Killing with Kindness: The Attack on Public-Use Data

Martin H. David, University of Wisconsin

■ A Case of the Jitters

Attacks of disclosure anxiety hit the Federal data-producing establishment at least once every ten years. They are precipitated by the legal responsibilities of data producers and the inability of data producers to audit every use of public data products. The tension produced by risks of disclosing data on respondents is real. And informed scientists have been aware since the 1970's that determined users can frustrate attempts to prevent reidentification of respondents.

Tension is also mounting because of public resistance to supplying data. That resistance means lower response rates on surveys and censuses. It means Congressional reluctance to fund the mission of agencies that collect statistics from administrative records. For example, funding for the Statistics of Income Division of the IRS was seriously impaired during this decade, as the information function of SOI, creating quality statistics from administrative records, got overwhelmed by concerns about enforcement activity of the IRS. The public has no clear picture of "statistical use" of its administrative records. It is hardly aware of the firewalls that separate Census and tax documents from use by Government employees in general. In contrast, daily assault by telemarketing and promotional mailings reminds the public of their inability to control information about themselves. It is clear to many recipients of bulk mailing, credit card offers, and telephone merchandising that a vast network links innocent private pursuits--having a child, moving, or graduating from college--to a massive intrusion of unsolicited commercial advertising. The public rightly infers that much private activity is scanned, linked, evaluated, and targeted to the benefit of commercial interests.

All Federal data collectors seek to overcome increasing public resistance to supplying data. Some members of the Federal data establishment, I fear, hope that restricting release of public-use data will offset nonresponse engendered by intrusive commercial activity. That hope appears unwarranted, when one considers the scope and frequency of private marketing relative to data collection by governments. In the comments below, I argue two theses:

- Eliminating public-use data will create more damage to the quality of Federal statistics than any benefit which might emerge from a more positive attitude of data suppliers.[1]
- Protecting the privacy and interest of data suppliers lies in creating institutions that regulate access to data, rather than in expanding technical devices for disclosure limitation.

I develop these points under three headings:

- Research use of data benefits the public by hybridization.
- "System Software" outperforms "statistical disclosure limitation" in reducing risk.
- Public use generates stewardship for the data warehouse.

■ Research use of data benefits the public by hybridization.

The knowledge industry includes data collectors and users of data in the research sector outside of data-collecting agencies. Both rely on advanced statistical science. The missions of collectors and users differ markedly. The statistical establishment has a legal mandate to measure certain aspects of society. That mandate has induced periodic data collection and timely release of indicators--from retail sales to population counts in congressional districts. The indicators are essential for many policy decisions and for the orderly administration of government. The mission is "sponsored" by legislative appropriations to specific data-collection agencies.

Research users have a different mandate. They engage in the discovery of paradigms for interpreting measures, estimating models that display relationships among measures, and making inferences. Each activity requires timely, reliable, and accurate data. Research use adds to the stock of knowledge, albeit erratically and with content that is occasionally revolutionary.

Public-use microdata join user analysts to the experts who collect the data. That join creates a hybrid understanding of statistical science used for data collection and analysis. Both analysts and experts have highly developed skills and experience in pursuing their respective missions. But the outlook of experts and analysts differs because their missions differ. Their experience and knowledge do not coincide. Those differences create feedback for the data-collecting agencies. User analysts identify errors in the data that may not be detectable within the scope of the collecting agency's mission (e.g., Scholz, 1994, discovered the difference between persons using Earned Income Tax Credits and persons legally entitled). Analysts provide new paradigms to link measurements. For example, economists have clarified the relationships between income and value of corporate entities. Analysts have identified the difference between the distribution of survival probabilities in a population and the distribution of ages in the population at a point in time. Such differences in paradigm have been critical to the continuing evolution of data collection methodologies. Cross-section studies reveal the distribution of age in a population at one point in time. Panel data are required to observe "entry and exit" probabilities. Panel capabilities have been important in debate on welfare policies, and they reveal new insights into the viability of US business enterprise. Analysts have identified heterogeneity among business entities that leads to "churning" or gross flows larger than net flows--whether the domain of interest is job creation or change in inventories.

Hybridization--the exchange of information between data collectors and data users--generates valuable (and unpredictable) improvements in the quality, coverage, and specification of data collections. This by-product of public use data cannot easily be programmed into the data-collecting agency's mission. This by-product of data use is a benefit *in addition* to the scientific publication of users' analyses.

The creation of a community of data users also works to defend the mission of the data-collecting agencies against short-sighted micromanagement from outside (e.g., the Congress), which can compromise the long-term development of professional data collection.

Reducing Disclosure Risk--Secure Sites and Secure Data

Historically, disclosure risk has been reduced by limiting access to the data to vetted persons, typically special sworn employees, in *secure locations*. Special sworn employee status binds the user to the same penalties for disclosure as employees of the data-producing agency. Limiting disclosure risk by operations on microdata came to the United States with the *1960 Census of Population*. Public access microdata (*secure data*) were released for widespread use.[2] Public access implied no limits on the user of data. Many of the hybridization benefits that we have obtained since 1960 in the US derived from secure data. Now, we know that eliminating risk of disclosure is impossible and that the potential for making inferences about individuals in microdata is substantial because many data bases are maintained by public and private owners.

It would be counter-productive to deny research access to data on the existing scale because of a logical argument that demonstrates increased disclosure risk. Instead, we need to find risk-reducing alternatives that can be combined with increasing research use at *secure sites* and continuing research use of not so *secure data.*

In the past, many data sets that warrant intensive scientific scrutiny have been almost inaccessible. Disclosure risks are cited as the primary reason. I review two kinds of data briefly to illustrate the problem. Analysis of tax returns has been restricted to secure sites and a limited number of individuals who are employed by the Office of Tax Analysis, the US Treasury Department, the Joint Committee on Finance, and the Congressional Budget Office. Despite great scientific interest in wealth, access to data on estate and inheritance re-

turns has been restricted because of the high risk of disclosures about identifiable wealth holders. Though economic debate has raged about the behavioral responses to tax law, access to the corporate tax model prepared by Statistics of Income has been restricted because some business entities are included with certainty and the concentration of sales makes it easy to identify industry leaders. Each of these data sources warrants scientific analysis beyond the substantial accomplishments of the Treasury, the Congressional Budget Office, and the Joint Committee on Taxation. These microdata cannot be accessed except by special sworn employees. Be clear that the risks associated with the estate data include the invasion of privacy of heirs. The risk associated with corporate return data do not have a privacy dimension but may fuel insider trading or insight into proprietary information. Access needs to be controlled by a process that is secure, most likely involving a secure site, but is not limited by collecting agency funding, priorities for publication, or excessively narrow constructions of mission.

A similar story can be told about data from the Economic Censuses, where research access to microdata was extremely limited before the 1990's. For more than a decade, the Bureau of the Census has pioneered in developing its Center for Economic Studies and its regional Data Centers as secure sites for research access.

These examples leave us with a paradox. The research community can exploit microdata. Data collectors incur disclosure risk by providing access to those data. How can we minimize disclosure risk without compromising access? Economists answer with a slogan: "Internalize the externalities," that is, place the liability for the risks of disclosure on users. This can be done with a relevant "system software" for public data release. I call it contracted access.[3] Contracted access can be used to augment the use of secure sites; it can also be used to formalize the responsibilities of persons who access public-use data.

■ "System Software" outperforms "statistical disclosure limitation" in reducing risk.

The institutions through which public-use microdata are disseminated affect the risk of disclosure. Our past policy sought no credentials for access to public-use data. Past policy almost never regulated "redissemination" or copying of data for use by others. Contracted access appears to be less risky. It is time to test the feasibility of an alternative.

Procedures for disseminating public-use files should encompass two principles. Public-use microdata should be freely available for research. Uses of microdata other than research--promotion, surveillance, and further data dissemination should be prohibited. Research use should be defined by three criteria:

- The activity is potentially capable of producing a publishable report.
- No report may disclose the identity of any data providers.
- All products of the activity must become public goods that can be widely accessed within a year of their creation.

The first criterion eliminates unstructured scanning of data sets for titillating combinations. The second assures protection of the data providers. The third excludes the production of analyses that will become the permanent private property of a single client. For example, an analysis of market shares could be performed for a single client, but would fulfill the public goods rule if it were published in less than a year after delivery.

One additional restriction on use is required:

- Public users may not redisseminate data. All users must register with the agency that releases public data. Redissemination will be penalized.[4]

Every user of public microdata can be asked to affirm that he or she will abide by these four restrictions. Users' behavior can be monitored by requiring regular reports on titles of research products and evidence that products are (or will become) publicly available. This kind of reporting is required by grant-giving agencies, and is used by producers of privately collected research data (the *Panel Survey of Income Dynamics*, the *Lux-*

emburg Income Studies, and the *Socio-Economic Panel Study*). Control of resdissemination has already been incorporated in household microdata that were linked to the Death Index by the University of Michigan. Unauthorized users can be identified by their data use (Juster, 1991).

Establishing eligibility for the use of public data entails some inconvenience on the part of users and administrative cost for the disseminating agency. These costs are not large. For example, registration for permission to use public access data can be accomplished on the Internet; the validity of the registration can then be verified through e-mail correspondence. "Boilerplate" that goes with the agreement can specify stipulated damages that the user incurs for non-compliance. Penalties should include fines and exclusion from future access to microdata resources for a period that would clearly impair the career of the offender.

Contracted access also has benefits. Collecting citations of data analyses will uncover and trace many items that are not captured by bibliographic data bases. Contracting for access will trigger awareness of liability on the part of users and the organizations to which they belong. That awareness is appropriate and should serve as an incentive to responsible data use.

Contract access for public-use data and the associated penalties are not sufficient to make the system of data release self-enforcing. Data producers and users must commit to three additional activities: Education, professional standards, and the creation of institutions that have the mission of monitoring access to data and the attending use of data.

Why is education needed? Many users have not contemplated the damage that could be caused by inadvertent or intentional disclosures. They do not understand that confidentiality pledges are essential to motivating respondents to supply data. Many users know nothing about privacy rights and the data suppliers' expectation of being protected by fair information practices. Users need to understand that businesses have withheld data in the past, in situations where concern about disclosure is a likely cause (Manser, 1993). Both data collectors and professional training can assist in helping users learn the appropriate use of microdata products. Professional training certainly must include learning about responsible data use. Data collectors can embed training in many of the procedures that are used in disseminating data.

Professional standards? If statisticians, economists, and social scientists cannot agree on methods, why should they agree on standards for handling data? The professional organizations to which data producers and analysts belong can help to offset the "surfing" culture of the Internet--many users of the Internet grab goodies, enjoy an information high, and take no interest in the physical, financial, and intellectual infrastructure that makes Internet possible. The professional community must create institutions that maintain norms for responsible user behavior by data analysts.

Institutions for monitoring access to data and attending uses of data are needed because the value of public access to particular data increases over decades. In that long time span, the original data collectors may be reorganized and their memory for public data extinguished, though the data continue to be available. A consortium of data producers and users should examine how they can benefit from an institution that is at arm's length from both data users and data collectors. That institution can be organized to offer advice, standards, and prototypes for solutions to the problems of access to data within a socially acceptable level of disclosure risk.

I suggest that widespread use of microdata be enabled through contracted access governed by criteria for research use. Ancillary education, standards, and creation of institutions for monitoring access and data use complement contracted access. This portfolio does not eliminate all risks of disclosure. No alternative can do so. It is especially important to discount the fantasy that secure data sets can be created by technical operations on data. For example, disseminating synthetic microdata is impractical. The procedure required to simulate real data will be exorbitantly costly because it absorbs the scarcest resources in the statistical system. Extremely knowledgeable professionals will be required to use their understanding of data solely to encrypt the real world. For the encryption to work, conditional distributions of each variable in relation to all other variables must be

known for all possible analysis populations that might be drawn from the data. The task of discovering those distributions is orders of magnitude larger than the task facing research analysts who add to the stock of knowledge. Resources used in creating synthetic data sets are better put to use in analyzing the original data. Furthermore, the time required to create synthetic data sets assures that public-use data will never be timely. The Rubin-Trieman effort to impute occupational classifications to the 1980 Census is an example.[5] The best brains in statistics were tied up in a procedure that produced viable output after six or eight years of development. Public access to synthetic data at such a late date lost most of their policy relevance. Other sources of information were used as substitutes for the capabilities of the Census.

■ Conclusion

Hybridization of the research-user and data-collector communities through widespread access to microdata produces serendipitous improvements in measurement. Access to those microdata can be successfully continued through "contracted access" that will reduce risks of disclosure. Contracted access disciplines users to engage in appropriately risk-free analysis and publication. Contract access commits users to four verifiable standards of research use. Professionals in the research community can agree to implement this new "system software" for access to public-use microdata. Implementing a contractual obligation will create a demonstrable mechanism for enforcing responsible data use. Incentives for responsible data use appear far more desirable than curtailing public use or squandering scarce resources on manipulating microdata to reduce disclosure risks.

In conclusion, I would like to suggest one additional and extremely important outcome from a program for disseminating public-use microdata.

■ The Last Word: Public use generates stewardship for the data warehouse.

Observational sciences are built on a base of historical data. The dependence of astronomy on the observations recorded by Copernicus makes that clear. Understanding human behavior in an evolving social system is clearly a task for observational science, as experiments can never be conducted on a sufficient scope or time scale to provide definitive information on all aspects of the social system we seek to understand. If the social sciences are to progress, it is clear that we need to keep microdata in an ever-growing warehouse for continuing re-examination.

Keeping the warehouse has not been done well in the past. Decades of data from the Census of Manufactures in the last century were lost (Bohme, 1987). The *1950 Survey of Consumer Expenditures* was lost (David, 1980). The existence of public-use data increases the probability that data we now have in hand will not be lost in the future. The system required for producing public-use microdata increases the likelihood that data will be available. Disseminating data for research generates metadata that describe the measurement process and data produced. It creates a process for the dissemination of data, as well as estimates. Dissemination increases the number of places where data are stored, protecting against loss of data.

Without metadata, public use of the data is impossible. Increasing the number of users increases the complexity of their analyses. Both the scale and scope of use induce the creation of more complete metadata: metadata that include error notifications, bibliographies of past work, and contextual variables that have been added in prior study.

The institution for the dissemination of data assures that someone is responsible for archiving the data. In the last decade, it has become clear that archiving entails continuing renewal of the data resource to be compatible with changing technologies for storage of data. If we do not attend to creating an adequate archive, we will be unable to unlock secrets from past data collections that could be unlocked through meta-analysis, study of repeated measures, etc.

Even with an institution that is responsible for archival data, the hazards of nature and wars imply that we will lose data irrevocably unless there are multiple storage vaults. Using the user community as an archival resource is extremely attractive and can be encouraged.

References

Bohme, Frederick G. (1987), US economic censuses, 1810 to the present, *Government Information Quarterly,* 4/3: pp. 221-243.

Clogg, Clifford C.; Rubin, Donald B.; Schenker, Nathaniel; Schultz, Bradley; and Weideman, Lynn (1991), Multiple imputation of industry and occupation codes in Census public-use samples using Bayes logistic regression, *Journal of the American Statistical Association*, 86: pp. 68-78.

David, Martin H. (1980), Access to data: The frustration and utopia of the researcher, *Review of Public Data Use.* 8: pp. 327-337.

Duncan, George T.; Jabine, Thomas B.; and deWolf, Virginia A. (eds.), *Private Lives and Public Policies*, Washington, DC: National Academy Press.

Juster, F. Thomas (1991), Discussion, *Proceedings of the Social Statistics Section of the American Statistical Association.*

Manser, Marilyn (1993), *Price measurements and their uses,*Betsock, Thomas and Gerduk, Irwin. The problem of list prices in the PPI: the Steel Mill products case, pp. 261-274 and Foss, Murray F. (1993). Does government regulation inhibit reporting of transactions prices by Business?, pp. 275-301.

Marsh, C.; Dale, A.; and Skinner, C. (1994), Safe data versus safe settings: access to data from the British Census, *International Statistical Review,* 62/1: pp. 35-54.

National Research Council (1979), *Privacy and Confidentiality as Factors in Survey Response.*

Scholz, Karl (1994), The earned income tax credit: participation, compliance, and anti-poverty effectiveness, *National Tax Journal*, pp. 59-81.

Treiman, D. J.; Bielby, W.; and Cheng. M. (1988), Evaluating a multiple imputation method for recalibrating 1970 U.S. Census detailled industry codes to the 1980 standard, in C.C. Clogg (ed.), *Sociological Methodology 1988.*

Footnotes

1 Cognitive research that can be supported in the laboratories of the Bureau of the Census, the Bureau of Labor Statistics, and the National Center for Health Statistics is needed to understand whether there is *any* correlation of worries about privacy to the existence of public-use data products.

2 Marsh, Dale, and Skinner (1994) coined the distinction between secure data and secure locations.

3 Statistics of Income offers public-use and special sworn employee (SSE) access. Restricted access (RA) is less stringent than special employee and has worked well for the National Center for Educational Statistics, and in other settings. Contracted access (CA) is an alternative to public use.

Alternative	Secure data?	Secure site?	User obligations
Existing			
Public-use microdata	Essential	No	None
SSE	None		Liability identical to other employees
New			
CA	Possible	Possible	Obligations of user defined by contract. Violation subject to damages.
RA	None	Possible	Bonding, institutional liability for employer

4 At least one exception to this rule might be useful. Publishers of scientific journals may require submission of the data used in published articles.

This rule facilitates constructive criticism. A serious critic needs to access the author's transformed data, as well as source data, necessitating some redissemination of data.

[5] See Treiman, D. J.; Bielby, W.; and Cheng, M. (1988) and Clogg, Clifford C.; Rubin, Donald B.; Schenker, Nathaniel; Schultz, Bradley; and Weideman, Lynn (1991).