Data and Software Should Be First Class Contributions

Being on a new tenure track has led me to think a lot lately about "what counts" in an academic career. I won't bother with the specifics of the rules governing my current position, but as almost all political scientists know there are very few things that count for much: articles in a specific set of general field journals, articles in an even narrower tier of subfield journals, mythical (non-fraudulent) "unicorn" publications in general academic outlets like *Science* or *Nature*, and academic books from a couple of university presses. The breadth of each of those lists, of course, depends on where you are and who has control over your academic future.

In this post I want to make a very brief argument about two types of contributions that are clearly outside the scope of that "what counts" list: namely, data and software. I make a lot of software. My discussion of why I think it should count for something is almost entirely selfish. I wish the time I spent on software counted, but I know that for the most part it doesn't. But, I can also offer what I think is a reasonable defense of putting more weight on software (and data) and that's what you're about to read.

An article is a one-off endeavor. When the text comes out in print and you finish archiving your reproduction files at Dataverse (or wherever), that publication is essentially finished. You can joyfully sit around waiting to accrue citations on Google Scholar, perhaps write a spin-off paper using new data or expand it into a book. But at its core, the article is done. You likely won't run new analyses on the data, few if any people will ever write to ask you questions about it (this being likely inversely proportional to degree of transparency in your work), and the only future interaction with the document will be citations. Software and data are different.

Developing software is rarely a one-off project. I didn't realize this when I started developing R packages, but it's something that has become incredibly clear in the three years since my first R package (MTurkR) was published on CRAN. MTurkR started out narrowly as a way to send

follow-up emails to MTurk workers. That was it. Now, nine published versions later, MTurkR does much, much more. I've given talks about it (at useR!2015), I've written newsletter articles about it, there's a pretty comprehensive wiki showing all kinds of wisdom I've learned about it and MTurk generally. I have 200+ email conversations in my gmail account related to MTurkR. I probably answer 1-2 questions a week about it via email. I've answered 33 questions about it on StackOverflow. I've closed 95 bug reports, suggestions, and improvements on GitHub. And I've posted 455 times on the MTurk developer forum, helping others use the platform, reporting problems to the AWS staff, and so forth. Of all of the things I've done in academia, this is by far the thing that I've spent the most time on; and I keep spending time on it. I suspect I will spend time on it long into the future. (All of this says nothing about the other, probably less useful software projects I've worked on. Rmonkey is starting to become a major time commitment, as well, but we'll see how it progresses.) And yet, despite all of this time, none of this - not even MTurkR itself - "counts".

No pity needed here. It was my decision to work on MTurkR (and all other projects).

But, hopefully this discussion highlights the significantly different nature of software (and large-scale data production) versus article publishing. When you create software or novel data, you don't just make a one-time contribution. Instead, you commit yourself to *maintaining* that contribution, cultivating it, nurturing its user base, and constantly improving it. For that reason, software and data can appear to be relatively modest contributions but they can - particularly when they are useful - become massive efforts with relatively substantial impacts on other researchers' work. When we don't "count" software and data, we miss the implicit commitment involved in such contributions that often far exceeds the commitment to a given article, chapter, or even a book.

When a researcher produces, publishes, and maintains useful software or useful data, I think we need to treat those as "first class" contributions on par with articles, chapters, and books. In rare cases (say, SPSS or ggplot2), I would also be willing to argue that those contributions far, far exceed the value of any traditional academic publication. Writing useful software or creating a novel dataset isn't just a publication, it's an indefinite commitment to helping other researchers. That should be worth something.

Published: 2015-12-09

[Feed]

← Older post

Newer post →
