THE CENTER FOR
OPEN DATA ENTERPRISE

Open Data Roundtable on Privacy:
KEY TAKEAWAYS

In 2016, the White House Office of Science and Technology Policy and the Center for Open Data Enterprise co-hosted four Open Data Roundtables to identify case studies, lessons learned, and best practices in open data across the federal government. Open data from government is free, publicly-available data that anyone can use and republish. We have prepared the key takeaways from each Roundtable, which brought together experts from inside and outside of government with technical, policy, and legal backgrounds. The Center will publish a full report of Roundtable findings in fall 2016.

BACKGROUND

On March 24th, 2016, the White House and the Center for Open Data Enterprise co-hosted a Roundtable to address a key issue: ***How to open granular information while protecting privacy.*** As more open data has become available, data users have come up against a conundrum. Many datasets in health, education, housing, and other areas may have the most value when they are released with "microdata" that can be analyzed at the level of individual records. However, releasing data at that level carries the risk of exposing personally identifiable information (PII) that could threaten individuals' privacy if it were released openly. Individual privacy should be treated in the context of public good, recognizing that many datasets with PII also contain information that provide great public benefit.

This privacy-focused Roundtable brought together 75 experts from government, nonprofits, academia, and the private sector to address this important privacy concern. Participants were not asked to develop consensus recommendations but to share their own observations and suggestions.

CURRENT CHALLENGES

Participants identified a number of challenges in protecting privacy while using open data.

- There is continuing concern about **the Mosaic Effect**, through which disparate datasets can be combined to identify individuals. It can be a particular problem when small groups of individuals are being studied. Potential negative impacts include breaches of individual privacy and the use of data that discriminates individuals.

- **De-identification technology** has limits as a solution. While technologies exist to remove identifying information from datasets, they are not fully effective. It is difficult to apply de-identification technology to the wide range of data types now available, such as geospatial,

medical and genomic, and body-camera data. It may also become possible to re-identify individuals more easily in the future as technology evolves.

- Some agencies are **risk averse** and very hesitant to release data that poses any risk of personal identification. The data breach at the Office of Personnel Management, for example, has been a cautionary example for agencies even though it did not occur due to intentional data release. At the same time, agencies may not fully recognize the benefits of opening data that justify some risk. Issues include attitudinal concerns at agencies and a lack of trust in agency safeguards among high-level decision-makers.

- Participants noted that there are **legal and policy** challenges in managing privacy concerns. Laws that are in place do not reflect current information and technology. Agencies must be conservative when the laws are too vague or not on point. There is also a lack of clarity on how to assess and mitigate risk as policies for handling PII data are inconsistent.

- When data is released to a new agency or organization, it is difficult to **track and honor agreements** regarding privacy. Consent for data to be used for a specific purpose does not anticipate future possible uses. For example, health patient data may be collected now and used in five years by companies that don't even exist yet. MOUs between agencies are seen as a limited, cumbersome, and ineffective. Additionally, resources for data management are scarce.

STRATEGIES FOR PROTECTING PRIVACY

While many government agencies are concerned about the privacy risk of opening data, policymakers can create programs and assessment tools that reduce these risks to release data for the public good. The following describe strategies discussed by participants to successfully release granular data in ways that address privacy concerns. Suggestions included the following.

- **Balancing tests:** Agencies can use strategies to balance the risks of releasing data against the potential for public good. This is the approach taken by the Consumer Financial Protection Bureau in carrying out its mandate to release data under the Home Mortgage Disclosure Act, which can be used to show whether mortgage lenders are discriminating in their loans. Following a recent rulemaking, the CFPB will use a "balancing test" with public input to determine the right balance of serving the public good and protecting individual privacy in this data release.

- **Differential access:** Participants discussed the need to consider gradations of openness under different circumstances. For example, some kinds of data could be made "open" only for sharing between federal agencies under certain conditions, or sharing only with qualified and vetted researchers. Some approaches include:
    - A federated model using a secure cloud repository and limiting access to trusted users.
    - Tiered access data-sharing programs to allow levels of access for multiple types of users.

Published by The Center for Open Data Enterprise, www.OpenDataEnterprise.org (2016)

- Opt-in and permission-based mechanisms that enable individuals to make their data more widely available. For example, individual patients have an incentive to share their medical data in the hope that it will be used to discover better treatments.

- Researchers can **use de-identification appropriately** as part of a strategy of privacy management. There are many situations where a certain level of de-identification is sufficient, even if it does not provide absolute, 100% privacy protection. Sensitive data can be de-identified by assigning individuals unique ID numbers to analyze their data without revealing their identities. Dropping non-critical information can also make re-identification more difficult.

- **Government can provide fundamental resources** to support a comprehensive, strategic privacy program across agencies. The following tools and programs were suggested by participants:
  - An open data resource website, which would help assess privacy risks, open data benefits, technical approaches, data provenance, and applicable policies.
  - An electronic tool for managing privacy issues in data. This resource could be based on the TurboTax platform and designed to manage sensitive datasets.
  - Centralized and shared services across agencies. Each agency may not need to have its own group responsible for privacy.
  - Training for government staff on managing privacy in open data. Based on the success of past programs, such as The San Francisco Data Academy's standardized curriculum, participants suggested establishing an Open Data University.

- **Develop new governance models** for addressing data privacy issues. Participants suggested taking action to:
  - Review and strengthen the role of Disclosure Review Boards. Develop a set of standard questions to develop a framework and guide their decision making processes.
  - Use the office of the Chief Data Officer to centralize each agency's data management.
  - Create model infrastructure – a virtual central data hub where access to data and APIs is managed by a common set of metadata (security and sharing licences) and user agreements.
  - Practice data lifecycle management to "bake in" privacy protection early on.

ABOUT THE CENTER

The Center for Open Data Enterprise is an independent nonprofit 501(c)3 organization, based in Washington, DC, whose mission is to maximize the value of open data as a public resource. We thank our Open Data Partner Microsoft and Open Data Supporter Booz Allen Hamilton for supporting the Center's work on the Open Data Roundtables. We welcome feedback on this report; please send comments and inquiries to Katherine Garcia, Director of Communications and Outreach, at katherine@odenterprise.org.

Published by The Center for Open Data Enterprise, www.OpenDataEnterprise.org (2016)