Journal of
open research software

Katz, D S 2014 Transitive Credit as a Means to Address Social and Technological Concerns Stemming from Citation and Attribution of Digital Products. *Journal of Open Research Software,* 2(1): e20, pp. 1-4, DOI: http://dx.doi.org/10.5334/jors.be

ISSUES IN RESEARCH SOFTWARE

Transitive Credit as a Means to Address Social and Technological Concerns Stemming from Citation and Attribution of Digital Products

Daniel S. Katz*

The pursuit of science and engineering research increasingly relies on activities that facilitate research but are not currently rewarded or recognized, such as development of products and infrastructure. In research publications, citations are used to credit previous works. This paper suggests that a modified citation system that includes the technological idea of transient credit could be used to recognize the developers of products other than research publications and that if this were done in a systematic manner, it would lead to social and cultural changes that would provide incentives for the further development of such products, accelerating overall scientific and engineering advances.

Keywords: citation; credit; attribution; software

Introduction

The pursuit of science and engineering research increasingly relies on activities that facilitate research but are not currently rewarded or recognized. This includes 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] that focus on changing needs and mechanisms for attribution and citation of digital products, from the use of alternative metrics [2] that track reports of research impact apart from research publications, to work on data [3]. About half of 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].

To promote and advance pursuit of activities that facilitate research, 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 approaches to these other needs.

Note: this is an expanded version of a paper [5] that was part of the First Working Towards Sustainable Scientific Software: Practice and Experiences (WSSSPE1) workshop.

* National Science Foundation, Arlington, VA, USA
dkatz@nsf.gov

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 [6, 7]. In the early 1700s, citation also appears in the legal system as a method of understanding precedents [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 [9]. Note that paper authorship in science really is used to note both the actual authors of the paper as well as the contributors to the project [10].

Looking for the predecessors of an idea can be called "backward citing." In cases in which multiple groups claim credit for the same advance, backward citing may be used—by looking at which groups are cited and how this changes over time—to ascertain how the larger research community assigns credit [11].

The idea of "forward citing" has also been used in cases where 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 [12]. However, 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 [13].

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 that

should be cited—a "name" that refers to a unique product—is not clear. Additionally, if a cited library depends on another library, the contribution of this second library is not captured. 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.

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 [14]. 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 research 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 [15] 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, it's likely that 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.

In the commercial world, the idea of credit is often monetized, with software and data commercialized as products that must be purchased. This provides an alternative solution to recognizing the producers of such products, though it does not help in understanding their use in later scientific discoveries, i.e. forward citing.

Issues of motivation are of particular concern today [16] as science becomes more collaborative (team science), and this leads to more—and better—science [17]. The average number of authors per publication is growing, and collaborative projects are increasingly common, which is part of the cause for the growing number of publication authors.

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 first, register digital products and those who should be credited for those products, and second, 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 [18]. The question of credit for these products, however, will be as much an issue in the future as it is for publications today [19, 20].

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 [21]. 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 [22, 23]. (Note, a thorough survey of the different practices in various fields and by various publishers is needed.)

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. Recent work by Allen et al. [24] studied one taxonomy under which the role of each author of a paper was classified by the corresponding author. In general, the corresponding authors were satisfied with this and found it beneficial, though they also pointed out changes that could make this better. An additional step that would be needed for fractional credit to be used is to apply weights to the roles in the taxonomy.

Methods for doing this weighting, whether using a taxonomy or a more traditional list of authors, and analysis of these methods and their impact would likely happen if this overall idea moves forward. An example of work in this direction is the spliddit service [25], which applies Sperner's Lemma to the problem of how to fairly split [26] credit by choosing an author order that is envy-free.

Just as publications today are submitted by one person who is responsible for making sure all authors are listed (and perhaps assigned roles in a taxonomy) and the publication is complete, the submitter would also be responsible for registering this fractional credit, no matter how the values are determined.

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 that 50 percent of the credit for this should go the lead developer, 20 percent to the second developer, and 10 percent to the third developer. In addition, 5 percent should go to each of the four libraries that are needed to run the code. When this product is created and registered, this credit map is registered along with it. Product B is a paper that obtains new science results, and it depended on Product A. The person who registers the publication also registers its credit map, in this case 75 percent to her/himself, and 25 percent to the software code previous mentioned. Credit is now transitive, in that the lead software developer of the code can be given credit for 12.5 percent of the paper. If another paper is later written that extends the product B paper and gives 10% credit to that paper, the lead software package developer will also have 1.25% 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, due to the cost (in time or something else). If they were captured, this disincentive would be replaced by an incentive, which for software and data would mean to publish and share them in a reusable form.

Tools to measure product usage are needed, some of which are being developed or used today, such as provenance systems, the Thompson Reuters' Data Citation Index, article level metrics, especially when used with software and data papers, and many altmetrics. Provenance systems in particular may be used to help developers track their activities (such as publication and data views, or software usage), so that they can select those that were related to 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 publication, 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 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.

Outcomes

With such a universal transitive credit system, we could quantify the contribution of the code developer to research by summing over all the products where the code is used. This information could be used in multiple ways, for example, by employers in making decisions about hiring, promotions, or raises. It could also be used by funding agencies to help decide what products to support.

An additional benefit of such a system is its application to provenance. If a failure in a product, such as a bug in a code, is discovered, this system would easily allow discovery of later products (including publications) that used the faulty product, which the failure may have invalidated.

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 research and a better understanding of research. Additionally, while many incentives toward better citation practices may be social, funding agencies also have a role to play. Judging from the recent discussion in the US 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 and the research enterprise.

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 WSSSPE series (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.

References

1. **National Science Foundation** 2011 ACCI Task Force Reports on Campus Bridging; Data & Visualization; and Software for Science & Engineering. Available at: http://www.nsf.gov/od/oci/taskforces [Last accessed 02 April 2014].
2. **Priem, J, Taraborelli, D, Groth, P** and **Neylon, C** Altmetrics Manifesto. Available at: http://altmetrics.org/manifesto [Last accessed 02 April 2014].

3. **National Research Council** 2012 For Attribution—Developing Data Attribution and Citation Practices and Standards. National Academies Press.
4. **National Science Foundation** 2013 Grant and Proposal Guide. NSF 13–1.
5. **Katz, D S** 2013 Citation and Attribution of Digital Products: Social and Technological Concerns. *figshare*, 791606. DOI: http://doi.acm.org/10.6084/m9.figshare.791606
6. **Lipsius, J** 1595 De Militia Romana.
7. **White, R** 1597–1607 Historiarum (Britanniæ) libri (1–11) ... cum notis antiquitatum Britannicarum.
8. **Raymond, Lord R** 1743 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.
9. **Bettig, R V** 1996 Copyrighting Culture: The Political Economy of Intellectual Property. Westview Press, pp. 9–30.
10. **Teixeira da Silva, J A** 2011 The ethics of collaborative authorship. *EMBO Reports*, 12(9): 889–893. DOI: http://dx.doi.org/10.1038/embor.2011.161
11. **Lindahl, B I, Elzinga, A** and **Welljams-Dorof, A** 1998 Credit for discoveries: citation data as a basis for history of science analysis. *Theoretical Medicine and Bioethics*, 19(6): 609–620. DOI: http://dx.doi.org/10.1023/A:1009944903620
12. **Weinberg, B H** 2004 Predecessor of scientific indexing structures in the domain of religion. In: Second Conference on the History and Heritage of Scientific and Technical Information Systems, pp. 126–134.
13. **Thompson Reuters** History of Citation Indexing. Available at: http://thomsonreuters.com/products_services/science/free/essays/history_of_citation_indexing/ [Last accessed 02 April 2014].
14. **Lewin, K** 1943 Defining the "Field at a Given Time". *Psychological Review*, 50: 292–310. DOI: http://dx.doi.org/10.1037/h0062738
15. **Hirsch, J E** 2005 An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the USA*, 102(46): 16569–16572. DOI: http://dx.doi.org/10.1073/pnas.0507655102
16. **Howison, J** and **Herbsleb, J D** 2013 Incentives and integration in scientific software production. In: Proceedings of the ACM Conference on Computer Supported Cooperative Work. San Antonio, TX, pp. 459–470.
17. **Wuchty, S, Jones, B F** and **Uzzi, B** 2008 The Increasing Dominance of Teams in Production of Knowledge. *Science*, 316(5827): 1036–1039. DOI: http://dx.doi.org/10.1126/science.1136099
18. **Hafer, L** and **Kirkpatrick, A E** 2009 Assessing open source software as a scholarly contribution. *Communications of the ACM*, 52(12): 126–129. DOI: http://dx.doi.org/10.1145/1610252.1610285
19. **Schiermeier, Q** 1999 Europe's young researchers seek proper rewards. *Nature*, 397(6721): 640–641. DOI: http://dx.doi.org/10.1038/17660
20. **Tarnow, E** 1999 The Authorship List in Science: Junior Physicist's Perceptions of Who Appears and Why. *Science and Engineering Ethics*, 5(1): 73–88. DOI: http://dx.doi.org/10.1007/s11948-999-0061-2
21. **Martinson, B C, Anderson, M S** and **De Vries, R** 2005 Scientists behaving badly. *Nature*, 435(7043): 737–738. DOI: http://dx.doi.org/10.1038/435737a
22. **Verhagen, J V, Wallace, K J, Collins, S C** and **Scott, T R** 2003 QUAD system offers fair shares to all authors. *Nature,* 426(6967): 602. DOI: http://dx.doi.org/10.1038/426602a
23. **Molla, M** and **Gardner, T** 2007 Roll Credits: Sometimes the Authorship Byline Isn't Enough. PLOS Blog. Available at: http://www.plos.org/cms/node/285 [Last accessed 02 April 2014].
24. **Allen, L, Scott, J, Brandt, A, Hlava, M** and **Altman, M** 2014 Publishing: Credit where credit is due. *Nature*, 508: 312–313. DOI: http://dx.doi.org/10.1038/508312a
25. spliddit Available at: http://www.spliddit.org [Last accessed 02 May 2014].
26. **Su, F E** 1999 Rental harmony: Sperner's lemma in fair division. *The American Mathematical Monthly*, 106(10): 930–942. DOI: http://dx.doi.org/10.2307/2589747

How to cite this article: Katz, D S 2014 Transitive Credit as a Means to Address Social and Technological Concerns Stemming from Citation and Attribution of Digital Products. *Journal of Open Research Software,* 2(1): e20, pp. 1-4, DOI: http://dx.doi.org/10.5334/jors.be

Published: 9 July 2014

]u[*Journal of Open Research Software* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS