Internalizing Externalities: Designing Effective Data Policies

By RYAN HILL, CAROLYN STEIN, AND HEIDI WILLIAMS*

* Hill: Department of Economics, MIT; 77 Massachusetts Avenue, E52-301, Cambridge, MA 02139 (ryanhill@mit.edu). Stein: Department of Economics, MIT; 77 Massachusetts Avenue, E52-301, Cambridge, MA 02139 (cstein@mit.edu). Williams: Department of Economics, Stanford University; 579 Jane Stanford Way, Office 323, Stanford, CA 94305 (hlwill@stanford.edu).

In tandem with a broader trend across disciplines, many economics journals have recently adopted policies requiring authors to archive and curate research data, in an effort to promote reproducible research. For example, the American Economic Association (AEA)'s data and code posting policy is intended to create a minimal framework from which to replicate empirical findings by requiring the data and code to be available to others.[1] The importance of such efforts should not be understated. However, we are concerned that the economics profession has – relative to many other scientific disciplines – focused too little attention on the related question of what types of institutions and incentives might encourage and subsidize the creation and sharing of datasets that are likely to facilitate novel follow-on research of high social value. We argue in this paper that this question deserves more attention.

As a preliminary step in that direction, we briefly describe some examples from other scientific fields of institutions and incentives designed to promote subsequent research, and we speculate on some potential reforms that could be undertaken within the field of economics to encourage the type of data collection, curation, and sharing that is most likely to encourage socially valuable follow-on research.

I. What is the problem?

The current status quo in economics can be characterized in simple terms as follows: new data is collected or constructed when it is in the interests of a particular researcher to do so, and such researchers – naturally, given the incentives they face – focus attention on collecting whatever data they need for their particular research project.

As an example, consider a recent research project one of us (Williams) published with Bhaven Sampat (Sampat and Williams 2019).

1 See also Vilhuber (2019) on the recently created AEA Data Editor position.

We spent nearly ten years working on this paper, and much of that time was focused on wrangling the publicly available data from the US Patent and Trademark Office (USPTO) into a format that would let us analyze around 1,500 patent applications claiming human genes. One of the empirical approaches in that paper leveraged the fact that patent applications are quasi-randomly assigned to patent examiners at the USPTO. We developed an instrumental variables approach to investigate how (quasi-randomly assigned) patents on human genes affected subsequent medical innovation. That empirical approach is potentially applicable across many contexts, and we frequently receive requests for our "examiner fixed effect" estimates from other researchers who would like to re-use these estimates in other contexts.

For the purposes of our research project, we only needed to build a dataset of the 1,500 patent applications claiming human genes, and estimate variation in examiner grant propensities for the relatively small sample of patent examiners who reviewed those applications. But from a social perspective, a relatively small increment of additional work during the course of our project would have created a dataset covering all patent applications (of which human gene applications are of course a subset) that would enable other researchers to easily re-use and apply our empirical approach in other contexts.

What is the problem? Many datasets are expensive to curate but cheap to share, and – given the current set of incentives facing economists – most researchers optimize their data curation to maximize individual utility rather than social benefit. A passive version of this distortion is that with only slightly more work, many researchers could produce a more general version of their data that would be much more useful to a much broader set of follow-on researchers. More active (and self-serving) versions of this distortion sometimes also arise: some scientific researchers, seeking to maintain a competitive advantage over their peers, may strategically choose not to disclose their data in the form that would be most useful for follow-on research.

This problem is in many ways a classic public good provision problem. But given that many public good problems exist, why should we focus attention on trying to fix this one? We feel frustrated by and concerned about the potential inefficiencies this problem generates. Researchers in economics are constantly forced to re-do work that has been done by other research teams in order to build on those discoveries. While an optimally designed system could purposefully choose to allocate researchers with a proprietary window of

access to data they invest in curating, it seems clear to us that the status quo in economics was not purposefully designed to optimize over the relevant trade-offs. From a practical perspective, there is reason to think these policy choices really matter, as there is a growing body of evidence suggesting that less science happens – and fewer products are commercialized – when access to data and basic research materials are limited (Furman and Stern 2011; Williams 2013).

If you are willing to grant that this is a problem worth trying to solve, what should we do about it? Relying on goodwill seems likely to be insufficient: researchers have scarce time and resources, and expecting public goods to be provided in the absence of providing incentives or institutions to support their creation seems unlikely to be the best path forward.

II. Have other fields "solved" this problem?

Drawing on our past work analyzing the fields of observational astronomy (Hill 2019), structural biology (Hill and Stein 2019), and genetics (Williams 2013; Sampat and Williams 2019), here are three examples from other scientific fields of institutions and incentives which address, in various ways, this problem.

A. Observational Astronomy

The field of observational astronomy records and analyzes data about the observable universe. The traditional research model in observational astronomy is that a central organization supplies a large upfront capital investment to build a telescope, and then allocates short periods of observing time to individual astronomers through some type of peer review process. For private organizations (such as the Keck Institute for Space Studies owned by CalTech), they can restrict access to appropriate as much value out of their investment as possible for their own scientists. For public organizations (such as the Hubble Telescope), conditional on passing a peer-review bar any researcher can collect observations. At all public and most private observatories, all data is publicly shared after a 6-18 month proprietary window; this is intended to give a priority advantage to the researcher who generated the data, but also foster follow-on research through public release.

A new research model was pioneered by the Sloan Digital Sky Survey (SDSS) in the year 2000. The SDSS was motivated by the idea that the traditional model created a patchwork of short and disparate observations which was likely not the most efficient way of creating publicly useful data: because each astronomer

has her own agenda, a year's worth of observations on the Hubble Telescope might create data for hundreds of papers on hundreds of different topics, yet the resulting archive may be too disconnected to enable meaningful follow-on research. In response to this perceived problem, SDSS embarked on a very broad mission, spending multiple years with a dedicated telescope mapping the sky while focusing on a few specific but broadly useful scientific goals. The Sloan Foundation staked the first 30% of funding, and then solicited proposals from universities around the world to invest. Each university that paid the entrance cost was able to help design the mission, participate in the implementation, and have first access to the data to write the papers they had proposed. Each wave of data was publicly released, and each wave of data had a companion working paper (co-authored by participants) that should be cited each time the public data is used. Over the past two decades, the SDSS has become the most cited data source in astronomy.

B. Structural Biology

The field of structural biology analyzes the three-dimensional structure of biological macromolecules such as proteins. Understanding the structure of proteins is essential for understanding their function and role in health and disease. Protein folding and macromolecule structure has had a series of important applications in medicine, and around fifteen Nobel Prizes have been awarded for advances in structural biology.

As of the early 1970s, very few protein structures had been solved, and the community of researchers using x-ray crystallography – the technique most frequently used to solve protein structures – was small. However, scientists understood the importance of providing unified access to these growing data for follow-on basic research and applied use. The Protein Data Bank (PDB) was established in 1971 as a uniform repository for data on protein structures. At the time it was founded, the PDB contained only seven structures, but the PDB is now one of the most widely used data resources in biology, containing over 150,000 entries. Depositors today submit data to the PDB through a deposition portal, and each entry is checked, validated, and annotated by PDB staff, with no charge to depositors. At the time of deposit, 4-character PDB IDs are assigned to each structure, which serve as unique, immutable identifiers of each entry in the PDB. The resulting database is organized, searchable, and easily accessible to the broad scientific community.

Observers of this field have argued that deposits to the PDB accelerated after an ad hoc

committee of scientists published a formal set of guidelines for data deposition in 1989. The original guidelines stated that structures should be deposited at the time of publication, and should be publicly released within a year. Most major journals subsequently adopted these guidelines, and the National Institute of General Medical Sciences (NIGMS) at the US National Institutes of Health (NIH) made federal funding contingent on depositing to the PDB. In the late 1990s, *Nature*, *Science*, and other leading journals revised their policies to require verification that the structure had been deposited in the PDB prior to paper acceptance, and that the data be released prior to or at the time of publication. Virtually all journals in the field now follow some version of this policy.

C. Genetics

The field of genetics analyzes genes, genetic variation, and heredity. Traditionally, individual teams of genetics researchers pursued a "targeted" effort to map and sequence genes based on research that systematically traced the roots of specific diseases thought to have a genetic basis, such as Huntington's disease. That traditional model was upended in 1990, when a large publicly-funded effort called the Human Genome Project was launched with the goal of sequencing the entire human genome by 2005.

The Human Genome Project leveraged its position to enforce a regime of open science (in part motivated by a desire to limit researchers' ability to patent genes sequenced under the project). In 1996, the heads of the largest labs involved in the Human Genome Project agreed at a Bermuda-based meeting to the so-called "Bermuda rules," which required that data sequenced under the Human Genome Project be posted on an open-access website within 24 hours of sequencing.[2] The stated goal of the Bermuda rules was "to encourage research and development and to maximize [the data's] benefit to society."

Under the Bermuda rules, genes sequenced by the Human Genome Project were required to be deposited in GenBank – an open-access archival sequence database. By nature of being archival, GenBank entries can be very redundant and are not catalogued in any systematic way (GenBank records are "owned" by the original submitter and cannot be altered by third parties). The National Center for Biotechnology Information (NCBI) at the US National Library of Medicine uses data derived

2 These rules replaced a previous US policy which required that data be made available within 6 months. The largest labs involved in the Human Genome Project were the US DOE Joint Genome Institute (Walnut Creek, CA), Baylor College of Medicine Human Genome Sequencing Center (Houston, TX), the Wellcome Trust Sanger Institute (UK), Washington University School of Medicine Genome Sequencing Center (St. Louis, MO), and the Whitehead Institute/MIT Center for Genome Research (Cambridge, MA).

from GenBank and similar data sources to create a separate open-access sequence database called RefSeq. In contrast with GenBank, RefSeq aims to provide a validated, non-redundant, well-annotated set of sequences that can provide a stable reference for gene identification and characterization (including a unique ID number), and to summarize the current state of scientific knowledge of known genes. For example, the only known mRNA for the RAX2 gene has RefSeq ID number NM_032753, and that RefSeq entry is curated with information about four scientific publications studying that gene.[3] RefSeq records are owned by NCBI and can be updated as needed to remain current.

III. Lessons for the field of economics

Against the background of these three examples, let us speculate on some potential lessons for the field of economics. We focus our discussion here on three broad classes of policies: (1) changing the reporting requirements of institutions such as journals or funding agencies; (2) changing the incentives facing researchers; and (3) having public, non-profit, or philanthropically-oriented for-profit firms provide public goods.

In many ways, changing reporting requirements is the "easiest" type of reform. For example, rather than requiring economists to post the code and data required to replicate just the tables and figures of their paper, one could instead require researchers to disclose all components of the data and code used in their analysis – including the full raw data files, access to which would presumably go some distance towards encouraging follow-on research. However, this path is often quite costly (e.g. to host very large raw data files), and by nature of not directly addressing the underlying incentive problem would require monitoring and enforcement. For example, in the case of structural biology, scientists are supposed to report experimental details (exact methods, environmental conditions) to the PDB to allow other scientists to replicate their experiments. However, anecdotal evidence suggests that some researchers provide minimal (or even insufficient) details in an effort to maintain trade secrecy.

Changing the incentives facing researchers – by, for example, granting researchers more credit or recognition for creating public goods – seems, a priori, to be a more promising way to align private incentives with social contributions. For example, a very well-cited

3 See https://www.ncbi.nlm.nih.gov/nuccore/NM_032753.4.

"paper" in economics is Bronwyn Hall, Adam Jaffe, and Manuel Trajtenberg's (unpublished) documentation file describing their linkage of the Compustat data with the US patent grants data. However, the field of economics currently has very poor norms around citing datasets – even though this Hall-Jaffe-Trajtenberg Compustat-patent linkage documentation has nearly 4,000 citations (as of November 2019), we would conjecture that the underlying data has in fact been used much more broadly than those citations would suggest. The example of the Sloan Digital Sky Survey, which formally instituted norms that researchers using the Sloan Digital Sky Survey data cite the companion working paper (co-authored by participants), is one example of how data citations could be encouraged. Journals could also require that as part of the publication process, datasets used in the analysis are cited appropriately.

The final class of policies we consider – having public, non-profit, or philanthropically-oriented for-profit firms provide public goods – is more expensive, but also provides much more scope to address the underlying incentive problems. While fixed costs are frequently lamented as a barrier to data creation, in several ways the existence of fixed costs can be leveraged in a way that enforces open science and maximizes the potential for follow-on research. In the case of observational astronomy, one benefit of expensive telescopes is that only the government or large collaboratives can afford to purchase them, and once they are built the founding organizations control access to the data. If these organizations have a pro-social objective function – as did the Sloan Digital Sky Survey, the Protein Data Bank, and the Human Genome Project – they can set policies in a way that meaningfully encourages follow-on research.

First, centralized institutions can set policies for priority protection. For public and most private observatories for observational astronomy, data is publicly shared after a 6-18 month proprietary window. In the case of the Protein Data Bank, data is kept secret until the paired paper is published. In general, we would expect longer priority windows to encourage data creation, and shorter proprietary windows to hasten follow-on innovation by outside researchers. A centralized collaborative can weigh these trade-offs in a way that seems appropriate given the specifics of the context at hand. A natural example for the field of economics would be an effort to systematically link and provide streamlined access to public administrative datasets such as those available in Scandinavian countries which link information on births, deaths, demographics,

education, employment, social benefits, and other records.

Second, centralized institutions provide a natural mechanism through which scientific goals can be coordinated. In observational astronomy, individual telescope observers typically try to maximize their limited observing time for the improvement of the one or handful of papers they plan to write. In medical genetics, researchers typically focus on understanding their hereditary disease of interest largely in isolation. As with the Sampat and Williams example above, economists similarly typically focus their attention on curating the data they need for their particular papers. In each case, individual goals are set based on private costs and benefits without much, if any, effort to account for what additional social benefits could be realized for a given change in private costs. In contrast, an effort like the Sloan Digital Sky Survey aims to maximize the social value of data by soliciting input from a broad set of researchers on what data should be systematically collected and curated. The required monetary stake produces revealed preferences for worthwhile scientific goals, but these goals have to be coordinated with other stake-holders, thus maximizing the social value of the data. In the field of economics, one example of where such coordination could be valuable is in the field of health economics. Teams of researchers often access administrative records on Medicaid beneficiaries for individual states, because Medicaid is a state-run program and the structure of the Medicaid claims data differs across states in a way that is difficult to harmonize. However, precisely *because* Medicaid is a state-run program, it seems ideal to combine Medicaid data from *all* states, so that state-by-year variation in Medicaid policies, and linked data allowing researchers to follow Medicaid beneficiaries as they move across states, could be used to assess Medicaid policies. Having a centralized institution undertake the – admittedly heroic – task of harmonizing the Medicaid claims data from all states in a way that linked individuals and created a centralized data resource which could be reused by others seems likely to be an incredibly valuable endeavor.

Third, centralized institutions provide a natural mechanism through which to standardize the format in which data is collected and curated. In the field of structural biology, the Protein Data Bank (PDB) invests time and resources into standardizing new protein deposits, and ensuring that the data remain consistent over time. In the field of medical genetics, RefSeq – which verifies, annotates, and curates genetic data in a standardized format with unique records for

each gene – has likely been immensely more valuable than GenBank, which archives sequenced genes in an open-access database riddled with redundancies. In contrast with RefSeq, individual human genomes are typically sequenced with little attention to standardization, and researchers frequently need to essentially start from scratch when creating a useable dataset. The current norms in economics are similar to those for individual genomes and for GenBank – researchers do sometimes choose to archive their data in repositories such as ICPSR or Harvard's Dataverse, but there are few centralized efforts such as PDB or RefSeq which aim to curate such data in a way that enhances opportunities for re-use. Going forward, one potential opportunity in the field of economics would be to advocate for the inclusion of unique identifiers for individuals and institutions to be added to administrative datasets. For example, PubMed – which catalogues biomedical research – began recording ORCID IDs for researchers in 2015, which are meant to provide stable unique identifiers of researchers. One can imagine encouraging other institutions like the US Patent and Trademark Office (USPTO) to adopt ORCID IDs as well.

Of course, real barriers would arise in pursuing any of these types of efforts. As one example, astronomers often express concern that long author lists make it hard for publications to signal potential in scientific careers. More unique to the social sciences are thorny barriers related to data security and privacy. But it seems well worth the effort for an organization like the AEA to invest in assessing potential investments in this space.

REFERENCES

Furman, Jeffrey and Scott Stern (2011) "Climbing Atop the Shoulders of Giants," *American Economic Review* 101(5): 1933-63.

Hill, Ryan (2019) "Searching for Superstars: Research Risk and Talent Discovery in Astronomy," mimeo.

Hill, Ryan and Carolyn Stein (2019) "Scooped! Estimating Rewards for Priority in Science," mimeo.

Sampat, Bhaven and Heidi Williams (2019) "How Do Patents Affect Follow-On Innovation? Evidence from the Human Genome," *American Economic Review* 109(1): 203-236.

Vilhuber, Lars (2019) "Report by the AEA Data Editor," *AEA Papers and Proceedings* 109: 718-729.

Williams, Heidi (2013) "Intellectual Property Rights and Innovation: Evidence from the Human Genome," *Journal of Political Economy* 121(1): 1-27.