Data.gov and De-Identification Considerations for the Open Government Directive

CENTER FOR DEMOCRACY & TECHNOLOGY
Keeping the Internet Open, Innovative and Free

1634 I Street, NW Suite 1100
Washington, DC 20006
202.637.9800
fax 202.637.0968
http://www.cdt.org

One of the first projects adopted under the Obama administration's push for transparency is the announcement of the Data.gov Web site. The intended purpose of Data.gov is to gather, bundle, and make publicly available various types of raw data produced by the federal government. By bringing together feeds of all non-private, non-classified data, this site will make it easy for individuals and the private sector to use this information, the same way that GPS navigation services and weather reports use free government data now. CDT, along with many other organizations, is excited about the promised deluge of data, as it is likely to spur new and innovative services and to provide third-party oversight of federal government spending. While the Web site has the potential to be a valuable resource for information, there are important privacy implications associated with data disclosure that need to be addressed -- namely, protecting the privacy of individuals.

Data.gov is expected to bring together a great deal of data that has not been previously released in raw form and thus has been effectively opaque to the public. Some of these data sets will be updated routinely in order to release the most current information. While the new presumption that data will be proactively released is laudable, those readying data sets for data.gov will need to ensure, prior to release, that their data does not contain personally identifiable information, sensitive information, or other information that could be used to link the released data to individuals. The Data.gov team must move forward cautiously in handling data sets containing such information in order to adequately address the corresponding privacy risks.

Privacy Implications of Data.gov

The government collects and produces information across sectors as diverse as scientific research and internal government functioning. Information is collected about economic indicators, health, product recalls, and government services such as Medicaid. Many databases are already made available in processed formats, but not in raw form.

Different data sets will have different qualitative privacy implications. Data about internal government functioning will tend to contain information about government employees, while other kinds of data will likely include information about private citizens and businesses. Each of these data types could contain personal information explicitly, or could be used to infer identity. For this reason, each data set will need its own specialized review before it can be published to data.gov.

This holds true for data sensitivity as well – certain kinds of data that have historically warranted higher privacy protections will require special care before they may be release through data.gov. While there is no firm consensus about what kinds of information should be considered "sensitive" in bulk, an array of existing statutes, self-regulatory guidelines

and policy proposals provide some basis for deciding what kinds of information about individuals should be granted some measure of special treatment. CDT has compiled a list of such proposals,[1] which may be helpful in determining the privacy implications of the release of particular data sets.

De-identification

Data should be released if, on balance, it can be put to good use. Yet in the absence of appropriate privacy and security safeguards, increased data liquidity can compromise individual privacy. One way to facilitate the use of data while protecting privacy is through the de-identification of data. De-identification is the process of stripping data of nearly all types of information that can be used to identify an individual, including names, social security numbers, addresses, and telephone numbers.

De-identification can be used to mask most any type of data, but the process is commonly associated with health data, which is generally considered to be highly sensitive. Under the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule, health data that is fully identifiable ("Protected Health Information") is subject to restrictions on access, use and disclosure. However, data that qualifies as "de-identified"[2] is not regulated at all. It may be used by anyone for any purpose, as long as it is not re-identified. Entities covered by HIPAA may access and share a limited data set[3] only for research, public health and health care operations purposes, and must enter into data use agreements with data recipients.

Under the HIPAA Privacy Rule, data can be de-identified in two ways. One, the "statistical method," requires that an experienced statistician or, more particularly, someone with "appropriate knowledge of and experience with generally acceptable statistical and scientific principles and methods for rendering information not individually identifiable" must determine that the "risk is very small that the information could be used, alone or in combination with other reasonably available information, by an anticipated recipient to identify an individual who is a subject of the information."[4] The statistician/expert must document the methods and results of his or her analysis.

The second, the "safe harbor" method, relies on the removal of 18 specific data elements that could uniquely identify an individual (for example, name, telephone numbers, social security numbers, and email addresses, among others). In employing this method, an

---

[1] Compendium of "Sensitive" Information Definitions, March 2008, http://www.cdt.org/privacy/20080324_info_compendium.pdf

[2] Under the HIPAA Privacy Rule, de-identified data is "health information that does not identify an individual and with respect to which there is no reasonable basis to believe that the information can be used to identify an individual." 45 C.F.R. §164.514(a).

[3] A limited data set is stripped of many categories of identifying information, but information often needed for public health and health research (such as birth dates and dates of treatment, and some geographic data) can be retained in the data.

[4] 45 C.F.R. §164.514(b)

entity covered by HIPAA must not have any "actual knowledge" that the remaining information can be used, alone or in combination with other data, to re-identify individuals.

Not all uses of de-identified data or limited data sets require identical levels of masking. Ideally, a broader spectrum of data de-identification options would meet the needs of different contexts and assure that data in the least identifiable form possible is accessed or disclosed for any given purpose. This principle holds true for other types of data, even data that may not be as sensitive as health information. Therefore, in assessing data de-identification options for any data type, it is critical to assess the nature of the data itself, and its intended use.

Release of government data has posed a challenge to federal entities in the past. For example, in an audit of court records obtained from PACER, Public.Resource.Org found that almost two thousand appellate decisions were released containing social security numbers or other non-public government identifiers that are required to be removed prior to record release.[5]

Re-identification

Just as technology allows for data to be masked and individual privacy to be protected, it also poses opportunities for these protections to be undone. Re-identification is the process of linking de-identified data to the actual identity of an individual person. Re-identification is aided by the existence of public records or commercially available databases. Because data.gov will house an enormous amount of raw information from various federal government agencies in one location, understanding the risk of re-identification is critical.

Unexpected re-identification has often caused a backlash against those who released the original data. For example, in August of 2006, a researcher at AOL de-identified search logs for 650,000 users and over 20 million search terms by replacing IP addresses with randomized numerical identifiers. Searchers were quickly re-identified using the released search queries, leading to lawsuits, firings, and significant public backlash.[6]

Re-identification has proven to be a major problem with a wide range of data sets. In one example, researchers positively identified one third of users of the social network Twitter using only linkage information within Twitter and Flickr.[7] In a similar experiment, data

---

[5] Letter from Carl Malamud to the Judicial Conference Committee, Confidential – 1,718 Personal Identifiers Found in Apellate Opinions, http://public.resource.org/scribd/7512583.pdf

[6] "Researchers Yearn to Use AOL Logs, but They Hesitate", Katie Hafner, August 23, 2006, The New York Times, http://www.nytimes.com/2006/08/23/technology/23search.html?ex=1313985600&en=cc878412ed34dad0&ei=5088&partner=rssnyt&emc=rss

[7] De-anonymizing Social Networks, Arvind Narayanan and Vitaly Shmatikov, http://www.cs.utexas.edu/~shmat/shmat_oak09.pdf

released by Netflix about user ratings in order to create a better matching algorithm, de-identified data was used to match users with IMDB profiles.[8]

These examples show how easily de-identified data can be used to re-identify individuals, often in conjunction with other publicly available data. Data does not exist in a vacuum. Any data released will be available in the context of all other public data, and re-identification must be addressed within that framework. Those posting data to data.gov should make every attempt to avoid the kind of backlash incurred by previous data releases by proactively checking data sets not only for their own identifiability, but also in the context of all information that is publicly available. Each data.gov data set will require its own specialized considerations based on the type of information being released and its relationship to other public data.

Key Principles for De-Identification and Use of Data Sets

CDT has spent considerable time studying the de-identification of health data. In September 2008, CDTs Health Privacy Project invited some of the nation's best thinkers on data security and policy to take part in a workshop on the de-identification of health data.[9] The participants discussed and weighed alternatives to resolve the problems of data liquidity and protection, and several key principles were put forth. These principles may also provide guidance in assessing data liquidity and protection of other types of data, including those to be made available on data.gov.

- Different levels of data protections are appropriate in different contexts. Limiting options by imposing a high degree of anonymity of data across the board also limits the value that can be derived from data.

- De-identification guidelines should be adaptable over time: it does not make sense to develop new de-identification guidelines that will become obsolete within a few years as technology and the data marketplace evolve. Any new mechanisms to protect de-identified data should instead be designed to incorporate a regular review process.

- De-identification rules must provide for ease of use for the entities engaged in de-identification of data.

- Any staff involved in de-identifying data or working with data that has been de-identified should participate in basic training about how best to protect privacy and security through organizational and technical means. Basic training, perhaps supported by data stewardship entities, would help to minimize the likelihood of

[8] Two researchers at the University of Texas at Austin used the Netflix Prize data to match anonymized Netflix users with their profiles at the popular movie information site IMDB.com. Robust De-Anonymization of Large Sparse Datasets, Arvind Narayanan and Vitaly Shmatikov, http://www.cs.utexas.edu/~shmat/shmat_oak08netflix.pdf

[9] CDT will be releasing a paper on the de-identification of health data and the findings of the workshop in Spring 2009.

breaches and other misuses of data.

- New standards may be strengthened by applying recent learning and technological developments, such as the use of limited access databases. That is, rather than a system in which data is used or disclosed either in de-identified form or fully identified form, data can be presented to a recipient in the most general form that is useful to that person. Data fields within a database can be analyzed to determine which are vulnerable to re-identification inference strategies, and data in those fields may then be aggregated, substituted or removed.[10]

For more information: Ari Schwartz, ari@cdt.org, 202-637-9800 x107
Alissa Cooper, acooper@cdt.org, 202-637-9800 x110
Heather West, heather@cdt.org, 202-637-9800 x315

[10] Sweeney presentation. Sweeney, "Weaving Technology and Policy Together to Maintain Confidentiality," Journal of Law, Medicine & Ethics, 25 (1997): 98-110.