NISS

Why Data Availability is Such a Hard Problem

Alan F. Karr

Technical Report 186
February 2014

National Institute of Statistical Sciences
19 T.W. Alexander Drive
PO Box 14006
Research Triangle Park, NC 27709
www.niss.org

Why Data Availability is Such a Hard Problem

Alan F. Karr
National Institute of Statistical Sciences
Research Triangle Park, NC 27709, USA
karr@niss.org

Abstract

If data availability were a simple problem, it would already have been resolved. In this paper, I argue that by viewing data availability as a public good, it is possible to both understand the complexities with which it is fraught and identify a path to a solution.

1 Data Availability as a Public Good

Those who view data availability as a black-and-white issue—the *purist* view, as in the left-hand panel in Figure 1, are ignoring or attenuating not only reality, but also fundamental principles of economics and human behavior. Instead, data availability is composed of infinitely many shades of gray, as in the right-hand panel in Figure 1—the *realist* view.

My fundamental point is that *data availability is a public good* (Varian, 1992). As are other public goods, it is extremely complex. Strikingly, however, much of the current conversation about data availability ignores, in many cases willfully, this complexity. To purists who disagree, I submit that the empirical evidence is overwhelming. If data availability were a simple problem, it would have been resolved long ago. Solutions imposed by fiat are inefficient at best, and generally ineffective. Many proposals overlook the multiplicity of stakeholders (§3), as well as the complex, competing incentives to which they are subject.

Without going into economic depth, the classic example of a public good is national defense. It is there for everyone because no one can sensibly, let alone efficiently, defend only himself or herself. There can be only one military in a country, and it results from a collective societal decision. Individual sacrifices are necessary, in the form of taxes to fund the military, and (in many countries) compulsory service in it. For some people, tragically, the sacrifice is extreme. Opting out is not possible: even symbolic acts such as not paying taxes on the basis that some of them support the military does not deprive anyone of protection. Other relevant economic terms are *non-excludable*—individuals cannot be excluded from use—and *non-rivalrous*: use by one individual does not affect availability to others.

A multitude of choices is necessary: How large? How expensive? Who shall serve? What actions by the military permissibly serve the national interest? In the U.S. (the only country that I know well), the complexity of these decisions is enormous, and the mechanisms to cope with the complexity—in particular, the President's being the Commander-in-Chief[1]—can be traced to the Constitution.

Closer to my subject, Stiglitz (1999) portrays knowledge as a public good. Publicly available knowledge is clearly non-excludable, and all knowledge is non-rivalrous. Equally close, Abowd (2013) analyzes population statistics and privacy as public goods, using economic arguments to show that under some scenarios, privacy is over-supplied and information is under-supplied. See also §3.3.

Data availability is not the military, but like other public goods, it cannot be achieved without sacrifice, some of which must be compensated. To continue the analogy, having a military is deemed a benefit to soci-

[1] Article II, Section 2: "The President shall be commander in chief of the Army and Navy of the United States, . . ."

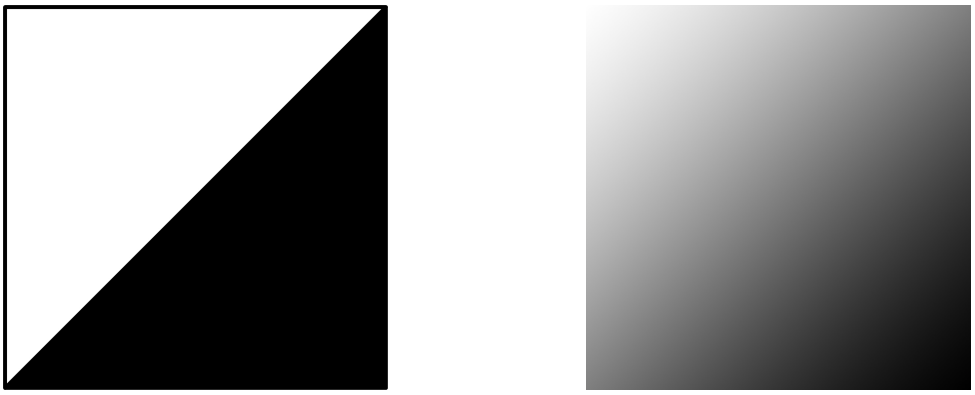

Figure 1: Left: The purist view of data availability. Right: The realist view of data availability.

ety. Therefore, those who serve in it, which is at least in part not in their self-interest, receive compensation, in the form of pay and various benefits, both during and after service.[2]

By viewing data availability as a public good, a central tenet becomes clear: compensation is both appropriate and necessary. Data availability, while patently in the social interest, is not always in the self-interest of the individuals and institutions being asked or required to implement it. We do not ask members of the military to serve without compensation, and it is equally unfair—not to mention unrealistic—to view data availability differently. So long as this perspective is not part of the conversation, while some positive steps can take place, creating a prevailing climate of data availability will remain intractable.

2 Two Contextual Complications

In the presentation on which this paper is based (Karr, 2013), I suggested, only half jokingly, that other than being unable to define either "data" or "availability," we are in fine shape. In fact, inability to define the fundamental terms is a major impediment to progress. In this section, I discuss them briefly, in the opposite order.

Lurking in the background is the delicate question of responsibility. In an attempt to keep this paper concrete rather than philosophical, I have downplayed responsibility, possibly non-constructively. When I make data publicly available, what responsibility do I assume? To whom? Without pursuing them in depth, the parallels to software are insightful. For open source software, availability is total, cost is zero,[3] and responsibility is nil: let the user beware. For commercial software, by comparison, availability is at a cost that at least conceptually, reflects the resources used to create the software. Even notwithstanding the disclaimers in the Terms and Conditions, there is some level of responsibility.

I have read only a few, and heard fewer, discussions of responsibility and liability in the context of data availability. What if the data are not exactly what they are purported to be? What if Professor X publishes an analysis of flawed data from Professor A? What assurances are necessary, and which are desirable? I

[2]The benefits can be very costly. In the U.S., the costs of the military education, health and retirement systems far exceed the salaries paid to active members of the military.

[3]At least monetary cost; time costs may be substantial, even disproportionately so.

have no answers to these questions, and I don't believe that ready answers exist. My point is that excluding such central issues from the conversation is retarding progress.

2.1 What is Availability?

In a purist world, "Availability" means in every format, at no cost, and documented to extent of being usable without assistance by anyone, but also with human assistance provided when there are problems. Even if this state of nirvana is the agreed-on goal, it will not materialize by magic, without resources. In §3, I discuss the implications for various stakeholders. Here the view is somewhat broader.

A major issue is that for many people, "Availability" is not the ultimate goal. I have argued elsewhere (Karr, 2012) that datasets are not, except to a few aficionados, the ultimate product of official statistics agencies. That product is, rather, the decisions by governments, businesses and individuals that are based on the data. Analogously, data availability has no intrinsic value to most people, whereas things that can be accomplished as a result of availability do have definable value.

Two of the most often articulated secondary values of data availability are replicability and reproducibility of research. That these are often confused does not help. I view the latter as narrower. For instance, can the results in a paper be reproduced by someone other than the authors who is given access to the data (but see §2.2 below) and the code? Reproducibility is, of course, of particular interest to journals. Replicability is broader: can a comparable (perhaps broader or narrower) experiment be performed? Precisely how access to data supports replication is sometimes nebulous. Even so, it seems clear that detailed understanding of the data generated in one experiment does usefully inform design of the next one.

Somewhere between these two sits the question of alternative analyses of extant data. For both good and bad reasons, many statisticians are convinced that they can improve published analyses. For them, "What are the data" is perhaps more salient than "What is availability?"

2.2 What Are The Data?

In the purist world, data are clearly defined, static over time, and free from errors and other quality problems. This is a fantasy. In reality, every working dataset involves most of:

Edits, to fix faulty or unreasonable data values, to the point of dropping some data points or variables.

Imputation, both of missing values and as part of the editing process.

Creation of new variables, by combining extant variables and linking to other datasets, as well as generated by the analysis process itself. Examples of the latter are predictions and residuals.

Many datasets also have undergone some form of **statistical disclosure limitation**, in order to protect data subjects or sensitive attributes.

At which step in this iterative, intellectually challenging, resource-intensive process, do "The Data" exist? What if some steps are taken, then reversed? Equally important, what about the metadata?

Why does any of this matter? The consequences are more practical than might be realized. Simply by defining "The Data" to mean "the data as I first received them," whether from someone else or by generating them myself, I could simultaneously make them available and thwart anyone else who wanted to use them. Such a person would almost surely be unwilling or unable (in part because I also did not make available the metadata and other key information) to duplicate the data development process I had performed.

If this all sounds abstract, let me attempt to make it concrete. Suppose I have conducted a survey, and define "The Data" to be the raw responses, *sans* edited or imputed values. Suppose I also define "The Data" not to include frame variables (used to select the sample), design variables (used for stratification) or weights. Suppose I also exclude not only metadata but also *paradata*, for instance, the number of times a person is contacted before he or she responds. Essentially every survey statistician I know of would say that such data are virtually useless.

3 Stakeholders and Incentives

Other than because it is a public good, data availability is a hard problem because there are many stakeholders, each subject to a multitude of incentives. It is facile, but also true, to answer the "Why is data availability such a hard problem?" question by simply saying "Look at the incentives." People are people, and organizations are organizations. For the most part, they behave rationally, and therefore predictably, with respect to prevailing incentives. Economically irrational organizations generally do not survive. To the extent that making data available interferes with such incentives, change is very problematic.

To make my own potential biases clear, I belong to several of the sets of stakeholders discussed below. I am a researcher. I am the CEO (Director) of a private-sector research organization, the National Institute of Statistical Sciences (NISS), whose existence depends on data. I have been a data subject, and not just because we all are. I have never worked for, but have worked with, funding agencies and journals.

3.1 Researchers

To a purist, for a researcher, learning the truth is its own reward. Saying that there is an immutable truth is, in itself, a gross simplification (Kuhn, 1962),[4] akin to the left-hand panel in Figure 1. Statements based on data are, instead, "to the best of our current knowledge and the best of our current ability to analyze the data." That learning the truth is enough is hopelessly inconsistent with reality.

Every researcher that I know is constantly seeking funding, publications and research team members (e.g., graduate students and postdoctorals). These are palpable incentives. Anything that diverts resources—principally, time and money—that could be used to pursue them carries a powerful disincentive. No one should be, and I suspect few people are, surprised when researchers act in their own self-interest. Just think: if there is a choice between devoting a month's time to writing proposals and devoting a month's time to writing documentation necessary to support making data available, what will a rational researcher do?

Failure to account for incentives also makes difficult distinguishing reluctance to provide data from unwillingness to incur the costs associated with doing so, leading to gratuitous criticism and suspicion. As a result, already low-content conversations become even noisier.

That researchers themselves routinely seek other researchers' data seems not to make them more responsive to requests for data, confirming my points.

[4]Data availability may parallel Kuhn's scientific revolutions in deeper ways than are articulated here. There is need for a paradigm shift, because there is accumulating evidence that the current paradigm is inadequate. What has not (yet) emerged is consensus on a new paradigm.

3.2 Research Organizations

Most discussions about data availability overlook the role of research organizations. Not very many researchers are self-employed, and almost all high-level research takes place within organizational settings. Organizations are complex, but organizational commitments, for instance, to data availability if other factors could be dealt with, are easier to obtain, easier to enforce and longer-lived than commitments involving individuals. As Director of NISS, I can commit everyone employed by NISS to certain kinds of behavior, and be held responsible if they do not comply. It seems to me apparent that the conversation about data availability should focus on organizations, not on individual researchers; see also Young and Karr (2011).

At the same time, organizations are dizzyingly complex, with many incentives and multiple stakeholders of their own. As Director, I am responsible to the Board of Trustees of NISS for the financial health and continued existence of the organization. I must safeguard our assets, including our corporate reputation, as well as see that investments (primarily of employees' time) yield as much return as possible. Data are an asset. As an aside, the same is true of for-profit corporations, a fact that constantly seems to surprise some people. Corporations are sometimes portrayed as profit-hungry monsters that will resort to anything. In fact, a CEO who fails to make every effort to maximize profits may be derelict in his or her responsibility to the shareholders.

What does this mean for data availability? As do all research organizations, NISS needs researchers, revenue and visibility. We have made value-added modifications to every dataset we have ever worked with, only some these modifications were paid for by the sponsor of the research. Such datasets are legitimate intellectual property. If providing them to others involved no further investment (e.g., in documentation), doing so might be rational. In the absence of compulsion, investing additional resources is simply irrational. Although perhaps exaggerated, the thought of having to translate data into multiple formats and run a 24/7 help desk for those attempting to use our data is truly frightening.

I should note that NISS, as do many other organizations, works with many restricted datasets provided to us by U.S. government agencies under license, in research data centers or under non-disclosure agreements with both non-profit and for-profit corporations. It is not our prerogative to release such data. Broad-brush discussions of data availability that fail to consider this component of the problem are incomplete at best.

3.3 Data Subjects

Not all data are about human subjects. Higgs bosons and sea surface temperatures have few rights, and so this subsection does not apply to some datasets.

In a purist world, data subjects participate for the common good. In reality, some (e.g., students in psychology courses) are compelled to participate. Others, including participants in some government surveys, are paid. Still others see direct personal benefit, such as treatment of a disease or access to medical care.

Privacy is a central problem. As noted in §1, privacy can be viewed as a public good, even though is it not non-excludable. Most data subjects cannot assess whether their data are being protected. Many would happily allow someone else's privacy to be compromised if they perceived this as benefitting them. To be sure, failure to observe privacy promised to subjects is wrong. But, privacy has become a hiding place (in the extreme, the "last refuge of scoundrels") far out of proportion to the issues. A plethora of techniques for statistical disclosure limitation (SDL) exists, mainly in the official statistics literature. Many of these explicitly balance disclosure risk and data utility (Cox et al., 2011). At the conceptual level, data subject privacy is a completely addressable issue.

The same is not true operationally. Relatively few organizations and only a relative handful of researchers possess the expertise to make an informed selection and implementation of SDL for a given dataset. Expending their own resources to fill this gap is, once more, irrational. The potential solutions discussed in §4 address this problem by placing SDL expertise in data archives.

Discussions of data availability might not be driven by the incentives operating on data subjects, but do need to be cognizant of them. Radically different approaches, such as paying subjects to relinquish privacy, may be relevant. Lest readers be appalled by this suggestion, let me point out the obvious: mobile telephone users, voluntarily and with no compensation, relinquish massive amounts of their privacy.

Trust by data subjects is also crucial, and often absent from conversations. Consent forms that are explicit about availability may drive subjects away, with unpredictable but negative consequences for data quality.

3.4 Funding Agencies

Most of this section deals with public (that is, government) funding agencies such as the U.S. National Science Foundation (NSF) and National Institutes of Health (NIH), with briefer comments on private agencies (foundations).

In a purist world, a funding agency with perfect predictive powers would allocate its resources in a way that maximizes social benefit, as measured for instance by global economic competitiveness or job creation. Under some interpretations of social benefit, this might lead to under-funding of what is usually termed basic research, an issue of which agencies are clearly aware.

In reality, as organizations, funding agencies have an imperative for continued survival. Most immediately, this means maintaining the support of the groups who formulate and approve their budgets. (In the U.S., funding agencies are either part of executive branch departments or so-called "independent agencies;" in either case, their budgets must be approved by the legislative branch.) Crucial to this process is producing enough big-time research successes to sustain the favorable attention of those who control budgets. U.S. agencies vary substantially and substantively in their tolerance and presentation of failed research. At one extreme, the Defense Advanced Research Projects Agency (DARPA) aggressively promotes its undertaking high-risk, high-payoff research. Other agencies protest, not entirely incorrectly, that there is no such thing as failed research, because something is always learned.

In addition to incentives associated with budgets, funding agencies also face incentives with respect to their other principal set of stakeholders—the researchers whom they fund. Without proposals, the agency has no case for continued existence. To some extent those submitting proposals are a captive market, but this may not always be so. In any event, driving them away is dangerous.

The link to data availability is again through incentive-based reasoning. Several U.S. funding agencies have proposed and implemented measures to increase data availability. These are a step in the right direction. (Albert Bowker, the first chair of the Board of Trustees of NISS, said at the inaugural meeting of the board that "It is easier to take a step in the right direction than to be able to specify the destination." In the case of data availability, we do need also to think hard about the destination.) However, the details reflect the lack of clear definitions of data and availability, rather than the relevant incentives. Loopholes abound, especially delays and claims of confidentiality.

The current agency measures are also *unfunded mandates*. Although it seems to have lost currency, this term was in widespread use in the U.S. 10–15 years ago, to describe government mandates (for instance, for student performance on educational assessments, in the Bush administration's 2004 No Child Left Behind

legislation) for which no funds are provided to achieve. I will return to this point in §4, but I think that unfunded mandates for data availability are uniformly destined to be problematic, or simply to fail.

Foundations, the principal private sector funders of research, do not seem notably more advanced in their thinking about data availability than public agencies. Moreover, they lack some options, such as regulations, available to government agencies. Indeed, private entities cannot create public goods. There may be opportunities, however, for foundations to catalyze creation of the data archives discussed in §4, for instance, by means of seed money and support for creation of governance structures.

3.5 Journals

In a purist world, journals are selfless guardians of scientific truth, publishing (only) what is important, novel and correct. And, indeed, many journals do publish what they (that is, their editors and referees) deem is important, novel and correct, albeit sometimes with incomplete success. But journals are subject to other, equally powerful, incentives. Even those published by nonprofit professional societies are businesses, competing for authors, readers, visibility, and in some cases, advertisers.

These incentives are not perverse, and their consequences should not be surprising. Given the incentives, it is perfectly rational for a journal to seek to publish papers that will be covered in the *New York Times*' Tuesday science section, even if it means "cutting corners" on importance, novelty, correctness or data availability. A retraction, if covered in the media at all, appears on inside pages, or at least below the fold. Expecting journals to ignore such incentives is optimistic to say the least. That some do is encouraging.

I believe, however, that something deeper is also relevant. In some fields (one of which, computer science, is a neighbor of statistics), publication seems to be less an assertion of correctness than one of potential importance. If the ideas and results are novel, interesting and possess sufficient potential impact, then they should be published, in order that they can be subjected to broader scientific scrutiny. (The extreme opposite is true for some mathematics journals, which seem to encourage line-by-line refereeing in order to ensure correctness, a process that is demonstrably ineffective.) Publication is a step (ideally, in right direction), not a destination.

4 Potential Solutions

I have been accused of being a cynic about data availability. In return, without being fully able to define the difference, I argue that I am a realist, and I have tried in the previous sections to be specific about what being a realist constitutes. Moreover, I do not believe that there is no solution to the problem, only that workable solutions do not look a lot like what most people are thinking about currently.

4.1 The Case for Data Availability

I am a proponent of data availability. Consistent with previous sections but unlike some other people, I believe that the argument for data availability must be made on grounds of efficiency, not principle.

The discussion in §3.5 implies that journals are the central mechanism in replicability. Many authors, notably Ioannidis (2005), have noted what they assert to be unreasonably high rates of failure to replicate, especially in contexts such as observational medical studies. In the absence of any knowledge of what the rate of failure to replicate rate should be, it is arguable that failure to replicate actually means merely that

scientific process is working perfectly. What is potentially important is made public in order to allow others, especially skeptics, the opportunity to test it.

However, a functioning process need not be an efficient one. If the problem lies in the analysis of the data (and there are examples where this has been demonstrated, e.g., Young et al. (2009)), then conducting a new experiment (say, a randomized clinical trial rather than an observational study), can be a monstrously and avoidably expensive way of securing data than can be analyzed differently.

4.2 A Possible Scenario

Attempting to tie things together, my central theses are first, that data availability is a public good, which can be attained only if people and organizations undertake actions that directly contradict other incentives they face. Second, and therefore, unfunded mandates for data availability are predestined to fail.

Historically, public goods are provided through centralized mechanisms, which allows creation of focused expertise and can foster efficiency. (Centralization can also produce ponderous, unresponsive systems—"bureaucracies"—that thwart rather than advance the original purposes.) Clearly this approach has not always worked as well as hoped, but history shows that essentially nothing else has worked at all. Whether the internet will prove an exception remains to be seen; I predict that it will not.

The implication, I believe, is that the mechanism to achieve data availability is (one or more) government-operated or government-sponsored data archives. Funding would come from government. If data availability is truly a public good, then the source of the funds should be general revenues, but it is also possible to envision models in which there are also (explicit or implicit[5]) user fees. Data archives would be of sufficient scale to operate efficiently, and possess the expertise necessary, for instance, to create high-quality metadata, characterize data quality meaningfully (Karr et al., 2006), deal with data versioning, perform principled disclosure limitation, and disseminate data in multiple formats. They will have credible longevity.

If use of archives is mandated, who does the mandating? Because they are few in number and do hold genuine sway, I think that funding agencies are the most efficient choice. If their mandate is no longer unfunded, it will be much more potent. For reasons laid out in §3.2 and Young and Karr (2011), the mandates should be imposed on organizations, not individuals. When an organization risks losing all of its government funding if it does not comply, it will pay attention.

No system is perfect, however. Mandated data archives will, inevitably, be massively inefficient. Most datasets placed in them will never be accessed, in the same way that most papers are never cited by anyone other than their authors. I see no ready solution to this problem.

There is also the issue of equity. National defense benefits everyone approximately equally, so everyone pays approximately the same amount for it, adjusted for ability to pay by means of progressive taxation. There are no user fees for defense. Roads do not benefit everyone equally, which is reflected in the existence of user fees (tolls and gasoline taxes) for them. Where data availability lies in this spectrum seems still unclear.

4.3 Final Remarks

I am not under any illusion that what I propose is straightforward to achieve. I do believe that my proposed solution is cognizant of the complexities and incentives discussed here, more so than most other proposed

[5] An example of the latter is partial support of road construction by means of taxes on gasoline. Roads become, in effect, partially public goods.

solutions. Public goods are never dealt with easily. On a completely different scale, at the time this is written (December 2013), the U.S. Congress and populace are engaged in divisive, shrill and necessary debate concerning the extent to which medical care and elementary/secondary education are public goods.[6] Opinions regarding the nature and role of central governments differ profoundly. Only with trepidation do I cast data availability into these waters, but I am completely convinced that what I am advocating is necessary.

Acknowledgements

This work was supported by the National Science Foundation grant SES–1131897 to Duke University and the National Institute of Statistical Sciences. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author and do not necessarily reflect the views of the National Science Foundation.

I am grateful to Murray Cameron for organizing the 2013 World Statistics Congress session at which Karr (2013) was presented, and for inviting me to participate in it. I also thank Stanley Young of NISS for numerous discussions of data availability issues.

References

Abowd, J. M. (2013). Presentation: Revisiting the economics of privacy: Population statistics and privacy as public goods. Available on-line at http://digitalcommons.ilr.cornell.edu/cgi/viewcontent.cgi?article=1009&context=ldi.

Cox, L. H., Karr, A. F., and Kinney, S. K. (2011). Risk-utility paradigms for statistical disclosure limitation: How to think, but not how to act (with discussion). *Int. Statist. Rev.*, 79(2):160–199.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8): e124. Available online at medicine.plosjournals.org/perlserv/?request=get-document&doi=10.1371%2Fjournal.pmed.0020124.

Karr, A. F. (2012). Discussion on statistical use of administrative data: Old and new challenges. *Statist. Neerlandica*, 66(1):80–84.

Karr, A. F. (2013). Why data availability is such a hard problem. Paper presented at IPS 108 of the 2013 World Statistics Congress of the International Statistical Institute, Hong Kong.

Karr, A. F., Sanil, A. P., and Banks, D. L. (2006). Data quality: A statistical perspective. *Statistical Methodology*, 3(2):137–173.

Kuhn, T. (1962). *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago, IL.

Stiglitz, J. E. (1999). Knowledge as a global public good. In Kaul, I., Grunberg, I., and Stern, M. A., editors, *Global Public Goods: International Cooperation in the 21st Century*. Oxford University Press, New York.

[6] Many other countries have long since decided that they are, of course.

Varian, H. R. (1992). *Microeconomic Analysis*. W. W. Norton, New York. Third edition.

Young, S. S., Bang, H., and Oktay, K. (2009). Cereal-induced gender selection? most likely a multiple testing false positive. *Proc. Royal Soc., Series B*, 654:1211–1212.

Young, S. S. and Karr, A. F. (2011). Deming, data and observational studies: A process out of control and needing fixing. *Significance*, 8(3):116–120.