Incentives not to Share

Josh Rolnick

This past year, for the first time, I have had to construct my own datasets for research projects. My two projects have quite distinct data needs. The first project uses aggregate data on post-abortion complication rates in different states across time. Working with another student, I have spent the past eight months piecing together data from different public sources. It has been time-consuming and we have had to proceed carefully in selecting and combining sources. Data collection for the second project, however, is even more time-consuming. For this project, I am administering a survey to inmates in more than 30 state prisons. Data efforts began more than two years ago. I have spent countless hours this fall finishing the effort—securing state approval, coordinating with prison staff, tinkering with the survey.

I plan to make my data and code publicly available in a manner that will permit others to reproduce and extend my work. I will do this because I believe strongly in the values of data sharing and reproducibility. From a purely self-interested perspective, however, I am not sure that sharing is to my benefit. Here I think it is helpful to distinguish between the values of reproducibility and extensibility, both quite important. From self-interest, reproducibility is unlikely to be harmful—provided the results are, indeed, reproducible, of course. Allowing others to reproduce results may endow them with more legitimacy. It may encourage citations to my work. Extensibility, the ability to use the data for further work without legal or technological barriers, seems less clear. On the one hand, it might also increase citations. On the other hand, it allows other researchers with larger staffs and more resources to use the data. If I kept the data to myself, I would have the exclusive ability to produce additional papers. And—regrettably—it is still the case that hoarding the data would be unlikely to prevent me from finding a good place to publish results.

Perhaps I am wrong. Authors do, after all, generally receive something of value for providing data. Presumably, work based on my data might cite a paper of mine. In some set of circumstances, I might be a co-author. There are various forms of credit that a researcher might receive for supplying useful data—source attribution in a footnote, co-authorship, publication-style credit toward career advancement. I do not feel that these norms are entirely clear or consistent, as far as I can tell, in many fields. Right now, facing uncertainties about benefits and worries about competition, a self-interested junior (or senior) researcher may still decide not to share, at least in some circumstances.

The problem of self-interested hoarding may solve itself. If—as I hope—strictly enforced rules about supplying reproducible, extensible data do take hold, then researchers will need to share, whether they want to or not. And then it is likely that norms of credit will become more clear and consistent on their own. But as a research community, we should make sure that such rules do emerge, rules that properly reward researchers for the valuable contribution of gathering and organizing data. For rigorous data collection is itself an important part of scholarship.