



DIGITAL ACCESS TO SCHOLARSHIP AT HARVARD

Ensuring the Data-Rich Future of the Social Sciences

The Harvard community has made this article openly available.
[Please share](#) how this access benefits you. Your story matters.

Citation	King, Gary 2011. Ensuring the Data-Rich Future of the Social Sciences. <i>Science</i> 331, no. 6018: 719–721.
Published Version	doi:10.1126/science.1197872
Accessed	February 16, 2015 5:20:46 PM EST
Citable Link	http://nrs.harvard.edu/urn-3:HUL.InstRepos:12724029
Terms of Use	This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA

(Article begins on next page)

Ensuring the Data Rich Future of the Social Sciences¹

Gary King²

Abstract

Massive increases in the availability of informative social science data is making dramatic progress possible in analyzing, understanding, and addressing many major societal problems. Yet, the same forces pose severe challenges to the scientific infrastructure supporting data sharing, data management, informatics, statistical methodology, and research ethics and policy, and these are collectively holding back progress. I address these changes and challenges and suggest what can be done.

Fifteen years ago, *Science* published predictions from each of 60 scientists about the future of their fields (1). The physical and natural scientists wrote about a succession of breathtaking discoveries to be made, inventions to be constructed, problems to be solved, and policies and engineering changes that might become possible. In sharp contrast, the (smaller number of) social scientists did not mention a single problem they thought might be addressed, much less solved, or any inventions or discoveries on the horizon. Instead, they wrote about social science scholarship, how we once studied *this* and in the future we're going to be studying *that*.

Fortunately, the editor's accompanying warning was more prescient: "history would suggest that scientists tend to underestimate the future" (2).

Indeed. What the social scientists did not foresee in 1995 was the onslaught of new social science data, enormously more informative than ever before, and what this information is now making possible. Today, huge quantities of digital information about people and their various groupings and connections are being produced by the revolution in computer technology, the analogue-to-digital transformation of static records and devices into easy-to-access data sources, the competition among governments to share data and run randomized policy experiments, the new technology-enhanced ways people interact, and the many commercial entities creating and monetizing new forms of data collection (3).

Analogous to what it must have been like when they first handed out microscopes to microbiologists, social scientists are getting to the point in many areas where enough information exists to understand and address previously major intractable problems that affect human society. Want to study crime? Whereas researchers once conducted victimization surveys, huge quantities of real time geocoded incident reports are now available. What about the influence of citizen opinions? Adding to the venerable random survey of 1,000 or so respondents, researchers can now harvest more than 100 million social media posts a day and use new natural language methods to extract relevant information (17). At the same time, parts

¹My thanks to Micah Altman and Merce Crosas for helpful comments on an earlier version.

²Albert J. Weatherhead III University Professor (Institute for Quantitative Social Science, 1737 Cambridge St, Harvard University, Cambridge, MA 02138; <http://GKing.Harvard.edu>, king@Harvard.edu; 617-500-7570).

of the biological sciences are effectively becoming social sciences, as genomics, proteomics, metabolomics, and brain imaging produce huge numbers of person-level variables and researchers in these fields join in the hunt for measures of behavioral phenotypes. In parallel, computer scientists and physicists are delving into social science data with their new methods and data collection schemes.

The potential of the new data is huge, and the excitement in the field is palpable. The fundamental question is whether researchers can find ways of accessing, analyzing, citing, preserving, and protecting this information. Although information overload has always been an issue for scholars (11), today the infrastructural challenges in data sharing, data management, informatics, statistical methodology, and research ethics and policy risk being overwhelmed by the massive increases in informative data. Many social science data sets are so valuable and sensitive that, when commercial entities collect them, they grant external researchers almost no access. Even when sensitive data is collected originally by researchers, or acquired from corporations, privacy concerns sometimes lead to public policies that require the data be *destroyed* after the research is completed - a step that obviously makes scientific replication impossible (15) and some think will increase fraudulent publications (16).

Indeed, we appear to be in the midst of a massive collision between unprecedented increases in data production and availability about individuals and the privacy rights of human beings worldwide, most of whom are also effectively research subjects.

Consider how much more informative to researchers, and potentially intrusive to people, the new data can be. Researchers now have the possibility of continuous time location information from cell phones, Fastlane or EZPass transponders, IP addresses, and video surveillance. We have information about political preferences from person-level voter registration, primary participation, individual campaign contributions, signature campaigns, and ballot images. Commercial information is available from credit card transactions, real estate purchases, wealth indicators, credit checks, product RFIDs, online product searches and purchases, and device fingerprinting. Health information is being collected via electronic medical records, hospital admittances, and new devices for continuous monitoring, passive heart beat measurement, movement indicators, skin conductivity, and temperature. Huge quantities of information in unstructured textual format are being produced in social media posts, emails, product reviews, speeches, government reports, and other web sources. Satellite imagery is increasing in resolution and scholarly usefulness. Social everything -- networking, bookmarking, highlighting, commenting, product reviewing, recommending, and annotating -- has been sprouting up everywhere on the web, often in research-accessible ways. Participation in online games and virtual worlds produce even more detailed data. Commercial entities are scrambling to generate data to improve their business operations through tracking employee behavior, website visitors, search patterns, advertising click-throughs, and every manner of cloud services that capture more and more information.

Efforts in the social sciences that make data, code, and information associated with individual published articles available to other scholars have been advancing through software, journal

policies, and improved researcher practices for some time (8,9). However, this movement is at risk of collapsing unless the improvements in methods for sharing sensitive, private, or proprietary data (13) are able to be modified fast enough to keep up with the changes in the types and quantities of data becoming available, and unless public policy adapts to permit and encourage researchers to use them. The necessary technological innovations are more difficult than it may seem. For example, the venerable strategy of anonymizing data is not very useful when, for example, date of birth, gender, and ZIP code alone are enough to personally identify 87% of the U.S. population (4). And the cross-classification of ten survey questions of ten categories each contains more unique classifications than there are people on the planet. And now think of the challenges of sharing continuous time cell phone location information from a whole city, or biological information with hundreds of thousands of variables. The political situation is also complicated, with a media storm generated by each new revelation of how personal information becomes publicly available, but at the same time citizens are voluntarily giving up more privacy than ever, such as via the rapid transition from private email to public or semi-public social media posts.

If privacy can be protected in a way that still allows data sharing, considerable progress can be made for people everywhere without harm coming to any one research subject. This seems easier than, for example, the situation with most randomized medical experiments, where if everything works as expected those in one treatment arm will be harmed relative to those in the other arms. Moreover, most concern about data sharing involves individuals whereas social scientists usually seek to make generalizations about aggregates, and so spanning the divide is often possible with appropriate statistical methods.

What can we do to take advantage of the new data, while facilitating data sharing and at the same time protecting privacy? First, before we try to convince other parts of society to give us some leeway, social scientists need to get our own act together. At present, large data sets collected by social scientists in most fields are routinely shared, but the far more prevalent smaller data sets that are unique or derived from larger data sets, are regularly lost, hidden, or unavailable -- often making the related publications unreplicable. In most cases, many data sets associated with individual publications, and the related computer code and other information necessary to reproduce the published tables and figures from the input data, is not available unless you obtain permission of the original author, with no rules governing when access must be provided. This deserves serious reconsideration and action. We need to devolve web visibility and scholarly credit for the data to the original author while ensuring that the data are professionally archived with access standards formalized in rules that do not require ad hoc decisions of or control by the original author (5,6).

Second, we need to nurture the growing replication movement (7,12). More individual scholars should see it as their responsibility to deposit data and replication information in public archives, such as those associated with the Data Preservation Alliance for the Social Sciences (10). More journals should encourage or require authors to make data available as a condition of publication, and granting agencies should continue to encourage data sharing norms. More importantly, when we teach, we should explain that data sharing and replication is an integral

part of the scientific process. Students need to understand that one of the biggest contributions they or anyone is likely to be able to make is through data sharing (8).

Third, we need to continue research into privacy enhanced data sharing protocols (13) and to communicate better what is possible to government officials. Modern technology allows hundreds of millions of people to do electronic banking, commerce, and investing on the web; to view their personal medical records; to store their photographs, videos, and personal documents on line; and to share with selected individuals their most private thoughts and secrets. So why, when analyzing these and other personally identifiable sensitive data for the public good, does policy regularly require researchers (through university Institutional Review Boards) to do their work in locked rooms without access to the Internet, other data sources, electronic communication with other researchers, or many of their usual software and hardware tools? Surely we can develop policies, protocols, legal standards, and computer security so that privacy can be maintained while data sharing and analysis proceeds in far more convenient, efficient, and productive ways. Progress in social science research would be greatly accelerated if policies merely allowed researchers more often -- as it does corporations, governments, and private citizens -- to analyze sensitive data using appropriate digital rather than physical security.

Fourth, even when privacy is not an issue, data sharing involves more than putting the data on your web site. Scientists, and editors of scholarly journals, are not professional archivists, and many home grown one-off solutions do not last long. Data formats have been changing so fast that archiving standards require special preservation formatting, using internationally agreed upon metadata protocols and appropriate data citation standards. Social scientists need to continue to build a common, open source, collaborative infrastructure that makes data analysis and sharing easy (9,10). However, unless we are content to let data sharing work only within disciplinary silos, which of course makes little sense in an era where social science research is more interdisciplinary than ever, we need to develop solutions that operate, or at least interoperate, across scholarly fields.

Finally, social scientists could use additional help from the legal community (14). Standard intellectual property rules and data use agreements need to be developed so that every data set does not have its own essentially artisan legal work that merely increases transaction costs and reduces data sharing. The federal government should reconsider and relax the rules that prevent academic researchers from collecting, sharing, and publishing from data that those in other sectors of society do routinely.

Of course, social scientists have plenty to do even before we publish and share data. We must find ways of educating students about nonstandard data types, computational methods that scale, legal protocols, data sharing norms, and statistical tools that can take advantage of the new opportunities. Data are now arriving fast enough that the work life of many current social scientists is observably changing: whereas they once sat in their offices working on their own, rates of co-authorship are increasing fast, and a collaborative laboratory-type work model is emerging in many subfields. These trends would be greatly facilitated by universities and

funding agencies recognizing the need to build the infrastructure to support social science research.

For the first time in many areas of the social sciences, new forms and quantities of information may well make dramatic progress possible. Will we be ready?

REFERENCES AND NOTES

1. H. Weintraub et al., *Science*, **267** (1995)
2. D.E. Koshland, *Science*, **267** (1995).
3. See Chapter 38, in G. King, K. Scholzman, and N. Nie eds., *The Future of Political Science: 100 Perspectives*. Routledge (2009).
4. L. Sweeney. *J of Law, Medicine and Ethics* **25** (1997).
5. G. King. *Soc Meth and Res.* **36** (2007).
6. M. Altman and G. King. *D-Lib Mag.* **13** (2007).
7. G. King. *PS: Political Science and Politics.* **28** (1995).
8. G. King. *PS: Political Science and Politics.* **39** (2006).
9. The Dataverse Network, <http://TheData.org>.
10. DATA-Pass, <http://www.icpsr.umich.edu/icpsrweb/DATAPASS/>.
11. A.M. Blair. *Too Much to Know: Managing Scholarly Information before the Modern Age*, Yale University Press (2010).
12. R.G. Anderson, W.H. Green, B.D. McCullough, H.D. Vinod. *J of Economic Methodology*, **15** (2008).
13. C.C. Aggarwal and P.S. Yu, eds., *Privacy-Preserving Data Mining: Models and Algorithms* (2008).
14. V. Stodden. *International Journal of Communications Law and Policy*, **13** (2009).
15. C. Mackie and N. Bradburn, eds, *Improving Access to and Confidentiality of Research Data*. National Research Council (2000), p.49, http://www.nap.edu/catalog.php?record_id=9958.
16. R.F. White. *The Independent Review*, XI, n. 3, (2007), p.551, <http://www.independent.org/publications/tir/article.asp?a=630>.
17. D.Hopkins and G.King. *American Journal of Political Science.* **54** (2010).