Data Collection and Analysis: Balancing Individual Rights and Societal Benefits

Ramon Barquin
and
Clayton Northouse

Computer Ethics Institute

Paper presented at the
Panel on Access to Research Data: Assessing Risks and Opportunities
Workshop

October 16-17, 2003

The National Academies
2101 Constitution Avenue, NW
Washington, DC

TABLE OF CONTENTS

I. Framing the Issue

What if we lived in a world where schools went up in neighborhoods with no children or roads were built to places no one wanted to go? There would obviously be many complaints about who was running that world and how things might be done ***better***. This is a key concept, since to do better usually entails comparison, improvement, assessment of what is now, review of options, evaluation of approaches, in short, analysis of givens or observations, analysis of data. As scientists, we have learned that in order to proactively improve any activity or aspect of life we need to do analysis of data that has either been extracted from relevant transactions or gathered specifically, experimentally, for that purpose.

As a matter of fact, nothing has stirred more debate in the post-9/11 world than the question of whether the attacks on the World Trade Towers and the Pentagon could have been detected and prevented. Much of that discussion has centered on whether our intelligence and law enforcement communities had the adequate information technology (IT) tools to have identified the terrorist plans early on, located the planners, and taken the necessary preemptive action. The potential basis for doing this hinges on the existence of databases with enough data about potential terrorists so that researchers could identify plans, intentions, individuals, dates and places expeditiously. But there continues to be an ongoing debate over what would have been possible without there being considerable damage to individual liberties and the cost of that loss to society as a whole.

Today, technology has made data collection, data integration and its analysis, an extremely powerful instrument that we have used to good stead in order to improve our lot. Yet we have also learned that there are potential unwanted second-order consequences involved. Gathering substantial amounts of data for analysis may also result in potential harm being done to the rights of individuals. How can we do everything we need to be safe and secure as a society and still keep our individual liberties?

The purpose of this paper is precisely to discuss the issue of data collection and analysis and how we can maximize the good obtained from it while minimizing potential harm.

Focus on government data collections

Any person or organization can collect data and they, or others, may engage in its analysis. Many corporations do so in the context of their own market research, as do scholars in educational institutions for their own academic pursuits. In addition, there are private sector entities that are engaged in the collection, packaging and merchandising of data as their principal business. These are all interesting and important in the context of our topic, but we want to focus this essay specifically on government data collections, since they in many ways represent seminal examples of the need to balance societal benefits with individual rights.

Why is this? First, we must differentiate between situations where providing data is mandatory or voluntary in nature. In the latter case, assuming awareness, the individual must consent and hence has the ability to make his or her own decision. In the former, there is no such choice, and for all practical purposes only government can truly mandate and enforce, under penalty of law, that individuals provide the sensitive information it decides to collect. Private sector and non-governmental institutions in general request data that is voluntarily given and there is always the choice to opt out of a specific data collection campaign. Data obtained in the private sector through integration of multiple public-domain sources need to be looked at carefully and where necessary dealt with in the context of many of the guidelines we will discuss in the other sections of this paper. Hence, the principal focus of this paper will be government source data collections.

In effect, one of the most significant contributions to the nation's governance lies in the wealth of information collected by government agencies. The Bureau of the Census, the Center for Medicare and Medicaid Services, the Energy Information Administration, the Bureau of Labor Statistics, the Centers for Disease Control and many other government agencies collect and disseminate information that serves as the basis on which decisions are made regarding legislation and its implementation and application. Without access to government data there would be no solid basis on which to analyze the factors involved in the critical issues of our times, such as poverty, health care, education, traffic, public safety or the environment.

The societal good generated from the use of data increases with the degree to which this data is shared with researchers, other government agencies and departments, distributed to the public, and combined with other data sets. Yet, there are concerns over what and how data can be shared. The potential value of data to the public is in direct proportion to its dissemination for purposes of research and decision-support. The most serious constraint to dissemination is at the same time the basis of trust that enables the government to collect accurate and timely data: its commitment to guarding the privacy and confidentiality of personal information in order to protect the rights of the individual.

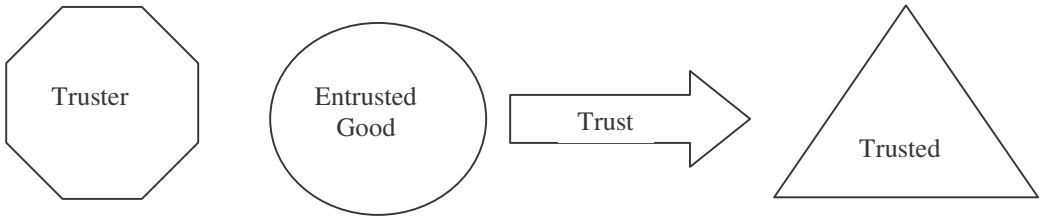
This paper focuses on the privacy and confidentiality concerns relative to data collection and dissemination that should be reviewed and analyzed in order to explore how the greatest societal good can be attained while minimizing the infringements on these rights of privacy and confidentiality. The paper will attempt to outline a framework for how government agencies can attempt to balance the privacy concerns of the individual with the societal good generated from the use of data.

II. The Role of Trust in Data Collection

In this section, we present a theory of trust, demonstrate how this theory is applied to data collection, and discuss how trust is dependent upon the preservation of privacy and confidentiality. Data can be collected directly through surveys or indirectly through extractions from transactions. The relationship that governs both of these forms of data collection is a relation between the data user and the public. For a member of the public

to willingly and truthfully give his or her personal information to the government, the individual must trust the government.

In the psychological and philosophical literature on trust, the concept is taken to be a three-part relation:[1] the truster (in this case, a member of the public), the trusted (the organization collecting personal information), and the entrusted good (the individual's personal information). When there is trust, the member of the public gives the data collector his or her personal information. This is sketched out as follows:



It is tempting to see "trust" as a Boolean element (an "or" gate) in the relationship. If there is trust the good will be passed on, if there is no trust it will not. But the concept of trust is inherently much more complex. Trust is broken down into three principles. The truster must see the trusted as, first, possessing a goodwill, second, encapsulating his or her interest,[2] and third, being competent in handling the entrusted good. The trustworthy person is then a person who has goodwill, encapsulates the interests of others, and is competent to handle the entrusted good. Trust is necessarily a matter of degree and can be represented by the following formula:

$$T_{\psi}(A:B) = f(G, I, C)$$

[1] See Annette Baier, "Trust and Antitrust." *Ethics*. Vol. 96, No. 2. (Jan., 1986), pp. 231-260; and Russell Hardin, "The Street-Level Epistemology of Trust." *Politics & Society*. Vol. 21, No. 4. (Dec., 1993), pp. 506-507.

[2] For the view of trust as encapsulated interest see: Russell Hardin. *Trust and Trustworthiness*. New York: Russell Sage Foundation, 2002.

Where,

$T_{\psi}(A:B)$ – Trustworthiness of person A (trusted), as perceived by person B (truster), in relation to good ψ.

G – Goodwill of person A

I – Degree to which A is able to represent the interests of B

C – Competence of A in handling entrusted good ψ

A number of questions arise when confronting trust in the context of data collection: What does it mean for the data collector to possess goodwill? And if this can be specified, how does the data collector affect the public perception of this goodwill? How does a data collector encapsulate a person's interests? Lastly, what does it mean for a data collector to be competent?

Before answering these questions, it is important to note that not only individuals but also institutions can be trustworthy, and both individuals and institutions can be trusted. This means that both individuals and institutions can possess and be seen to possess goodwill, interests that encapsulate the truster's interests, and competency. More specifically, "trusters" often assign higher values of G, I and C to individuals who are affiliated with certain institutions or professions. An unknown individual is often trusted if he or she is under the aegis of a trusted institution or is seen as a member of a certain trustworthy type. For instance, one trusts a dentist to operate on one's teeth even if he or she is relatively unknown, but has the appropriate credentials (e.g. the truster sees the

individual's diploma from dental school and thereby classifies the person as a member of the type "dentist" and transfers the trustworthiness of the school to the individual).

We must now clarify each of the three components of trust in the context of data collection. The possession of these three components constitutes the trustworthy person or institution and such a person or institution fundamentally comes down to being an ethical person or institution.

Goodwill

For a data collector to possess goodwill, G, is for one to have the goal of maximizing the public good in the collection, dissemination, and analysis of data while taking the necessary precautions of preserving the rights of the research participants and minimizing whatever harm may be caused to these individuals. In order to clarify the concept of goodwill, we specify a few attributes that those who have goodwill often possess:

- Part of the trustworthy person's motive in handling the entrusted good is to fulfill the truster's interests
- In fulfilling the truster's interests, the trustworthy person looks beyond the interests that would only benefit him or herself
- The trustworthy person has demonstrated that he or she cares about the management of other's entrusted goods in the past
- The actions, motives, and interests and the principles which guide the trustworthy person's actions are clear and easily understood

Therefore, in order to be seen as possessing goodwill, one is seen as having some or all of these attributes.

The attributes that apply to trustworthy people also apply to trustworthy institutions. And certainly when we are dealing with government as the data collector, then we must focus on trust in government and its specific institutions. For example, the Census Bureau has long been a model of how to collect large amounts of data that are extremely valuable to society. The data are used to inform public policy decisions of the utmost importance, while at the same time, the Bureau rigorously fights to maintain the confidentiality of the research participants. At the cost of not assisting other government investigations and not revealing personal information to researchers, members of the public, or other public sector agencies, the Bureau has upheld the principle that no individual should be identifiable in any publicly disseminated piece of information. The following is a set of attributes that apply to institutions that possess goodwill in this respect:

- The trustworthy institution's goal is to uphold the interests of trusters even though they may not have an interest in doing so and even if doing so conflicts with certain interests of the institution
- The institution has clear policies which guide the behavior of its members and demonstrate that the goal of the institution is to preserve the public's interests
- The institution's implicit or explicit code of conduct demonstrates that the institution upholds the interests of the public

It is through a long-term commitment of clearly abiding by the principle of confidentiality that the Bureau has demonstrated its goodwill in maximizing the public good while minimizing personal harm.

Institutions can also demonstrate that they are concerned about the ethical implications of using information technologies by subscribing to a voluntary code of ethics and administering programs that teach the ethical use of information technologies. One such code that has enjoyed widespread use, for example, is the 'Ten Commandments

of Computer Ethics'.[3] Using such a code and implementing courses that teach such codes promote the principles of the ethical use of computers. In a coming section, we specify the principles of privacy that should be upheld in data collection and analysis. For an institution to have a goodwill is for it to build these principles into its operational framework.

Encapsulating the truster's interests

The truster must see the trusted as encapsulating the truster's interests. This means that in fulfilling his or her own interests, the trusted fulfills the truster's interests. This aspect of trust means that the trustworthy person is seen as reliable, as one that can be depended upon. The following are attributes of the reliable person:

- The trustworthy person has demonstrated that he or she can be relied upon in the management of other people's entrusted goods
- The interests of the trustworthy person are clearly observable and understood
- How the trustworthy person's interests lead to the fulfillment of the truster's interests are clearly observable and understood

A data collector must encapsulate a research participant's interest. This means that the fulfillment of the data collector's interest must be seen as leading to the fulfillment of the research participant's interest. In other words, a person may be willing to give personal information to a government agency, because he or she sees it as serving a worthwhile function. The person may understand that his or her government can make

[3] <http://www.brook.edu/its/cei/overview/Ten_Commanments_of_Computer_Ethics.htm>

better and more informed decisions with this information, and the good of participating outweighs the potential costs of privacy violations.

For an institution to encapsulate a person's interest means that the fulfillment of the institution's interests should also lead to the fulfillment of the person's interests. The following are attributes of the reliable institution:

- The trustworthy institution has a history of reliability in the management of public's entrusted goods
- The interests of the trustworthy institution are clear, open and understood
- How the trustworthy institution's interests lead to the fulfillment of the truster's interests are also clear, open and understood

Competence

The third component of trust is that the data collector must demonstrate that they are competent enough to manage the truster's entrusted good, in our case, personal information. This is where the practical concerns of maintaining a secure data infrastructure and cracking down on data collectors who break with the policy of preserving confidentiality come into play. Even if the data collector is seen as having a goodwill and as encapsulating one's interests, an individual will be unwilling to trust if despite these good intentions, the data collector cannot competently handle one's personal sensitive information.

To be competent, Webster's Dictionary tells us, is to have "requisite or adequate ability or qualities", but also to be "legally qualified". We can understand that to be competent entails having the ability or quality to do something. A competent teacher

knows how to transmit knowledge so that his students are able to learn; a competent pilot can operate her airplane so that it takes off and lands smoothly and flies safely; a competent chef prepares food in a way that that it is tasty, attractive and timely. In all of these cases there are skills needed and tools involved.

What are the skills a data collector must have and what tools must he or she know how to use in order to be considered competent? If we think of the individual as a data collector, we would expect that they be well-organized, attentive to detail, mindful of the data's value, etc. In terms of the contemporary data collector, they should be able to handle the tools of the trade: survey forms, experiment design, relevant hardware and software, statistical techniques, etc. And if the data were sensitive, of course, we would expect competency in data security and safety. This means both keeping it physically secure as well as logically. In other words, ensuring that only individuals authorized to do so can see the data. And today this is done primarily through security technology such as firewalls, passwords and encryption.

When the data collected is both sensitive and mandatory, as in some of the examples we provided for government collections, then the stakes are so high that we often revert to the second definition from Webster and require "legal qualification", in effect testing the individual to assure his or her competence.

Institutions, such as government, must address the issue of competence with a combination of technology and policy. Tools and techniques will only accomplish so

much to provide competence in handling sensitive data; there must also be appropriate policies and procedures related to the handling, storing, processing and safekeeping of the data to complement all the tools in the toolkit and make sure they do what they're supposed to do.

Additional caveats

As mentioned, the model of trust when applied to data collection can be used to describe the relation between either governmental or non-governmental data collectors and the public. In this essay, we focus on governmental data collectors, but non-governmental data collectors still influence the government-public relationship in so far as the public perceives some common aspects of data collection and is inclined to merge them in their thinking. For example, spammers extracting email addresses from chat rooms or legitimate subscription lists will make the public wary of providing their email address for other legitimate collectors, such as an e-government application. In other words, negative publicity related to non-governmental data collectors may affect the level of trust the public gives to governmental data collectors even though the government may abide by different standards and different policies. Additionally, the level of trust given to the government in general or parts of the government that the public may associate with collection and use of personal information will affect the level of trust given to governmental data collectors.[4] Hence, though we are limiting the focus of analysis, these

[4] For an analysis of the level of trust in the government related to information technologies see Hart-Teeter for the Council for Excellence in Government, "The New E-Government Equation: Ease, Engagement and Protection," April 2003. For the level of trust given to parts of the government in general see The Harris Poll, December 16, 2002. In the latter, the military is rated the highest with 62% of the public trusting the military a great deal and Congress is rated near the bottom with 22% trusting Congress a great deal.

aspects need to be taken into consideration when the goal is trying to promote the public's trust in the government's data collection practices.

Again, we want to emphasize the importance of taking into consideration the distinction between mandatory and voluntary submission of data. Trust plays the largest perceptively causal role in the voluntary submission of data, since those who do not trust will not submit whereas those who do will submit sensitive information. The level of trust can then be analyzed as playing a significant factor in the level of returns of voluntary data collection surveys. In mandatory data collection surveys, the influence of trust will be muted due to the threat of criminal sanction. Though one may not trust the institution collecting the information, one will submit sensitive information in order to avoid a criminal sanction. Trust, though, still serves an important function in mandatory data submissions. If there is sufficient trust, there will be less of an incentive to provide false or deceptive data in order to cover sensitive personal information. Furthermore, if the institution collecting personal information is not considered trustworthy then it will lose its credibility over time and public pressure may force it to scale back its data collection activities. Hence, though the impact of trust is less perceptible in the case of mandatory data submissions, public trust affects the mandatory and voluntary data collection institutions' ability to function effectively.

In this section, we have demonstrated the importance of trust. If data collectors do not uphold the three principles of trust, then the public will fail to trust, and if the public fails to trust, then the data collector will not receive the personal information of

individuals. In all three of the principles, the preservation of privacy and confidentiality play a central role. Hence, in order for a data collector to preserve the trust of the public, one must be aware of the principles of privacy and confidentiality, and the data collector must understand how these can be preserved while maintaining the usefulness of the data.

As to the framing statement about trust in government, we just need to remember President Madison's words in Federalist #51 concerning the need for a separation of powers:

> "It may be a reflection on human nature, that such devices should be necessary to control the abuses of government. But what is government itself, but the greatest of all reflections on human nature? If men were angels, no government would be necessary. If angels were to govern men, neither external nor internal controls on government would be necessary. In framing a government which is to be administered by men over men, the great difficulty lies in this: you must first enable the government to control the governed; and in the next place oblige it to control itself."

In the next section, we outline the principles of privacy and confidentiality that constrain data collection and dissemination activities of statistical data centers.

III. Informational Privacy: Basic Issues

Information privacy is a concept developed to apply to the collection, use and maintenance of personal information with the advent of database technologies. Privacy is a broad term with many applications. For instance, there is surveillance privacy that concerns the right to have a private life free from external monitoring or unwanted observation. The search and seizure clause of the Fourth Amendment protects this right and prevents the government from intruding into one's private domicile. This form of

privacy has traditionally been defined by space: privacy concerns things in the personal as opposed to the public realm.

The development of technology has severely strained this model. For instance, in the use of wiretaps the government can listen to personal conversations without intruding into one's private space. While this can be highly beneficial to society in the apprehension of wrongdoers, it nonetheless irks anyone concerned with civil liberties. In addition, with the collection of personal information from credit card transactions, medical records, and travel reservations, the amount of data that can be stored about one person can be used to determine what a person eats, drinks, where they live and where they travel, how many children they have, how many crimes they have committed and what illness or disease they may have. With the integration of these disparate sources of information, which are traditionally thought to be public transactions because they occur in the public realm, a very powerful and revealing profile of a single individual can be constructed. The development of technology called for a different conception of privacy that was not based on this public/private distinction and hence we come to information privacy.

Alan Westin developed a compelling definition of information privacy. Information privacy refers to the standards for the collection, maintenance, use and disclosure of personal information. A central component of this is the power of an individual to control the use of sensitive information.[5] Now, we must investigate what constitutes sensitive information. When does information regarding a particular

[5] Alan Westin. *Privacy and Freedom*. New York: Atheneum, 1967.

individual become sensitive? Are there certain types of information that are sensitive and others that are not and if so what constitutes the sensitive type? Is sensitive information defined by the power it gives someone to harm another?

Sensitive Information and Identifier Information

Sensitive information is information that one does not want to be publicly known. The reason that people do not want sensitive information to be publicly known is usually because it would result in some form of emotional, physical, or financial harm. A person gives information to his or her doctor but does not want the doctor to broadcast that information, or a diagnostic, on national television that night. National security information is classified as sensitive because if it were to be publicly known it could result in harm caused to the nation. Keeping sensitive information private entails controlling who has access to the information. Medical information, for example, is kept private by not allowing anyone without the need to know (i.e., the patient's physician, other health care workers, pharmacist, medical insurance workers) to have access to this information. Restricting access to sensitive information preserves individual privacy.

The question of identity also figures into the equation of being able to harm someone with sensitive information. In order to harm someone with sensitive information, one must have access to identifier data. Hence to provide protection against the disclosure of personal information, one must know what it takes to identify a particular individual. With the multiplication of identifier information and the potential for fraud or corruption of breeder documents (e.g. Social Security numbers), this is a

difficult task. The question of how to preserve identifier information is an enormous problem, let alone the additional problem of how to keep this identifier information in the right hands. This also raises the valid question of whether identity itself is sensitive. Insofar as we are now reaching epidemic proportions in the cases of reported identity thefts, the answer seems to be yes. The amount of harm, financial and otherwise, that can be done to someone through the usurpation and exploitation of his or her identity is huge. For the purposes of this paper, though, we will leave these issues aside and assume that we have at least a working definition of identifier data and a way to limit access to it.

Establishing a Focus

Tying this back to the concept of trust developed in the previous section, the entrusted good in the relation between data collectors and members of the public may or may not be sensitive information. The degree to which the individual regards the information as sensitive, the higher level of trust needed to entrust the information to a data collector. Trust does not need to be present when giving a data collector information that is not at all sensitive. In the long run, this can become a complicated issue because the individual may be unaware of the sensitivity of the information until it is too late. Seemingly innocuous bits of data can become sensitive when integrated with other data that allow piecing together the informational "puzzle." So, what may have previously been conceived as non-sensitive information may come to be seen as sensitive when situations arise that make significant numbers of individuals wary. Such a situation is the current epidemic proportion of identity thefts.

The issue of identity must also be tied back to the concept of trust. Sensitive data, which remains detached from identifier information, does not pose as great a threat to cause harm as sensitive information which will be immediately matched with identifier information. The public can operate with a lower level of trust when interacting with institutions which preserve the anonymity of the collected data than when interacting with institutions which match the collected data with identifier information. Data collection institutions can collect highly sensitive data without the need to secure trust as long as the data is not matched to particular individuals. In our analysis of trust, we are concerned with the data collectors that have the power to match sensitive information with identifier information. Yet, we also have to remind ourselves that the achieving the greater good for the greater number may sometimes require identifying a physical individual, such as when we need to find and quarantine the carrier of a highly infectious virus.

We focus on how to preserve and establish trust when dealing with highly sensitive information that can be easily identified. In the remainder of this section, we outline the principles of privacy that should be upheld in data collection and dissemination in order to preserve the public's trust and minimize the potential for harm that can be caused to individuals.

Principles of Privacy

We have spoken somewhat loosely about privacy and confidentiality. They are clearly related but inasmuch as privacy is defined as "freedom from intrusion or public

attention," let us distinguish it from confidentiality which entails certain agreement not to disclose "entrusted secrets." Hence, we can say that when a good (i.e., private sensitive information) is entrusted by the truster to the trusted, the latter is expected to keep it confidential.[6]

In the government's use of data, preserving privacy comes in at two places. First, privacy figures in the data collection point. A principle of privacy that must be upheld is that the person submitting the information knows who is collecting the information, knows the purposes for which the information will be used, and knows not only if the information will be shared but to whom the information will be shared and for what purposes third parties might use that information. Second, privacy must be upheld in the use and dissemination of the data. The person's information must not be shared without the consent of the individual and the information must be stored in a secure environment that prevents unauthorized intrusion. The Center for Democracy and Technology has taken a number of policies regarding the protection of the right to information privacy and enumerated a set of seven principles:[7]

1. Openness: The existence of record-keeping systems and databanks that contain personal data must be publicly known, along with a description of the main purpose and uses of the data.

[6] See *The Problem of Definition -- Privacy and Confidentiality*, From the U.S. Congress, Office of Technology Assessment, Protecting Privacy in Computerized Medical Information, OTA-TCT-576 (Washington, DC: U.S. Government Printing Office, September 1993). Here the following points are made: *Confidentiality* will refer to how data collected for approved purposes will be maintained and used by the organization that collected it, what further uses will be made of it, and when individuals will be required to consent to such uses. *Privacy* will refer to the balance struck by society between an individual's right to keep information confidential and the societal benefit derived from sharing the information, and how that balance is codified into legislation giving individuals the means to control information about themselves.

[7] http://www.cdt.org/privacy/guide/basic/generic.html

2. Individual Participation: Individuals should have a right to view all information that is collected about them; they must also be able to correct or remove data that is not timely, accurate relevant, or complete.

3. Collection Limitation: There should exist limits to the collection of personal data; data should be collected by lawful and fair means and should be collected, where appropriate, with the knowledge or consent of the subject.

4. Data Quality: Personal data should be relevant to the purposes for which it is collected and used; personal data should be accurate, complete, and timely.

5. Finality: There should be limits to the use and disclosure of personal data: data should be used only for purposes specified at the time of collection; data should not be otherwise disclosed without the consent of the data subject or other legal authority.

6. Security: Personal data should be protected by reasonable security safeguards against such risks as loss, unauthorized access, destruction, use, modification or disclosure

7. Accountability: Record keepers should be accountable for complying with fair information practices.

The principles are developed in order to reduce the amount of harm that can be caused to individuals. Abiding by these principles and incorporating them into the fabric of a data collection institution increases the trustworthiness of the institution. When effectively promulgated, members of the public will thereby be willing to trust because they will see such institutions as embodying the attributes of trustworthiness. Hence, the establishment of robust policies that protect the privacy of individuals promotes the trustworthiness of institutions.

IV. Balancing Privacy and the Public Good: The Case of the Census Bureau

In this section, we outline how a relationship of trust was established between the public and government in the collection of Census data. The original purposes for conducting the Census, as mandated by the Constitution, was to reapportion seats in the

House of Representatives, assess the military capabilities of the country, and determine the taxes for each state.[8] As the potential benefits of collecting nation-wide information were demonstrated, the Census expanded the number of questions and applied the data to other purposes. Only then, did privacy and confidentiality concerns become an issue, and it was not until the mid-twentieth century that robust confidentiality protections were placed on the operations of the Bureau of the Census.

In the first censuses, the compiled data with identifier information was posted in public places so that citizens could inspect and verify the data.[9] During the early American period, there was little concern regarding privacy and confidentiality of Census data. People had little concern regarding any adverse purposes such data may be used for, and it was not until 1850 that a policy of confidentiality was imposed: Census employees could not disclose any information to non-governmental employees. The model of trust was employed in explaining the effectiveness of the policy. The Secretary of the Treasury, Thomas McKennan stated that

> "All marshals and assistants are expected to consider the facts intrusted to them as if obtained exclusively for the use of the Government, and not to be used in any way to the gratification of curiosity, the exposure of any man's business or pursuits, or for the private emolument of the marshals or assistants, who, while employed in this service, act as the agents of the Government in the most confidential capacity".[10]

Even at this early stage, the relationship between the public and government is seen as one of trust. The public gives the government its personal information for certain

[8] See "Census Confidentiality and Privacy: 1790-2002," Bureau of the Census, 1.

[9] Ibid., 4.

[10] Ibid., 6.

uses, and if the government uses the data for other unspecified purposes, the government has violated the public's trust.

President Howard Taft made the first public proclamation regarding the Census and stated that,

> "There need be no fear that any disclosure will be made regarding any individual person and his affairs. For the due protection of the rights and interests of the persons furnishing information, every employee of the Census Bureau is prohibited, under heavy penalty, from disclosing any information which may thus come to his knowledge."[11]

Here, again, the importance of trust is demonstrated. The public must be aware that it is in the interest of the government not to divulge private information, and knowing this, the public can enter into a relation of trust with the government. The Director of the Census Bureau saw himself as the "custodian or guardian, who is to see that [Census data] is used for the purposes for which it was gathered and not for private purposes to the harm or detriment of the person or persons from whom it was obtained under the implied promises that it would be considered confidential."[12] In 1930, policy was established limiting public access to data regarding oneself or that of one's minor child. In the implementation of this policy the government demonstrated that it was concerned about the confidentiality of Census data and that the public could trust the government's collection and management of their personal data. We believe that these strict privacy principles are conducive to generating public trust in the use of Census data.

[11] Ibid., 9.

[12] Ibid., 13.

The most important source of protection against privacy or confidentiality violations regards Title 13 of the United States Code. Title 13 assures the confidentiality of all records in the Bureau's custody. The goal of the Census is to "release as much statistically valid and useful data as possible without violating the confidentiality of the data as required by Title 13."[13] In addition, the Census has established a very strong "culture of confidentiality." Before any decision to publicly release data is made, the Bureau's Disclosure Review Board must approve the decision, and in making such approval, the released data must meet the standard that no data can be matched to any individual. The Bureau has successfully implemented the policy that before any decision is made regarding Census data, the privacy impact on the public must first be considered.

In part, the Census policies react to the successes and failures of the Bureau. Great harm can result when unbounded discretion to disseminate sensitive data is given to collecting agencies. This was vividly evident in the role that the Bureau of the Census played in the internment of Japanese American citizens during World War II. Vincent Barabba, a former director of the Census, noted that "the 1940 Census was the single most important source of information used for evacuation and resettlement purposes [of Japanese Americans]".[14] So, while data should be disseminated to create societal good, there need to be definite guidelines for the sharing of such information. The success story that is told by the Census is the time when the Federal Bureau of Investigation appeared with a search warrant wanting to confiscate the Census information in Denver, Colorado in order to determine whether a particular individual had forged survey data. The Census

[13] Ibid., 21.

[14] Ibid., 15.

Bureau immediately recognized the problem of data dissemination to other branches of the government. The Census Bureau stopped the FBI, and instead, researched the matter themselves. They ended up writing a report discussing the possible criminal actions without releasing the identity of the individuals and thereby preserved the confidentiality of the information. The promotion of this story within the Bureau tells the public that they have the goodwill to go to great lengths to keep the public's personal information confidential.

V. Benefits of Data Access

Though principles of information privacy should be vigorously upheld, there is a great deal of public good generated from the dissemination of data. In this section, we discuss the societal benefits that are created from sharing data with other branches of the government and with non-governmental researchers.

Societal Benefits

There are substantial benefits to society that come from data access for purposes of analysis. As we hinted in the beginning of this essay, how can we decide on the appropriate allocation of public goods or resources if we have no way of determining where they are needed? We could wind up building primary schools in areas where there are no school-age children, or roads that don't get people where they need to go. If we have no hard data on neighborhood demographics, or where households and workplaces are located, how would we know any different? How can we be sure that we don't give entry visas to terrorists without having substantial access to the relevant databases for

intelligence gathering and analysis? There are many obvious benefits to society, specifically tied to improvement or making things better, that derive from having adequate data to analyze.

Some of the data we need for analytical purposes may not be of a sensitive nature. For example, assuming a reasonable level of trust, most individuals should not have a problem with providing vital statistics, educational history, etc., if the outcome is going to benefit them or their communities. Furthermore, if we can find ways to decouple data from an individual's identity it becomes more likely still that there will be little objection to sharing the data. However, when it comes to sensitive data, as we described in the earlier section, it is usually another story.

Yet, how can we appropriately credential physicians in our hospitals without accessing the Department of Health and Human Service's National Practitioner Data Bank (NPDB) to check that individual doctor's record for cases of medical malpractice? Or how can we avoid the spread of certain epidemics without knowing the identity of the specific infected carrier? Here, the need for specificity is very substantial, as is the requirement to be able to link all the way back to a physical person.

Research Benefits

A study by the Committee on National Statistics, National Research Council[15] outlined a number of benefits to sharing certain research data in general (not specifically governmental data), which are relevant to our current focus: reinforcement of open scientific inquiry; verification, refutation, or refinement of original results; promotion of new research through existing data; improvements of measurements and data collection methods and analytic technique; and climate in which scientific research confronts decision making. In the same study, Terry Hedrick[16] lays out additional benefits of data dissemination: replications with multiple data sets; exploration of new questions; creation of new data sets through data file linkages; reduction in respondent burden.

Before discussing the main benefits of data access, we must address the issue of data quality. Researchers have long been aware of the GIGO principle, garbage-in, garbage-out. GIGO seems to be well ensconced in many databases integrated from other sources as the defects of data collected and efficiently stored in databases are transmitted and amplified through extractions and transformations. The problem often lies with the original sources of the data, and points back to the lack of a rigorous and methodical approach in many government agencies to ensuring information quality. In any case, while some may argue that bad data is a plus toward making it less desirable and hence less likely to lead to the disclosure of private or confidential data, it is nonetheless a real problem for anyone attempting to conduct research with that data. Hence, the capacity to

[15] See Stephen E. Fienberg, Margaret E. Martin, and Miron L. Straf, Editors; Committee on National Statistics, National Research Council. *Sharing Research Data*. Washington, DC: National Academy Press, 1985. The benefits discussed in this paper are selected from pages 9-15.

[16] See Terry E. Hedrick, "Justifications and Obstacles," in Stephen E. Fienberg, Margaret E. Martin, and Miron L. Straf, Editors; Committee on National Statistics, National Research Council. *Sharing Research Data*. Washington, DC: National Academy Press, 1985, pp. 124-132.

access and use data with dubious quality detracts substantially form the benefit that may accrue from that research. There are several proven approaches to achieving information quality that should be examined by serious data collectors and researchers.[17]

Now let us briefly discuss the five main benefits of data dissemination. The dissemination of governmental statistical data to the research community fosters an open research community. Policy disputes related to interpretation are common, and with the wide dissemination of data to researchers, these disputes are much more informative. In the end, there is a higher probability that the good or right policy will win, because the good or right policy will be the policies that are best supported by statistical data. Second, the findings that come from the analysis of governmental statistical data will undergo reexamination and reinvigoration when disseminated to the research community. This may expose errors that may have not been noticed. The policies that are then created based on these findings will then be more accurate.

Third, the same data sets that are used for one purpose can then be put to use for another purpose without substantial investments to perform the collection of data. "The same data that were gathered by researchers to answer one set of questions can be used by others to answer a new set."[18] Additionally, data sets can be combined and result in much more powerful tools for examining the problems facing society. This will give policy makers more definitive answers to the problems they currently face. This also will

[17] See Larry P. English, *Information Quality Improvement: Processes and Best Practices for Business Performance Excellence*, 2002 Ed., Brentwood, TN: INFORMATION IMPACT International, Inc., 1992-2002, pp1.2.
[18] *Ibid.* p. 10.

result in more studies with less investment in data collection and will in turn reduce the burden of respondents. Fourth, when research techniques are shared along with the data, the research community and other governmental statistical data centers improve and hone their own techniques. Techniques that may not have otherwise been acknowledged as faulty will be abandoned and techniques, which are effective, will be promoted. Fifth, when data is shared with the research community, it will actively involve researchers in the problems that are confronting the nation and policy makers. The quality of research would be advanced and the researchers would be actively involved in the decision making process.

Of course, these benefits must be achieved while upholding the privacy and confidentiality of the public, and in the next section, we discuss a few methods which are either being implemented or need to be implemented in order to preserve confidentiality while maximizing public good.

Degrees of data access and granularity of the data

As always, the devil is in the details; hence, the question of what degree of specificity yields the most benefit to research? It is easy to answer, that the more granular the data the better, and researchers will always want more access and a fine grain. Yet we know that there are limits. We are indebted to Jorge Luis Borges[19] for an insight into this topic when delves into the "exactitude of science" with his tale of the mythical empire whose emperor loved maps. "In that Empire," he explains, "the craft of Cartography

[19] See Jorge Luis Borges, "Of exactitude in science" in *A Universal History of Infamy,* E P Dutton, (November 1972), 146pp.

attained such Perfection that the map of a Single Province covered the space of an entire City, and the Map of the Empire itself an entire Province. In the course of Time, these Extensive maps were found somehow wanting, and so the College of Cartographers evolved a Map of the Empire that was of the same Scale as the Empire and that coincided with it point for point…"

So there are limits to granularity in cartography and there should be limits to granularity or atomicity in any research pursuit. Some of these are set by experiment design theory, others by elementary arithmetic, and others should be set by ethical, or practical, considerations revolving on our model of trust.

By the same token, degrees of access by researchers or examiners must also be addressed. Need to know becomes a key consideration. *Who* needs to know *what* in order to do *something truly critical* for society? This question has to provide the mandatory framework for deciding what degree of access can be given to individual researchers or examiners needing to work with collections of sensitive data.

VI. The Role of technology

Technology has an important role to play both in providing researchers with powerful tools for storing, processing, interpreting and displaying the data, as well as to protect the data itself, keeping it safe and secure. It has been the advances in information technology -- hardware and software -- that have allowed the collection and organization of large amounts of data into easily accessible and queriable databases. Advances in IT

hardware have made computing power a commodity and introduced very inexpensive storage devices allowing the collection of massive amounts of data. Parallel processors permit much speedier access and real-time manipulation of large databases kept permanently on-line.

On the software side we have seen the advent of data warehouses and database schemas specifically designed for OLAP (On-Line Analytical Processing) rather than traditional transaction processing. New and sophisticated indexing techniques have made databases much more malleable for analysis. Query tools have emerged that empower end-users to slice and dice data with ease in the process of problem solving and hypothesis testing; and data visualization software now permits researchers to display results of queries and analysis in ways that substantially enhance their ability to understand and interpret outcomes and consequences.

Lastly, data mining software tools have been developed that allow researchers to obtain substantial insights into data collections by looking for patterns and interrelations in a large number of variables. This is done through a number of approaches and techniques, such as applying advanced statistics, detailed decision trees, genetic algorithms and neural nets.

It has also been this wealth of technology that has increased the potential for privacy and confidentiality violations. Technology has made possible collecting and integrating these massive databases, many with information about individuals and their

identities, and with the tools to access and disclose sensitive data. Furthermore, the advent of the Internet increasingly provides a highway to the doorsteps of practically any database in the world.

Technology itself, however, has come partially to the rescue by providing us with tools to protect sensitive data. From the early days of data processing there has been an awareness of the need for security. Locks were installed to protect IT devices in the 1960's, and shortly thereafter security artifacts such as passwords emerged. With the emergence of networks, we saw the concept of the firewall arise and its Internet kin, the intranet. Most importantly, we have seen encryption emerge as a powerful approach to data protection. Furthermore, through the use of Public Key Infrastructure (PKI) there has been a strong move toward its adoption by many Federal agencies as a standard to protect e-government transactions where authentication and digital signatures are essential. [20]

Now, the move to network-centric architectures and knowledge-based environments is increasing the level of system complexity and hence the burden of system security. This is giving rise to identity management schemes that rationalize and facilitate system decision-making for managing access to networks and databases. Some of the solution categories for these technologies include single sign-on products, virtual directories, metadirectories, password management systems and provisioning systems. [21]

[20] The National Institute of Standards and Technology (NIST) has taken a leadership role on PKI within the Federal government. See http://csrc.nist.gov/pki/.
[21] See Tulu Tanrikorur, "Who Are You?" in ***Intelligent Enterprise***, September 1, 2003, pp33-36.

And there is substantial promise in work being done in privacy protection technology. Bayardo and Srikant recently reported on some of the areas being explored: Hippocratic databases, privacy-preserving data mining, anonymization, information sharing across private repositories, cryptographic protocols, secure coprocessors and privacy-preserving search. [22]

VII. The Role of Ethics

Ethics also plays a fundamental role in balancing individual rights and the public good. The role of ethics goes back to our discussions in section two of establishing goodwill in the model of trust. Those institutions that promote ethical practices within the institution promote a conception of goodwill amongst the public and thereby attain the trust needed to receive the public's personal information. Ethical practices can be established in one of three ways: from the top down through leadership, through voluntary codes of conduct, and through programs and classes.

The managers of data collection institutions need to be fully aware of the ethical practices of data collectors and know how to promote an understanding of these practices. The most effective means to this promotion is done through mission statements, speeches and corporate narratives. Ultimately, it needs to be reaffirmed through its incorporation where possible into performance plans. The Census Bureau established itself as a trustworthy institution and promoted ethical practices by strong declarative speeches on the subject, and by a series of explicit codes and statement of principles.

[22] See Roberto Bayardo and Ramakrishnan Srikant, "Technological Solutions for Protecting Privacy," in *Intelligent Enterprise*, September 1, 2003, pp33-36.

A code of conduct is implemented in order to instill an understanding of ethics and create consensus among the employees and stakeholders of an institution regarding the fundamental ethical principles involved in the practices of the group. Data collection institutions should therefore implement a code that incorporates the principles of informational privacy specified in section three. While a good code of conduct regarding the use of information technologies is helpful and to the point, any code that addresses the broader practices of data collection and sharing should also be embraced.[23]

Finally, internal programmatic activities can be implemented to establish ethical practices. These programs can be given to explain the code of conduct to the members of an institution through hands on activities. Additionally, walking through hypothetical ethical cases helps to demonstrate the practical importance of ethics in day-to-day activities of individual members of a data collection institution.

VIII. Finding a Balance

In finding a balance between public usefulness and confidentiality, we urge for greater dissemination while maintaining the same level of confidentiality. In this section, we offer a few ideas on how this can be done.

First, we believe that establishing contractual relations with the non-governmental researchers offers a wealth of opportunity without causing undue risks to privacy and

[23] For a wide array of codes of conduct see Appendix B, *Macro-Engineering: Global Infrastructure Solutions*. (Frank P. Davidson and C. Lawrence Meador, editors. New York: Ellis Horwood, 1992.)

confidentiality.[24] The process for applying and receiving unfettered access to limited sets of governmental statistical data should force the researcher to fully justify his or her project and should demonstrate why it is necessary for the researcher to have access to all the data as opposed to a restricted set of the data which has identifier information blurred or stripped.

The contract should also rigorously uphold the principles of informational privacy in section three. The following is an outline of principles that should be upheld:

1. Security: The data, when in the possession of non-governmental researchers, must be stored in secure databases that are up to the same standards as those required of government databases.
2. Accountability: The non-governmental researchers must be held accountable for any transgressions of privacy and confidentiality.
3. Consent: The research participants should be fully aware that their personal information might be shared with non-governmental researchers.
4. Finality: The data when shared with non-governmental researchers should be shared for finite periods of time, after which all non-public data must be destroyed.

Second, we believe that the Bureau of the Census's efforts to establish research data centers across the United States offer a fruitful opportunity to share Census data and provide a good model for the sharing of other types of governmental statistical data. If a researcher wants full access to Census data, he or she can make a proposal to the Census Bureau's Center for Economic Studies. If the proposal is approved, the researcher is sworn in to receive the status of a Census employee and is legally obligated to uphold the

[24] For the presentation of this view see G.T. Duncan and R.W. Pearson. "Enhancing Access to Microdata While Protecting Confidentiality: Prospects for the Future," *Statistical Science*. Aug. 1991, pp. 219-232.

confidentiality protections of the data. Instead of having to travel to Washington, DC, the Census Bureau has established seven research data centers across the United States. This program should be expanded to give access to Census data for more researchers across the country, and this model should be used by other governmental agencies with useful statistical data.

Beyond this we see a number of attempts to develop responses and frameworks to deal with the need for balance. In a recent Brookings Institution session related to balancing civil liberties and national security, the following framework was proposed by one of the discussion leaders in relation to the post-9/11 environment and the need for increased information sharing:[25]

(1) What are the new threats and what information is relevant to these threats?

(2) How can we get this information in the most effective manner?

(3) What institutions should take the praise or blame in gathering information regarding these threats?

(4) Who should provide oversight?

(5) How do we harness technology to meet these threats?

This is but one such attempt to develop a contractual relationship to govern these activities.

[25] This was a round table discussion on "The Challenge of Information Sharing: Balancing National Security and Civil Liberties" held at the Brookings Institution on June 5, 2003. The event was co-sponsored by the Brookings Institution ITS, the Computer Ethics Institute, and the Ascential Software Corporation, and was by invitation only and comments not for attribution.

Ultimately, we believe that the balance will have to lie in the skillful application of three sets of interrelated programs from the areas that we have mentioned here as having important roles to play: technology, policy and ethics.

First, we must continue to search for and develop increasingly robust tools to provide maximum protection for government-mandated and sourced collections of sensitive data. Whenever possible the research we do should utilize databases that are stripped of identity and guarantee anonymity. As much of this as we can do with technology, we should.

Second, government agencies that engage in the mandatory collection of sensitive data must establish clear and solid principles accessible to the public that specify their commitment to preserving confidentiality, implementing informed consent and limiting disclosure. These agencies must also invest in assuring they can handle and safeguard the data collections with competence. In addition, there must be penalties for violating any of the processes and procedures that underpin these principles and mechanisms for its enforcement.

Last, there has to be a commitment to ethics at the leadership level. This means that the mission is understood by all, and that voluntary codes are promoted and established. The data collecting institutions must strive to attain and preserve the trust of the public, and this is accomplished by acquiring the characteristics of possessing a goodwill and encapsulating the public's interests as outlined in section two. Through the

strict enforcement of the privacy principles specified in section three, data collection institutions can demonstrate that they are committed to upholding the rights of individuals while at the same time using the public's information for the public good.

In managing the balance between promoting public good and protecting individual rights, data collection institutions must effectively manage the three components of this balance—they must supply the technology, provide the correct policy, and cultivate an ethical environment of goodwill and trust. With this three-pronged effort, the institutions can hope to limit the number of negative unintended consequences that result from data collection and analysis, and thus strive to accomplish Raymond Bauer's dictum, when warned us that, "Briefly stated, the major task in control over our destiny is to make as many second-order consequences as possible intended, anticipated and desirable; and reduce to a practical minimum those that are unintended, unanticipated and undesirable."[26]

[26] Raymond Bauer. *Second-Order Consequences*, MIT Press, Cambridge, MA, 1969, p18.