NISTIR 8150

Government Data De-Identification Stakeholder’s Meeting June 29, 2016 Meeting Report

Simson L. Garfinkel

This publication is available free of charge from:
http://dx.doi.org/10.6028/NIST.IR.8150

National Institute of
Standards and Technology
U.S. Department of Commerce

NISTIR 8150

Government Data De-Identification Stakeholder's Meeting June 29, 2016 Meeting Report

Simson L. Garfinkel
Information Access Division
Information Technology Laboratory

This publication is available free of charge from:
http://dx.doi.org/10.6028/NIST.IR.8150

September 2016

U.S. Department of Commerce
Penny Pritzker, Secretary

National Institute of Standards and Technology
Willie May, Under Secretary of Commerce for Standards and Technology and Director

National Institute of Standards and Technology Internal Report 8150
9 pages (September 2016)
This publication is available free of charge from:
http://dx.doi.org/10.6028/NIST.IR.8150

Certain commercial entities, equipment, or materials may be identified in this document in order to describe an experimental procedure or concept adequately. Such identification is not intended to imply recommendation or endorsement by NIST, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose.

There may be references in this publication to other publications currently under development by NIST in accordance with its assigned statutory responsibilities. The information in this publication, including concepts and methodologies, may be used by Federal agencies even before the completion of such companion publications. Thus, until each publication is completed, current requirements, guidelines, and procedures, where they exist, remain operative. For planning and transition purposes, Federal agencies may wish to closely follow the development of these new publications by NIST.

Organizations are encouraged to review all draft publications during public comment periods and provide feedback to NIST. All NIST Computer Security Division publications, other than the ones noted above, are available at http://csrc.nist.gov/publications.

National Institute of Standards and Technology
Attn: Computer Security Division, Information Technology Laboratory
100 Bureau Drive (Mail Stop 8930) Gaithersburg, MD 20899-8930

Reports on Computer Systems Technology

The Information Technology Laboratory (ITL) at the National Institute of Standards and Technology (NIST) promotes the U.S. economy and public welfare by providing technical leadership for the Nation's measurement and standards infrastructure. ITL develops tests, test methods, reference data, proof of concept implementations, and technical analyses to advance the development and productive use of information technology. ITL's responsibilities include the development of management, administrative, technical, and physical standards and guidelines for the cost-effective security and privacy of other than national security-related information in Federal information systems.

Abstract

The first Government Data De-Identification Stakeholder's Meeting was held at the National Institute of Standards and Technology on June 29, 2016. This meeting featured 80 participants from 67 different government agencies. Following the keynote, five panels discussed agency case studies, agency needs, available solutions, governance, and evaluation of de-identification techniques. Eighteen presenters from eleven agencies spoke for 10-minutes each. After each speaker's presentation, audience members asked questions and elaborated on points that the speakers made. Overall, it was the sense of the attendees that there is a need for collaboration and the sharing of techniques for the de-identification of government data.

Keywords

De-identification; HIPAA Privacy Rule; k-anonymity; differential privacy; re-identification; privacy

Audience

This document is intended for use by employees of the US government, advocacy groups, researchers and other members of communities that are concerned with technical issues involving the removal of personal information from government datasets so the datasets can be released to the general public.

Government Data De-Identification Stakeholder's Meeting
June 29, 2016.
National Institute of Standards and Technology
Heritage Room, 101 Bureau Drive, Gaithersburg, MD

MEETING REPORT [1]

Executive Summary

The first Government Data De-Identification Stakeholder's Meeting was held at the National Institute of Standards and Technology (NIST) on June 29, 2016. This government-only meeting featured 80 participants from 67 different government agencies.

The meeting opened with an inspirational charge regarding the government-wide importance of de-identification techniques by Marc Groman, Chair of the Federal Privacy Council and Senior Advisor for Privacy at the Office of Management and Budget. Dr. Ron Jarmin, Assistant Director for Research and Methodology, Census Bureau, followed with a keynote presentation that explored the need to quantify the balance between data quality and the potential harm to individuals that results from the application of de-identification techniques.

Following the keynote, five panels discussed agency case studies, agency needs, available solutions, governance, and evaluation of de-identification techniques. Eighteen presenters from eleven different agencies spoke for 10 min each. After each speaker's presentation, audience members asked questions and elaborated on points that the speakers made. The audience was engaged and many important issues were raised; roughly 2.5 h of the 7-h program was devoted to audience participants.

Meeting attendees stressed the need for clear technical guidance for de-identification, the need for metrics, the lack of tools, and the lack of trained individuals. Several attendees expressed interest in helping NIST to evaluate de-identification technologies.

1 Background

President Obama's executive order "Making Open and Machine Readable the New Default for Government Information" [2] and its implementation in OMB Memorandum M-13-13[3] charges US Government agencies with making government information resources publicly available whenever possible and legally permissible while simultaneously safeguarding "individual privacy, confidentiality, and national security."

[1] The NIST Government Data De-Identification Stakeholder's Meeting was funded by a NIST Building the Future proposal by Simson Garfinkel, Sean Brooks, Naomi Lefkovitz and Mary Theofanos. Erin Kenneally (DHS) and Christa Jones (Census) assisted in planning and co-chaired the meeting. This report was prepared by Simson L. Garfinkel and Julie Ryan, with Julie Haney assisting with the creation and the administration of the attendee survey.

[2] "Executive Order—Making Open and Machine Readable the New Default for Government Information," Barak Obama, The White House, May 9, 2013. https://www.whitehouse.gov/the-press-office/2013/05/09/executive-order-making-open-and-machine-readable-new-default-government-

[3] "Open Data Policy—Managing Information as an Asset," Sylvia M. Burwell, OMB M-13-13, May 9, 2013. https://www.whitehouse.gov/sites/default/files/omb/memoranda/2013/m-13-13.pdf

Many government information resources contain personal data about private citizens. Consistent with both the open data policy and with the Privacy Act of 1974 (5 USC §552a), these data about individuals should not be made publicly available as part of an open data offering. Instead, personal data[4] should be removed from the government information resources prior to publication, a process known as *de-identification.*[5]

De-identification is an approach for removing personal identifying information from a dataset with the goal of making it difficult, if not impossible, to re-identify the data by linking the released data with specific identities. Because of the growing desire for access to data for a variety of purposes, such as program management, social science research and evidence-based policymaking, software tools are being developed and marketed to support the de-identification of datasets.

Some federal agencies, such as the Census Bureau and the Department of Education, have decades of experience in de-identification. Other agencies are relatively inexperienced. Some agencies have created detailed processes for vetting de-identified information prior to release to assure that it cannot be re-identified. Others are only now beginning to address the issue.

Thus, the meeting had multiple purposes, including:

- Create a neutral forum to discuss and share de-identification methodologies and approaches between the federal statistical agencies, federal agencies involved in statistics,[6] and non-statistical agencies.
- Allow the Census Bureau to share some of its best practices regarding de-identification with non-statistical agencies.
- Assess US Government (USG) community needs.
- Discuss data release models that can meet USG needs and requirements.

2 Meeting Contents

The meeting consisted of two introductory speeches, a keynote, and five panels consisting of short (10 min) talks followed by audience discussion.

2.1 Overview

The meeting began with a welcome from Charles Romine, Director of the Information Technology Laboratory, NIST. Dr. Romine spoke of NIST's goal of using measurement science to improve our understanding of privacy and our government's commitment to privacy protection. He stressed that data de-identification is an important privacy protection tool, but that we need a better understanding of how de-identification works and its limitations.

Marc Groman, Senior Adviser for Privacy at the Office of Management and Budget (OMB) and Chair of the Federal Privacy Council, followed Dr. Romine with an inspirational charge to the meeting attendees. Mr. Groman said that his goal has been to change the perception of privacy within the US government from a compliance issue to an issue of risk-based analysis. Instead of having government leaders say that they cannot

[4] Personal data includes any information that can be used to identify an individual, including personal names, addresses that can be tied to an individual, email addresses, identifying numbers, salaries, and other kinds of data.

[5] In Europe, the phrase *anonymization* is frequently used as synonym for de-identification.

[6] There are 89 federal agencies that have annual budgets of at least $500,000 for statistical activities. https://fedstats.sites.usa.gov/

proceed on programs because of privacy concerns, he wants them to perform a privacy analysis and to either decrease the privacy risk of a program, or to acknowledge and "own" the privacy risk. With regards to de-identification, Groman said that the Nation sometimes faces contradicting needs of privacy protections, research, and transparency in an era of increasing amounts of data being collected and used. He said that the topic of de-identification comes up on a weekly basis in high-level meetings. He stressed that the government needs to break out of the mold of classifying data as being personally identifiable information (PII) or not, and that instead that government officials need to be able to evaluate the potential risks that might result from a data release and the specific techniques that could be used to mitigate those risks. Decision makers then need to be able to weigh the societal benefit that might result from a data release against the residual risk to individuals after a de-identification has taken place. Finally, agency officials need to accept the residual risk after mitigation has taken place. Groman said that agency officials should not cite privacy concerns as a reason not to move forward on a project. Instead, he said, officials need to learn how to pursue their goals while protecting privacy, or else make the argument that their goals are worth the privacy risk. Groman stressed that approaches to dealing with privacy-related issues must be based on an analysis of the sensitivity of datasets from multiple perspectives and that the answers needed to come from science. Closing his remarks, he invited all attendees to join OMB's PRIVACY-COUNCIL mailing list[7][8].

Ron Jarmin, Assistant Director for Research and Methodology for the Census Bureau, spoke next about the problems and challenges facing the Federal Government with regards to de-identification. Dr. Jarmin introduced the concept of a privacy-loss budget as a way to understand how to think about the tensions between the quality of data and the privacy issues. This theme was repeated throughout the meeting: if the datasets are de-identified so thoroughly as to preclude any potential re-identification, the dataset may be functionally useless or using it for analysis may lead to misleading or wrong conclusions. The calculation of probability of re-identification (the *re-identification risk*) can be used to characterize a privacy-loss budget across a span of projects, which reflects policy decisions on how best to allocate resources and support the goals of de-identification. The policy decisions can be made from several different perspectives, depending on who owns the data. Commercial structures, such as paying for privacy, provide one such structure, while social norms or values, such as requiring seatbelts in cars, may provide another such structure.

The rest of the meeting featured a series of moderated panels focusing on different aspects of the de-identification problem space. These panels were:

- Agency Case Studies, featuring speakers from the Census Bureau, the Consumer Financial Protection Bureau, the National Institutes of Health, and the Department of Transportation;
- Understanding the Problem Space: Agency Needs, featuring speakers from NIST, the Census Bureau, and the Department of Veteran Affairs;
- Solutions Available Today and Gap Analysis, featuring speakers from the Department of Homeland Security, the Census Bureau, and the Federal Trade Commission;
- Governance, featuring speakers from the Department of Education, the Millennium Challenge Corporation, and the Agency for International Development; and

[7] For information about the Privacy Council, please see https://community.max.gov/display/Egov/CIO+Council+Privacy+Community+of+Practice

[8] Note: The Privacy Council's new web presence, https://privacy.org/, should be operational before the end of fiscal year 2016.

- Evaluating De-Identification, featuring speakers from the Department of Veteran Affairs, Health and Human Services, the Census Bureau, and NIST.

After introductory remarks by the panel speakers, the moderator engaged the speakers and the audience members in an exploration of the panel topic. Audience participation was active, not only during the formal sessions, but also during the breaks.

The meeting concluded with remarks from Simson Garfinkel from the NIST Information Access Division, with an invitation to join the government de-identification practice mailing list.

2.2 Data types

Panelists and attendees discussed the range of data that are currently being de-identified by federal agencies:

- **Tabular data** is the dominant format of data published by the Census Bureau. This information is published in summary statistical tables, in some sets containing microdata, as synthetic data created from microdata, and through a query interface. Census makes some data generally available, while other data are only made available to qualified researchers through a series of Census-approved research centers.
- **Textual narratives** consisting of consumer complaints are distributed by the Consumer Financial Protect Bureau (CFPB). With textual narratives, the challenge is to remove identifying names and other proper nouns without damaging the meaning of the document.
- **Geographic information**, such as map coordinates or street addresses, routinely appear in tabular data. Street addresses are not considered private when used to refer to buildings, but they are considered private if they are used to refer to individuals.
- **Geographic trip information** is a sequence of geographic information, sometimes with timestamps. Trip information made available to researchers by the Department of Transportation is challenging because some locations are highly identifying while others are not. DOT resolves this issue by not providing trip starts and stops.
- **Video that** can contain public or private information. A speaker from DOT discussed the challenge in making video available to researchers in a way that would obscure the identity of the subject while preserving information of interest, such as direction of a person's gaze, their expression, and other information. DOT contractors have developed software that replaces the human's face with a mask made from another person's image. This mask hides identity while preserving facial expression and the eye direction.

Although the Federal Committee on Statistical Methodology has issued guidance to Statistical Disclosure Limitation,[9] that technique is limited to the protection of personal information in tabular and microdata files. No similar guidance exists for these other modalities. Furthermore, some participants expressed concern that traditional statistical disclosure limitation techniques offer less protection in a world where large amounts of data are available for correlation with officially released datasets. For example, two speakers from the Census Bureau were looking to adopt new disclosure limitation techniques that offered formal privacy guarantees.

[9] Report on Statistical Disclosure Limitation Methodology (Revised 2005), 1994, Federal Committee on Statistical Methodology. https://fcsm.sites.usa.gov/files/2014/04/spwp22.pdf

2.3 Inconsistent practices regarding de-identified data within the Federal Government

Statements by speakers and attendees indicated that there are many different practices regarding the use of de-identification and the distribution of de-identified data within the Federal Government.

- While some agencies have adopted formal disclosure review boards, many have not.
- Different approaches are being used to calculate re-identification risk.
- While the Census Bureau attempts to gauge the effectiveness of de-identification through attempted re-identification, this approach does not seem to be commonly employed in the other federal agencies.
- There is a lack of guidance regarding the appropriate time horizon for the protection of de-identification: should de-identified data be warranted to be de-identified indefinitely, or should the de-identification determination be revisited within a certain time window, to assess whether the risk has changed? Is it acceptable for the data to be identifiable after a certain amount of time? Current policies seem to indicate that de-identified data should remain de-identified forever, even though it may not be necessary to do so.
- More generally, threat models for re-identification and improper data use are frequently not articulated as part of the data release. They may not be documented by the releasing agency.

2.4 Synthetic Data

An alternative to de-identification is the releasing of synthetic datasets that are similar to the original data but which are inherently privacy-protecting.

Several speakers from the Census Bureau stated that the Bureau is increasing the use of synthetic data as a tool for protecting privacy while still releasing information that can be broadly used by many constituencies. None of the data elements in a synthetic dataset map to actual humans, but the statistics performed on the synthetic data approximate the same statistics applied to the real data. Synthetic data offers strong privacy guarantees, but significant research on synthetic data generation and validation is required to make these methods generally usable within the federal government.

- Many attendees were unfamiliar with techniques such as field swapping, noise introduction, and the creation of synthetic data to satisfy a public release requirement.
- Those attendees who were familiar with the term "synthetic data" used the term inconsistently. Some used it to imply any process of noise addition or field swapping, while others used it to imply that there was no one-to-one mapping between individuals in the original dataset and the published dataset.
- There was a general discomfort expressed by many of the attendees regarding the distribution and use of synthetic data.

2.5 Research and Technology Challenges

The discussion of de-identification research and technology challenges identified several themes that have been underrepresented in the existing research literature. These include:

- Many techniques do not adequately protect against the problem of **self-recognition** — an individual recognizing their own data in a dataset. Many attendees felt that self-recognition was a harm that needed to be protected against, but many academic researchers do not recognize self-recognition as a disclosure harm.

- **Natural language processing** is an important part of de-identification, since many datasets contain free-format text. However, the de-identification tools that can perform natural language processing that were discussed only handle English text. This is a problem, because what needs to be de-identified comes in many human languages.
- **Longitudinal datasets** with successive data releases are significantly harder to de-identify reliably, because there are more data points for each user that can be used to re-identify an individual.
- There needs to be more clarification on **remediation strategies**. For example, Federal agencies have "pulled-back" data that was improperly de-identified. Are there strategies (beyond counting downloads) for monitoring use of the data that were improperly released? Can data be or somehow watermarked so that provenance can be proven in downstream data products?
- **Linkage attacks** rely on finding data available for correlation with the released de-identified data. "High risk features" are those that may be found in many places are publicly available, while "low risk features" are those that may be comparatively private. However, several attendees expressed concern that there was no clear way to determine if a feature is high risk or low risk, or how to detect if a low risk feature transitioned to a high-risk feature because of an external data release. Several suggested that de-identified data should have a "use-by" date, after which the de-identification would need to be re-certified.
- In addition to the link risk and inference attacks in which other sources of available data are used to re-identify de-identified data, there is a related risk that the **release of seemingly benign data** can be combined to create derived data that is deemed sensitive.
- There is a need for implementation of **reusable frameworks/models** that take a risk-based approach to the balance between re-identification risk and data quality, the application of various technical and policy techniques for de-identification, and that offer a consistent way to articulate residual risk and utility resulting from the application of de-identification approaches.

2.6 Agency Challenges

Several attendees from different agencies noted that it was difficult to hire individuals with an expertise in de-identification, mirroring the problems that the government has had hiring in information security, privacy, and data analytics. Specific workforce issues noted by speakers and attendees included:

- Several speakers noted the difficulty of bridging concepts between technologists and policy specialists.
- Several speakers were concerned about the growing amount of data that needs to be de-identified and the relative lack of expertise within their organizations to perform de-identification. This is especially troublesome because every dataset must be individually evaluated—information that is not identifying in one dataset may be identifying in another.
- Organizations are looking for simple, rule-based approaches for de-identification, similar to the Safe Harbor provision of the Heath Insurance Portability and Accountability Act (HIPAA) Privacy Rule. However, the Safe Harbor provision has an "actual knowledge" clause that requires the exercise of judgement in the face of context-specific notice of certain risk factors before rendering an opinion on de-identification.

These workforce challenges are having profound impact on agencies that are increasingly relying on de-identification. For example, many de-identification efforts consider the risk of re-identification, but they do not consider the potential harms to the individual that would result from a successful re-identification. Absent such consideration, agencies may make inappropriate decisions based on misconception about true risk levels.

2.7 Governance

There appears to be little understanding of governance approaches within the Federal Government that could be used to exploit de-identification while minimizing the risk from the release of improperly de-identified data.

- Different federal organizations have created different governance structures for de-identification and data release. These tasks can be owned by a Freedom of Information Act (FOIA) office, a Privacy Office, or a specially created Data Release Board or Disclosure Review board (DRB).
- Speakers from organizations that had adopted DRBs expressed confidence in their boards and in the procedures that had been created.
- It is useful for a DRB to have a charter, broad participation from throughout the organization, and the support of senior management.
- It is unclear where the liability exists if data that are properly de-identified are later re-identified through some kind of inference attack.
- Federal organizations may inadvertently release data that they deem to be "low-risk" but which provides linkage information allowing other datasets to be re-identified.

3 Opportunities for future work

Following the meeting, NIST created "Government Data De-Identification Collaboration Forum" on MAX.GOV that contains the workshop proceedings, relevant publications, and a page of events and notices for the Government De-Identification Community. This virtual collaboration space is located at https://community.max.gov/x/IYTSPw. Included in the meeting proceedings are the speaker biographies and presentation materials.

Potential next steps for researchers and practitioners in this area include:

- Propose a common **terminology and taxonomy** for de-identification.
- Align de-identification risk assessment with NIST SP800-30, "Guide for Conducting Risk Assessments," and the forthcoming NISTIR 8062, "Privacy Risk Management for Federal Information Systems."
- Formalize and socialize **data release models** that are workable within the Federal Government. For example, government agencies are limited by regulatory frameworks, policy objectives, and the availability of individuals who are technically able to perform de-identification and data release.
- Formalize **threat models** including types and methods of attacks.
- Formalize **metrics for evaluating de-identification**, including data quality, faithfulness to the original data, for attacker types, and attack types.
- Explore opportunities for federal agencies to release **synthetic datasets** as an alternative to microdata releases.
- Create reference **de-identification tools**.
- Develop a **catalog** of existing de-identification tools.
- Develop **test methods** for de-identification tools and algorithms.

- Creating **venues** for the exchange of research findings.

Potential next steps in governance include:

- Develop a reproducible framework for **data review boards (DRBs)**, so that federal agencies that do not have a DRB will have a model for creating one.
- Developing **sharable case studies** from agencies.
- Developing a **community of interest** for those involved in dataset de-identification and release.
- Developing **training materials**.

4 Conclusion

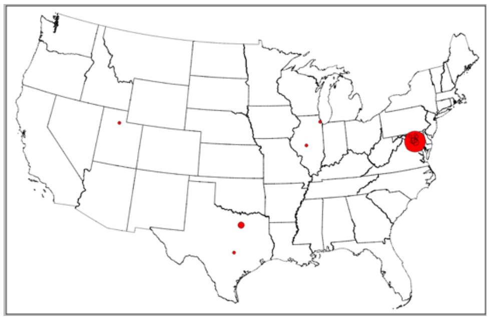
Overall, this meeting demonstrated that there is considerable heterogeneity within the US Government regarding needs, uses, and policy regarding the de-identification of government data. The findings of the presentations, and the comments of the attendees, and the follow-up survey indicates that there is a considerable need for policy on both matters of technology and governance. Such efforts can embrace existing NIST work on risk assessment, but will need to extend that work with attention to this specific set of issues.

A. Meeting Attendees and Survey Responses

NIST surveyed the government employee workshop attendees when they registered for the meeting and at the meeting's conclusion.

4.1 Registration Survey Results

As part of the meeting registration form NIST asked several questions of the attendees. Based on the results of those questions:



- 57 of the 106 registered attendees were from Washington DC.
- Of those from outside Washington, the most common worksites were Bethesda (7), Arlington (5), Suitland (4), Rockville (2), Silver Spring (2), College Park (2) and Reston (2).
- Those registered came from 67 different federal agencies.
- 42 % of the registered attendees (31 % by agency) stated that they were "involved in the release of datasets containing microdata."
- 50 % of the registered attendees (40 % by agency) stated that there was a software or methodology that they were currently employing to protect PII in microdata.
- The most popular de-identification techniques were swapping, noise addition, masking, and top-coding.
- 19 % of the registered attendees (16 % by agency) stated that they release synthetic datasets.

4.2 Survey Results (Summarized)

NIST requested that meeting attendees complete a post-meeting survey by providing a link to the survey at the workshop and by sending the link to attendees a week after the meeting had concluded. A total of 45 people visited the link, with 30 people completing the 2-page survey. Those responding to the survey overwhelmingly found value in the workshop, learned material that they said would be useful in their work, and indicated that they are interested in attending a future workshop on this topic.

Of those responding:

- 41 % stated that they were "personally involved in the release of datasets containing microdata."
- 76 % stated that their agency "currently use de-identification to remove PII from publicly released datasets."
- 82 % stated that their agency "have plans to use de-identification in the future."
- 63 % of the respondents strongly agreed with the statement "I found value in attending the workshop;" 31 % agreed with the statement.
- 52 % of the respondents strongly agreed with the statement "I gained new insights or learned about new approaches;" 42 % agreed with the statement.
- 68 % of the respondents strongly agreed with the statement "I would be interested in participating in future workshops on this topic;" 29 % agreed with the statement.