Government Information, Data.gov and Privacy Implications

July 13, 2009
Tags: Array
Policy Posts are in-depth analyses on current tech policy issues from CDT experts. Sign up to receive the latest Policy Posts here:

One of the first projects adopted under the Obama administration's push for transparency was the Data.gov Web site, currently adding data sets by the hundreds of thousands. Data.gov will gather, bundle, and make publicly available various types of raw data produced by the federal government. 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. As part of day-to-day operations, the government collects information about economic indicators, health, product recalls, and government services. Many databases of this information are already made available in processed formats, but not made available in raw form.

A full copy of the memo upon which this Policy Post is based is held here [1].

---

1) Government Information, Data.gov and Privacy Implications

One of the first projects adopted under the Obama administration's push for transparency was the Data.gov Web site, currently adding data sets by the hundreds of thousands. Data.gov will gather, bundle, and make publicly available various types of raw data produced by the federal government. 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. As part of day-to-day operations, the government collects information about economic indicators, health, product recalls, and government services. Many databases of this information are already made available in processed formats, but not made available in raw form.

While Data.gov has great potential, there are important privacy implications associated with data disclosure. Readying data sets for Data.gov will require, prior to release, checking 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. Thus far, the data sets on Data.gov have avoided data sets with information about people.

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.

Data.gov Web Site [2]

CDT Compendium of "Sensitive" Information Definitions [3] (March 2008)

---

2) De-identification ad Re-Identification of Data Sets

One of the key principles of government transparency is that data should be released proactively. One way to protect privacy while making data useful to the public is by de-identifying the data prior to release. De-identification is the process of stripping data of information that can be used to identify an individual, including names, social security numbers, addresses, and telephone numbers.

Release of government data has posed a privacy 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.

De-identification can be used to remove identifying information from most kinds of data, but the process is commonly associated with highly sensitive data like health data. Under the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule, identifiable health data is subject to restrictions on access, use and disclosure. However, data that has been de-identified is not regulated at all under the HIPAA Privacy Rule, as it is considered not to contain personal information.

Under the HIPAA Privacy Rule, data can be de-identified in two ways. The first requires an experienced statistician to determine that there is only a "very small" risk that information could later be re-identified, using other available information or otherwise. The second method of de-identification removes 18 specific data elements that could identify an individual - for example, name, telephone numbers, social security numbers, and email addresses.

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.

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 for personal information 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. Additional examples of re-identification have been demonstrated on social networks and using public data.

If not properly de-identified, data can easily 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

identifiably, 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.

CDT paper [4] on Deidentification of Health Data

---

3) Key Principles for De-Identification and Use of Data Sets

CDT has spent considerable time studying the de-identification of health data, which has produced a set of principles that can be used for de-identification of data sets in general, including those to be made available on data.gov.

- Different levels of data protections are appropriate in different contexts. Imposing a high degree of anonymity of data across the board is not appropriate, and 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 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.

-
- Open Government

Links:
[1] http://cdt.org/privacy/Open_Govt_Directive_Deidentification.pdf
[2] http://www.data.gov/
[3] http://www.cdt.org/privacy/20080324_info_compendium.pdf
[4] http://www.cdt.org/healthprivacy/20090625_deidentify.pdf