**Citation and Attribution of Digital Products:   
Social and Technological Concerns**

Daniel S. Katz, dkatz@nsf.gov

The pursuit of science increasingly relies on activities that facilitate science but are not currently rewarded or recognized. Of particular concern are the sharing of data; development of common data resources, software and methodologies; and annotation of data and publications. This situation has been documented in a number of recent reports[[1]](#endnote-2) that focus on changing needs and mechanisms for attribution and citation of digital products, from the use of alternative metrics[[2]](#endnote-3) that track popularity, to work on data[[3]](#endnote-4). About half the articles in many recent issues of *Science* describe research that depended on software and a larger fraction analyze data. Indeed, the US National Science Foundation recently updated its guide to proposers to instruct them to provide a list of their “products” (objects that are “citable and accessible including but not limited to publications, data sets, software, patents, and copyrights”) rather than publications.[[4]](#endnote-5)

To promote such activities, we must develop mechanisms for assigning credit, facilitate the appropriate attribution of research outcomes, devise incentives for activities that facilitate research, and allocate funds to maximize return on investment. In this article, I introduce the idea of *transitive credit*, which addresses the issue of crediting indirect contributions, and discuss potential solutions to these other problems.

**History of Citation**

Throughout history, most formal citation has been for authentication and authority, rather than for the provision of credit and acknowledgment or attribution. Scientific citation in Western history appears by the late 1500s.[[5]](#endnote-6),[[6]](#endnote-7) In the early 1700s, citation also appears in the legal system as a method of understanding precedents.[[7]](#endnote-8) The idea of copyright as recognizing authors’ rights also arises at this time, from the Statute of Anne in 1710, perhaps due to a slow societal trend to recognize intellectual property, an idea that appears to have developed alongside the printing press.[[8]](#endnote-9)

Looking for the predecessors of an idea can be called “backward citing.” When multiple groups claim credit for the same advance, backward citing can ascertain how the larger scientific community assigns credit by looking at which groups are cited and how this changes over time.[[9]](#endnote-10)

“Forward citing” has also been used when one wants to understand how an idea has been used. This is often done through citation indices, the earliest examples of which are to portions of the Bible from the 1100s.[[10]](#endnote-11) The common use of citations indices in science is much more recent, as exemplified by Garfield’s work in the 1950s that led to the Science Citation Index.[[11]](#endnote-12)

New knowledge clearly builds on past knowledge. Traditionally, an author cites a previous paper by adding a reference to the author, title, place of publication, and so on. However, this concept doesn’t work well for digital products such as software, which are often dependent on libraries (assembled software packages), code fragments, and algorithms. For many of these, the identifier (a “name” that refers to a unique product) that should be cited is not clear. Additionally, if a cited library depends on another library, the contribution of this second library is not captured. Similarly, citation of a dataset should perhaps give credit to the people who gathered the data, as well as those who curated it, but the paper author may not know or be able to find these details.

**A Robust System of Citation**

For citation of digital products to be robust and at least semi-automated, we need to develop and build a set of tools and practices that (1) register digital products and those who should be credited for those products and (2) track usage of the products, and tie this usage to future products.

Let’s initially focus on the first requirement. Papers traditionally have been registered by commercial publishers, who generally use peer-reviewers to validate the quality of the work, but often charge readers for access to the papers. Alternatives also have appeared in recent years, such as PLOS, an open-access, peer-reviewed, non-commercial publisher, and ArXiv.org, an open-access, non-peer-reviewed repository. These systems also have the technical capability to register (and peer-review, if appropriate) software and data.

There are, of course, additional issues related to digital products, many of which are social, such as the potential volume of products being produced, and the number of versions of those products. If we develop a culture that expects these products to have value similar to that of papers, in which a group produces a small number of products each year, and these products embody significant progress beyond previous products, these issues can probably still be handled with today’s systems[[12]](#endnote-13). The question of credit for these products, however, will be as much an issue in the future as it is for papers today.[[13]](#endnote-14),[[14]](#endnote-15)

Many sets of standards for authorship exist, often distinct across disciplines, but it seems that in many fields, a substantial number of papers do not follow these rules, particularly with regards to granting honorary authorship[[15]](#endnote-16). Some journals have tried to solve this problem by requiring that the contribution of each author be defined, and other systems have also been proposed.[[16]](#endnote-17),[[17]](#endnote-18)

A technologically simple solution is to give fractional credit to all authors, which can also be done for software and data. Arguably, determining how to weight credit of the authors may be difficult, but it should be possible. And just as papers today are submitted by one person who is responsible for making sure all authors are listed and the paper is complete, etc., the submitter would also be responsible for registering this fractional credit.

We can also envision combining the idea of credit to contributors, as currently listed in authorship lists, and credit to others, as currently listed in acknowledgements, and credit to predecessors, as currently listed in citations, into one single credit map for each product. The reason to do so is to allow *transitive credit*, which is part of the second step in developing a robust system of citation.

**Tracking Product Usage**

The idea of transitive credit is as follows: The credit map for product A, which is used by product B, feeds into the credit map for product B. For example, product A is a software package equally written by two authors and its credit map is just 50% credit to each of them. Product B is a paper that depended on this package, and the authors assign 10% credit to the package. The transitive nature of this credit system means that the two authors of the software package now can each fairly claim 5% credit for the paper. If another paper is later written that extends the product B paper and gives 10% credit to that paper, the software package developers will also have 0.5% credit for the new paper.

The primary value of transitive credit is in measuring the indirect contributions to a product, which today are not quantitatively captured. Because they aren’t captured, they aren’t rewarded, and there is a disincentive to perform them. If they were captured, this disincentive would be replaced by an incentive, which for software and data means to publish and share them in a reusable form.

Tools to measure product usage are needed, such as provenance systems, some of which are being developed today. Some altmetrics work may also be used to help developers track their activities (paper and data views, software usage), so that they can select those that were related to the new products. This second step in developing a robust system of citation is important because as more digital products become available, it will become increasingly difficult for the person who registers a new product (whether paper, software, or data) to remember what previous products were used.

**Implementation**

In order for transitive credit to be measured, we first need unique identifiers for products, which can be done by the existing Handle System (http://handle.net), as extended by the existing digital object identifier (DOI, http://doi.org) system. Next, we need unique identifiers for authors, which is a problem that ORCID (http://orcid.org) is attempting to solve. Third, we need a way to register the unique mapping of credit for each product, which would require a new service to map a DOI to a weighted lists of additional ORCIDs or DOIs, which is no more technically challenging than the existing DOI system. Finally, in order track product usage, we need easy-to-use automated provenance systems.

**Social Motivation**

To promote the creation, maintenance, and use of digital products, we must measure these activities, and provide credit to those who perform them. The current lack of credit for performing these activities acts as a negative force that stops sharing of digital products, following Lewin’s principal of force field analysis[[18]](#endnote-19). Providing a credit mechanism would both remove the negative force and create a positive force, creating an incentive for sharing. This would impact recognition and status, hiring and promotion, and funding agency decisions.

These ideas have the potential to change the culture, because the act of measuring an item and publicizing that measure leads to a focus on improving that measure, thus improving the item. This focus on improving the measure can be intentional, as in the Check portion of the Deming Cycle, or unintentional, as happens when teachers teach students to answer specific questions rather than the material that the questions cover. Similarly, the h-index[[19]](#endnote-20) is now being used (and gamed) in many ways that Hirsch did not foresee and, Google’s PageRank algorithm has had a substantial impact on the Web. A metric ‘D’ providing credit to the developers of digital products would lead to people trying to increase their D-value by developing more such products. And if it was clear that they received credit from others who used their products, they would likewise make it clear that they had used products from others and give those others credit, since such credit is not a zero-sum game.

Issues of motivation are of particular concern today[[20]](#endnote-21) as science becomes more collaborative (aka team science), and as collaboration leads to more — and better — science.[[21]](#endnote-22) The average number of authors per paper is increasing, and collaborative projects are becoming common, which is part of the cause for the increasing number of paper authors.

**Conclusion**

Overall, the issues related to software and data citation can be solved with a mix of adapting current systems for tracking citations, developing a new system to register the unique mapping of credit for digital products that is similar to existing systems for tracking citations, and building new tools to help developers identify the existing digital products that they used. The result could be an acceptance of transitive credit and incentives for developing and sharing new digital products, supporting both forward and backward citing, and ultimately leading to better science and better understanding of science.

Additionally, while many incentives toward better citation practices may be social, funding agencies also have a role to play. Judging from the recent US discussion about data management plans and access policies for the outputs of publicly funded research, it’s clear that government agencies are willing to add requirements if they think it will benefit the country.

The goal of this article is to start a conversation on these issues, which can continue at events such as the Research Data Alliance (http://rd-alliance.org/) and the WSSSP workshop at SC13 workshop (http://wssspe.researchcomputing.org.uk/).

**Acknowledgments**

Some initial discussions about this issue took place in a 2010 Institute for Computing in Science (ICiS) workshop breakout session with Jacob Foster (U Chicago) and Robert Stevens (U Manchester). Conversations with David Proctor (NSF) and Ian Foster (U Chicago) have also shaped this article.

1. National Science Foundation, ACCI Task Force Reports on Campus Bridging; Data & Visualization; and Software for Science & Engineering (2011); http://www.nsf.gov/od/oci/taskforces [↑](#endnote-ref-2)
2. Priem, J., et al. *Altmetrics Manifesto*; http://altmetrics.org/manifesto [↑](#endnote-ref-3)
3. National Research Council. *For Attribution—Developing Data Attribution and Citation Practices and Standards* (National Academies Press, (2012). [↑](#endnote-ref-4)
4. National Science Foundation, *Grant and Proposal Guide*, NSF 13-1 (2013). [↑](#endnote-ref-5)
5. Lipsius, J. *De Militia Romana* (1595). [↑](#endnote-ref-6)
6. White, R. *Historiarum (Britanniæ) libri (1-11) ... cum notis antiquitatum Britannicarum* (1597-1607) [↑](#endnote-ref-7)
7. Raymond, Lord R. *Reports of Cases Argued and Adjudged in the Courts of King's Bench and Common Pleas, In the Reigns of The Late King William, Queen Anne, King George the First, and His Present Majesty* (1743). [↑](#endnote-ref-8)
8. Bettig, R. V. *Copyrighting Culture: The Political Economy of Intellectual Property. Westview Press*. pp. 9–30 (1996). [↑](#endnote-ref-9)
9. Lindahl B. I., Elzinga A, Welljams-Dorof A. Credit for discoveries: citation data as a basis for history of science analysis. *Theoretical Medicine and Bioethics 19(6)*, pp. 609-620 (1998). [↑](#endnote-ref-10)
10. Weinberg, B. H. Predecessor of scientific indexing structures in the domain of religion. *Second Conference on the History and Heritage of Scientific and Technical Information Systems*. pp. 126-134 (2004). [↑](#endnote-ref-11)
11. Thompson Reuters. *History of Citation Indexing*; http://thomsonreuters.com/products\_services/science/free/essays/history\_of\_citation\_indexing/ [↑](#endnote-ref-12)
12. Hafer, L., Kirkpatrick, A.E. Assessing open source software as a scholarly contribution. *Commun. ACM 52(12)*, pp. 126–129 (2009). [↑](#endnote-ref-13)
13. Schiermeier, Q. Europe's young researchers seek proper rewards. *Nature 397(6721)*, pp. 640-641 (1999). [↑](#endnote-ref-14)
14. Tarnow, E. The Authorship List in Science: Junior Physicist’s Perceptions of Who Appears and Why. *Science and Engineering Ethics 5(1)*, pp. 73-88 (1999). [↑](#endnote-ref-15)
15. Martinson, B. C., Anderson, M. S., De Vries, R. Scientists behaving badly. *Nature* *435(7043)*. pp. 737-738 (2005). [↑](#endnote-ref-16)
16. Verhagen, J. V., Wallace, K. J., Collins, S. C., Scott, T. R. QUAD system offers fair shares to all authors. *Nature 426(6967)*. p. 602 (2003). [↑](#endnote-ref-17)
17. Molla, M., Gardner, T. Roll Credits: Sometimes the Authorship Byline Isn’t Enough. PLOS Blog; http://www.plos.org/cms/node/285 (2007). [↑](#endnote-ref-18)
18. Lewin K. Defining the "Field at a Given Time." *Psychological Review 50.* pp. 292-310 (1943). [↑](#endnote-ref-19)
19. Hirsch, J. E. An index to quantify an individual’s scientific research output. *Proceedings of the National Academy of Sciences of the USA. 102(46)*. pp. 16569–16572 (2005). [↑](#endnote-ref-20)
20. Howison, J., Herbsleb, J. D. Incentives and integration in scientific software production. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pp. 459–470 (2013). [↑](#endnote-ref-21)
21. Wuchty, S., Jones, B.F., Uzzi, B. The Increasing Dominance of Teams in Production of Knowledge. *Science 316(5827)*. pp. 1036-1039 (2008). [↑](#endnote-ref-22)