4] Alison G. Boyer,[5] Craig R. McClain,[6] Seth Finnegan,[7] Philip M. Novack-Gottshall,[8] Laura Sheble[9]

1 *Department of Geological and Environmental Sciences, Stanford University, Stanford, California 94305, U.S.A.*
2 *Department of Biology, University of New Mexico, Albuquerque, New Mexico 87131, U.S.A.*
3 *Florida Museum of Natural History, University of Florida, Gainesville, Florida 32611, U.S.A.*
4 *Institute of Geosciences, Johannes Gutenberg University, 55128 Mainz, Germany*
5 *Department of Ecology and Evolutionary Biology, University of Tennessee, Knoxville, Tennessee 37996, U.S.A.*
6 *National Evolutionary Synthesis Center, Durham, North Carolina 27705, U.S.A.*
7 *Department of Integrative Biology, University of California, Berkeley, California 94720, U.S.A.*
8 *Department of Biological Sciences, Benedictine University, Lisle, Illinois 60532, U.S.A.*
9 *School of Information and Library Science, University of North Carolina, Chapel Hill, North Carolina 27599, U.S.A.*