The attached DRAFT document (provided here for historical purposes) has been superseded by the following publication:

Publication Number: (2nd Draft) **NIST Special Publication (SP) 800-188**

Title: **De-Identifying Government Datasets**

Publication Date: **12/15/2016**

- http://csrc.nist.gov/publications/drafts/800-188/sp800_188_draft2.pdf

The following information was posted with the attached DRAFT document:

Aug 25, 2016

SP 800-188

DRAFT De-Identifying Government Datasets

NIST Requests Comments on a Draft Special Publication regarding the De-Identification of Government Datasets

De-identification removes identifying information from a dataset so that the remaining data cannot be linked with specific individuals. Government agencies can use de-identification to reduce the privacy risk associated with collecting, processing, archiving, distributing or publishing government data. Previously NIST published NISTIR 8053, *De-Identification of Personal Information*, which provided a survey of de-identification and re-identification techniques. This document provides specific guidance to government agencies that wish to use de-identification.

In developing the draft Privacy Risk Management Framework, NIST sought the perspectives and experiences of de-identification experts both inside and outside the US Government.

Future areas of work will focus on developing metrics and tests for de-identification software, as well as working with industry and academia to make algorithms that incorporate formal privacy guarantees usable for government de-identification activities.

Email comments to: sp800-188-draft <at> nist.gov(Subject: "Comments Draft SP 800-188")
Comments due by: **September 26, 2016**

DRAFT NIST Special Publication 800-188

De-Identifying Government Datasets

Simson L. Garfinkel

I N F O R M A T I O N S E C U R I T Y

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

DRAFT NIST Special Publication 800-188

De-Identifying Government Datasets

Simson L. Garfinkel
Information Access Division
Information Technology Laboratory

August 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

Authority

This publication has been developed by NIST in accordance with its statutory responsibilities under the Federal Information Security Modernization Act (FISMA) of 2014, 44 U.S.C. § 3551 *et seq.*, Public Law (P.L.) 113-283. NIST is responsible for developing information security standards and guidelines, including minimum requirements for federal information systems, but such standards and guidelines shall not apply to national security systems without the express approval of appropriate federal officials exercising policy authority over such systems. This guideline is consistent with the requirements of the Office of Management and Budget (OMB) Circular A-130.

Nothing in this publication should be taken to contradict the standards and guidelines made mandatory and binding on federal agencies by the Secretary of Commerce under statutory authority. Nor should these guidelines be interpreted as altering or superseding the existing authorities of the Secretary of Commerce, Director of the OMB, or any other federal official. This publication may be used by nongovernmental organizations on a voluntary basis and is not subject to copyright in the United States. Attribution would, however, be appreciated by NIST.

National Institute of Standards and Technology Special Publication 800-188
Natl. Inst. Stand. Technol. Spec. Publ. 800-188, 65 pages (August 2016)
CODEN: NSPUE2

Public comment period: *August 25, 2016* through *September 26, 2016*

National Institute of Standards and Technology
Attn: Information Access Division, Information Technology Laboratory
100 Bureau Drive (Mail Stop 8940) Gaithersburg, MD 20899-8940
Email: sp800-188-draft@nist.gov

All comments are subject to release under the Freedom of Information Act (FOIA).

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

De-identification removes identifying information from a dataset so that the remaining data cannot be linked with specific individuals. Government agencies can use de-identification to reduce the privacy risk associated with collecting, processing, archiving, distributing or publishing government data. Previously NIST published NISTIR 8053, "De-Identifying Personal Data," which provided a survey of de-identification and re-identification techniques. This document provides specific guidance to government agencies that wish to use de-identification. Before using de-identification, agencies should evaluate their goals in using de-identification and the potential risks that de-identification might create. Agencies should decide upon a de-identification release model, such as publishing de-identified data, publishing synthetic data based on identified data, and providing a query interface to identified data that incorporates de-identification. Agencies can use a Disclosure Review Board to oversee the process of de-identification; they can also adopt a de-identification standard with measurable performance levels. Several specific techniques for de-identification are available, including de-identification by removing identifiers and transforming quasi-identifiers and the use of formal de-identification models that rely upon Differential Privacy. De-identification is typically performed with software tools which may have multiple features; however, not all tools that mask personal information provide sufficient functionality for performing de-identification. This document also includes an extensive list of references, a glossary, and a list of specific de-identification tools, although the mention of these tools is only to be used to convey the range of tools currently available, and is not intended to imply recommendation or endorsement by NIST.

Keywords

privacy; de-identification; re-identification; Disclosure Review Board; data life cycle; the five safes; k-anonymity; differential privacy; pseudonymization; direct identifiers; quasi-identifiers; synthetic data.

Acknowledgements

The author wishes to thank the US Census Bureau for its help in researching and preparing this publication, with specific thanks to John Abowd, Ron Jarmin, Christa Jones, and Laura McKenna. The author would also like to thank Daniel Barth-Jones, Khaled El Emam and Bradley Malin providing invaluable insight in crafting this publication.

Audience

This document is intended for use by government engineers, data scientists, privacy officers, data review boards, and other officials. It is also designed to be generally informative to researchers and academics that are involved in the technical aspects relating to the de-identification of government data. While this document assumes a high-level understanding of information system security technologies, it is intended to be accessible to a wide audience.

Table of Contents

List of Appendices

Executive Summary

The US Government collects, maintains, and uses many kinds of datasets. Every federal agency creates and maintains internal datasets that are vital for fulfilling its mission, such as delivering services to taxpayers or ensuring regulatory compliance. Federal agencies can use de-identification to make government datasets available while protecting the privacy of the individuals whose data are contained within those datasets.[1]

Increasingly these government datasets are being made available to the public. For the datasets that contain personal information, agencies generally first remove that personal information from the dataset prior to making the datasets publicly available. *De-identification* is a term used within the US Government to describe the removal of personal information from data that are collected, used, archived, and shared.[2] De-identification is not a single technique, but a collection of approaches, algorithms, and tools that can be applied to different kinds of data with differing levels of effectiveness. In general, the potential risk to privacy posed by a dataset's release decreases as more aggressive de-identification techniques are employed, but data quality decreases as well.

The modern practice of de-identification comes from three distinct intellectual traditions:

- For four decades, official statistical agencies have researched and investigated methods broadly termed *Statistical Disclosure Limitation* (SDL) or *Statistical Disclosure Control*[3,4]

- In the 1990s there was an increase in the unrestricted release of microdata, or individual responses from surveys or administrative records. Initially these releases merely stripped obviously identifying information such as names and social security numbers (what are now called direct identifiers). Following some releases, researchers discovered that it was possible to re-identify individual data by triangulating with some of the remaining identifiers (now called quasi-identifiers or indirect identifiers).[5] The result of this

[1] Additionally, there are 13 Federal statistical agencies whose primary mission is the "collection, compilation, processing or analysis of information for statistical purposes." (Title V of the *E-Government Act of 2002*. *Confidential Information Protection and Statistical Efficiency Act* (CIPSEA), PL 107-347, Section 502(8).) These agencies rely on de-identification when making their information available for public use.

[2] In Europe the term *data anonymization* is frequently used as synonym for de-identification, but the terms may have subtly different definitions in some contexts. For a more complete discussion of de-identification and data anonymization, please see NISTIR 8053, *De-Identification of Personal Data*, Simson Garfinkel, September 2015, National Institute of Standards and Technology, Gaithersburg, MD.

[3] T. Dalenius, Towards a methodology for statistical disclosure control. *Statistik Tidskrift* 15, pp. 429-222, 1977

[4] An excellent summary of the history of Statistical Disclosure Limitation can be found in *Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics*, George T. Duncan, Thomas B. Jabine, and Virginia A. de Wolf, Editors; Panel on Confidentiality and Data Access, National Research Council, ISBN: 0-309-57611-3, 288 pages. http://www.nap.edu/catalog/2122/

[5] Sweeney, Latanya. Weaving Technology and Policy Together to Maintain Confidentiality. *Journal of Law, Medicine and Ethics*, Vol. 25 1997, p. 98-110.

research was the development of the k-anonymity model for protecting privacy,[6] which is reflected in the HIPAA Privacy Rule.

- In the 2000s, computer science research in the area of cryptography involving private information retrieval, database privacy, and interactive proof systems developed the theory of *differential privacy*,[7] which is based on a mathematical definition of the privacy loss to an individual resulting from queries on a database containing that individual's personal information. Starting with this definition, researchers in the field of differential privacy have developed a variety of mechanisms for minimizing the amount privacy loss associated with various database operations.

In recognition of both the growing importance of de-identification within the US Government and the paucity of efforts addressing de-identification as a holistic field, NIST began research in this area in 2015. As part of that investigation, NIST researched and published NIST Interagency Report 8053, *De-Identification of Personal Information*.[8]

Since the publication of NISTIR 8053, NIST has continued research in the area of de-identification. NIST met with de-identification experts within and outside the United States Government, convened a Government Data De-Identification Stakeholder's Meeting in June 2016, and conducted an extensive literature review.

The decisions and practices regarding the de-identification and release of government data can be integral to the mission and proper functioning of a government agency. As such, these activities should be managed by an agency's leadership in a way that assures performance and results in a manner that is consistent with the agency's mission and legal authority.

Before engaging in de-identification, agencies should clearly articulate their goals in performing the de-identification, the kinds of data that they intend to de-identify and the uses that they envision for the de-identified data. Agencies should also conduct a risk assessment that takes into account the potential adverse actions that might result from the release of the de-identified data; this risk assessment should include analysis of risk that might result from the data being re-identified and risk that might result from the mere release of the de-identified data itself.

One way that agencies can manage this risk is by creating a formal Disclosure Review Board (DRB) consisting of stakeholders within the organization and representatives of the organization's leadership. The DRB should evaluate applications for de-identification that describe the data to be released, the techniques that will be used to minimize the risk of disclosure, and how the effectiveness of those techniques will be evaluated.

---

[6] Latanya Sweeney. 2002. *k*-anonymity: a model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.* 10, 5 (October 2002), 557-570. DOI=http://dx.doi.org/10.1142/S0218488502001648

[7] Cynthia Dwork. 2006. Differential Pprivacy. In *Proceedings of the 33rd international conference on Automata, Languages and Programming - Volume Part II* (ICALP'06), Michele Bugliesi, Bart Preneel, Vladimiro Sassone, and Ingo Wegener (Eds.), Vol. Part II. Springer-Verlag, Berlin, Heidelberg, 1-12. DOI=http://dx.doi.org/10.1007/11787006_1

[8] NISTIR 8053, *De-Identification of Personal Data*, Simson Garfinkel, September 2015, National Institute of Standards and Technology, Gaithersburg, MD

Several specific models have been developed for the release of de-identified data. These include:

- **The Release and Forget model:**[9] The de-identified data may be released to the public, typically by being published on the Internet.
- **The Data Use Agreement (DUA) model:** The de-identified data may be made available to qualified users under a legally binding data use agreement that details what can and cannot be done with the data.
- **The Simulated Data with Verification Model:** The original dataset is used to create a simulated dataset that contains many of the aspects of the original dataset. The simulated dataset is released, either publically or to vetted researchers. The simulated data can be used to develop queries or analytic software; these queries and/or software can then be provided to the agency and be applied on the original data. The results of the queries and/or analytics processes can then be subjected to Statistical Disclosure Limitation and the results provided to the researchers.
- **The Enclave model:**[10,11] The de-identified data may be kept in some kind of segregated enclave that restricts the export of the original data, and instead accepts queries from qualified researchers, runs the queries on the de-identified data, and responds with results.

Agencies can create or adopt standards to guide those performing de-identification. The standards can specific disclosure techniques, or they can specify privacy guarantees that the de-identified data must uphold. There are many techniques available for de-identifying data; most of these techniques are specific to a particular modality. Some techniques are based on ad-hoc procedures, while others are based on formal privacy models that make it possible to rigorously calculate the amount of data manipulation required of the data to assure a particular level of privacy protection.

De-identification is generally performed by software. Features required of this software includes detection of identifying information; calculation of re-identification probabilities; performing de-identification; mapping identifiers to pseudonyms; and providing for the selective revelation of pseudonyms. Today there are several non-commercial open source programs for performing de-identification but only a few commercial products. Currently there are no performance standards, certification, or third-party testing programs available for de-identification software.

[9] Ohm, Paul, Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization. UCLA Law *Review*, Vol. 57, p. 1701, 2010

[10] Ibid.

[11] O'Keefe, C. M. and Chipperfield, J. O. (2013), A Summary of Attack Methods and Confidentiality Protection Measures for Fully Automated Remote Analysis Systems. *International Statistical Review*, 81: 426–455. doi: 10.1111/insr.12021

1 Introduction

The US Government collects, maintains, and uses many kinds of datasets. Every federal agency creates and maintains internal datasets that are vital for fulfilling its mission, such as delivering services to taxpayers or ensuring regulatory compliance. Additionally, there are 13 Federal statistical agencies whose primary passion is the collection, compilation, processing or analysis of information for statistical purposes."[12]

Increasingly these datasets are being made available to the public. Many of these datasets are openly published to promote commerce, support scientific research, and generally promote the public good. Other datasets contain sensitive data elements and, as a result, are only made available on a limited basis. Some datasets are so sensitive that they cannot be made publicly available at all. Instead, agencies may choose to release summary statistics, or even create synthetic datasets that resemble the original data but which do not present a threat to privacy or security.

Privacy is integral to our society, and citizens cannot opt-out of providing information to the government. The principle that personal data provided to the government should generally remain confidential and not used in a way that would harm the individual is a bedrock principle of official statistical programs.[13] As a result, many laws, regulations and policies govern the release of data to the public. For example:

- US Code Title 13, Section 9 which governs confidentiality of information provided to the Census Bureau, prohibits "any publication whereby the data furnished by any particular establishment or individual under this title can be identified."

- The release of personal information by the government is generally covered by the Privacy Act of 1974[14] and the E-Government Act of 2002.[15] Specifically, the E-Government Act states that "[d]ata or information acquired by an agency under a pledge of confidentiality for exclusively statistical purposes shall not be disclosed by an agency in identifiable form, for any use other than an exclusively statistical purpose, except with the informed consent of the respondent."[16]

- The Confidentiality Information Protection and Statistical Efficiency Act of 2002 requires that federal statistical agencies "establish appropriate administrative, technical, and physical safeguards to insure the security and confidentiality of records and to protect against any anticipated threats or hazards to their security or integrity which could result

[12] Title V of the *E-Government Act of 2002*. *Confidential Information Protection and Statistical Efficiency Act* (CIPSEA), PL 107-347, Section 502(8).

[13] George T. Duncan, Thomas B. Jabine, and Virginia A. de Wolf, eds., Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics. National Academies Press, Washington. 1993.

[14] Pub.L. 93-579, 88 Stat. 1896, 5 U.S.C. § 552a.

[15] Pub.L. 107–347, 116 Stat. 2899, 44 U.S.C. § 101, H.R. 2458/S. 803

[16] Pub.L. 107-347 § 512 (b)(1).

in substantial harm, embarrassment, inconvenience, or unfairness to any individual on whom information is maintained."

- On January 21, 2009, President Obama issued a memorandum to the heads of executive departments and agencies calling for US government to be transparent, participatory and collaborative.[17,18] This was followed on December 8, 2009, by the Open Government Directive,[19] which called on the executive departments and agencies "to expand access to information by making it available online in open formats. With respect to information, the presumption shall be in favor of openness (to the extent permitted by law and subject to valid privacy, confidentiality, security, or other restrictions)."

- On February 22, 2013, the White House Office of Science and Technology Policy (OSTP) directed Federal agencies with over $100 million in annual research and development expenditures to develop plans to provide for increased public access to digital scientific data. Agencies were instructed to "[m]aximize access, by the general public and without charge, to digitally formatted scientific data created with Federal funds, while: i) protecting confidentiality and personal privacy, ii) recognizing proprietary interests, business confidential information, and intellectual property rights and avoiding significant negative impact on intellectual property rights, innovation, and U.S. competitiveness, and iii) preserving the balance between the relative value of long-term preservation and access and the associated cost and administrative burden."[20]

Thus, many Federal agencies are charged with releasing data in a form that permits future analysis but does not threaten individual privacy.

Minimizing privacy risk is not an absolute goal of Federal laws and regulations. Instead, privacy risk is weighed against other factors, such as transparency, accountability, and the opportunity for public good. This is why, for example, personally identifiable information collected by the Census Bureau remains confidential for 72 years, and is then transferred to the National Archives and Records Administration where it is released to the public.[21]

De-identification is a term used within the US Government to describe the removal of personal information from data that are collected, used, archived, and shared.[22] De-identification is not a single technique, but a collection of approaches, algorithms, and tools that can be applied to

---

[17] Barack Obama, *Transparency and Open Government*, The White House, January 21, 2009.

[18] OMB Memorandum M-09-12, *President's Memorandum of Transparency and Open Government—Interagency Collaboration*, February 24, 2009. https://www.whitehouse.gov/sites/default/files/omb/assets/memoranda_fy2009/m09-12.pdf

[19] OMB Memorandum M-10-06, *Open Government Directive*, December 8, 2009, M-10-06

[20] John P. Holden, Increasing Access to the Results of Federally Funded Scientific Research, Executive Office of the President, Office of Science and Technology Policy, February 22, 2013.

[21] The "72-Year Rule," US Census Bureau, https://www.census.gov/history/www/genealogy/decennial_census_records/the_72_year_rule_1.html . Accessed August 2016. See also Public Law 95-416; October 5, 1978.

[22] In Europe the term *data anonymization* is frequently used as synonym for de-identification, but the terms may have subtly different definitions in some contexts. For a more complete discussion of de-identification and data anonymization, please see *NISTIR 8053: De-Identification of Personal Data*, Simson Garfinkel, September 2015, National Institute of Standards and Technology, Gaithersburg, MD.

different kinds of data with differing levels of effectiveness. In general, the potential risk to privacy posed by a dataset's release decreases as more aggressive de-identification techniques are employed, but data quality of the de-identified dataset decreases as well. Decreased data quality may result in decreased utility for some or all of the intended users of the de-identified dataset. Therefore, any effort involving the release of data that contains personal information inherently involves making some kind of tradeoff.

Some users of de-identified data may be able to use the data to make inferences about private facts regarding the data subjects; they may even be able to re-identify the data subjects—that is, to undo the privacy guarantees of de-identification. Agencies that release data should understand what data they are releasing and the risk of re-identification.

Planning is essential for successful de-identification and data release. Data management and privacy protection should be an integrated part of scientific research. This planning will include research design, data collection, protection of identifiers, disclosure analysis, and data sharing strategy. In an operational environment, this planning includes a comprehensive analysis of the purpose of the data release and the expected use of the released data, the privacy protecting controls, and the ways that those controls could fail.

Proper de-identification can have significant cost, where cost can include time, labor, and data processing costs. But this effort, properly executed, can result in a data that has high value for a research community and the general public while still adequately protecting individual privacy.

1.1 Document Purpose and Scope

This document provides guidance regarding the selection, use and evaluation of de-identification techniques for US government datasets. It also provides a framework that can be adapted by Federal agencies to frame the governance of de-identification procedures. The ultimate goal of this document is to reduce disclosure risk that might result from an intentional data release.

1.2 Intended Audience

This document is intended for use by government engineers, data scientists, privacy officers, data review boards, and other officials. It is also designed to be generally informative to researchers and academics that are involved in the technical aspects relating to the de-identification of government data. While this document assumes a high-level understanding of information system security technologies, it is intended to be accessible to a wide audience.

1.3 Organization

The remainder of this publication is organized as follows: Section 2, "Introducing De-Identification", presents a background on the science and terminology of de-identification. Section 3, "Governance and Management of Data De-Identification," provides guidance to agencies on the establishment or improvement to a program that makes privacy-sensitive data available to researchers and the general public. Section 4, "Technical Steps for Data De-Identification," provides specific technical guidance for performing de-identification using a variety of mathematical approaches. Section 5, "Requirements for De-Identification Tools," provides a recommended set of features that should be in de-identification tools; this information

may be useful for potential purchasers or developers of such software. Section 6, "Evaluation," provides information for evaluating both de-identification tools and de-identified datasets. This publication concludes with Section 7, "Conclusion."

This publication also includes three appendices: "References," "Glossary," and "Specific De-Identification Tools."

2 Introducing De-Identification

This document presents recommendations for de-identifying government datasets.

As long as any utility remains in the data derived from personal information, there also exists the possibility, however remote, that some information might be linked back to the original individuals on whom the data are based. When de-identified data can be *re-identified,* the privacy protection provided by de-identification is lost. The decision of how or if to de-identify data should thus be made in conjunction with decisions of how the de-identified data will be used, shared or released. Even if a specific individual cannot be matched to a specific data record, de-identified data can be used to improve the accuracy of inferences regarding individuals whose de-identified data are in the dataset. This so-called *inference risk* cannot be eliminated if there is any information content in the de-identified data, but it can be minimized.

De-identification is especially important for government agencies, businesses, and other organizations that seek to make data available to outsiders. For example, significant medical research resulting in societal benefit is made possible by the sharing of de-identified patient information under the framework established by the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule, the primary US regulation providing for privacy of medical records. Agencies may also be required to de-identify records as part of responding to a Freedom of Information Act (FOIA) request.

2.1 Historical Context

The modern practice of de-identification comes from three distinct intellectual traditions.

- For four decades, official statistical agencies have researched and investigated methods broadly termed *Statistical Disclosure Limitation* (SDL) or *Statistical Disclosure Control*[23,24] Most of these methods were created to allow the release of statistical tables and *public use files* (PUF) that allow users to learn factual information or perform original research, while protecting the privacy of the individuals in the dataset. SDL is widely used in contemporary statistical reporting.

- In the 1990s, there was an increase in the release of *microdata* files for public use, with individual responses from surveys or administrative records. Initially these releases merely stripped obviously identifying information such as names and social security numbers (what are now called *direct identifiers).* Following some releases, researchers discovered that it was possible to re-identify individuals' data by triangulating with some of the remaining identifiers (now called *quasi-identifiers* or *indirect identifiers).*[25] The

[23] T. Dalenius, Towards a methodology for statistical disclosure control. *Statistik Tidskrift* 15, pp. 429-222, 1977

[24] An excellent summary of the history of Statistical Disclosure Limitation can be found in *Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics*, George T. Duncan, Thomas B. Jabine, and Virginia A. de Wolf, Editors; Panel on Confidentiality and Data Access, National Research Council, ISBN: 0-309-57611-3, 288 pages. http://www.nap.edu/catalog/2122/

[25] Sweeney, Latanya. Weaving Technology and Policy Together to Maintain Confidentiality. *Journal of Law, Medicine and*

result of this research was the development of the k-anonymity model for protecting privacy,[26] which is reflected in the HIPAA Privacy Rule. Software that measures privacy risk using k-anonymity is used to allow the sharing of medical microdata. This intellectual tradition is typically called *de-identification*, although this document uses the word de-identification to describe all three intellectual traditions.

- In the 2000s, computer science research in the area of cryptography involving private information retrieval, database privacy, and interactive proof systems developed the theory of *differential privacy,*[27] which is based on a mathematical definition of the privacy loss to an individual resulting from queries on a database containing that individual's personal information. Differential privacy is termed a *formal method for privacy protection* because it is based its definition of privacy and privacy loss is based on mathematical proofs.[28] Because of this power there is considerable interest in differential privacy in academia, commerce and business, but to date there have been few systems employing differential privacy that have been released for general use.

Separately, during the first decade of the 21st century there was a growing awareness within the US Government about the risks that could result from the improper handling and inadvertent release of personal identifying and financial information. This realization, combined with a growing number of inadvertent data disclosures within the US government, resulted in President George Bush signing Executive Order 13402 establishing an Identity Theft Task Force on May 10, 2006.[29] A year later the Office of Management and Budget issued Memorandum M-07-16[30] which required Federal agencies to develop and implement breach notification policies. As part of this effort, NIST issued Special Publication 800-122, *Guide to Protecting the Confidentiality of Personally Identifiable Information (PII).*[31] These policies and documents had the specific goal of limiting the accessibility of information that could be directly used for identity theft, but did not create a framework for processing government datasets so that they could be released without impacting the privacy of the data subjects.

2.2 NISTIR 8053

In recognition of both the growing importance of de-identification within the US Government

---

Ethics, Vol. 25 1997, p. 98-110.

[26] Latanya Sweeney. 2002. *k*-anonymity: a model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.* 10, 5 (October 2002), 557-570. DOI=http://dx.doi.org/10.1142/S0218488502001648

[27] Cynthia Dwork. 2006. Differential Privacy. In *Proceedings of the 33rd international conference on Automata, Languages and Programming - Volume Part II* (ICALP'06), Michele Bugliesi, Bart Preneel, Vladimiro Sassone, and Ingo Wegener (Eds.), Vol. Part II. Springer-Verlag, Berlin, Heidelberg, 1-12. DOI=http://dx.doi.org/10.1007/11787006_1

[28] Other formal methods for privacy include cryptographic algorithms and techniques with provably secure properties, privacy preserving data mining, Shamir's secret sharing, and advanced database techniques. A summary of such techniques appears in Michael Carl Tschantz and Jeannette M. Wing, *Formal Methods for Privacy,* Technical Report CMU-CS-09-154, Carnegie Mellon University, August 2009 http://reports-archive.adm.cs.cmu.edu/anon/2009/CMU-CS-09-154.pdf

[29] George Bush, Executive Order 13402, *Strengthening Federal Efforts to Protect Against Identity Theft*, May 10, 2006. https://www.gpo.gov/fdsys/pkg/FR-2006-05-15/pdf/06-4552.pdf

[30] OMB Memorandum M-07-16: *Safeguarding Against and Responding to the Breach of Personally Identifiable Information*, May 22, 2007. https://www.whitehouse.gov/sites/default/files/omb/memoranda/fy2007/m07-16.pdf

[31] Erika McCallister, Tim Grance, Karen Scarfone, Special Publication 800-122, Guide to Protecting the Confidentiality of Personally Identifiable Information (PII), April 2010. http://csrc.nist.gov/publications/nistpubs/800-122/sp800-122.pdf

and the paucity of efforts addressing de-identification as a holistic field, NIST began research in this area in 2015. As part of that investigation, NIST researched and published NIST Interagency Report 8053, *De-Identification of Personal Information*. That report provided an overview of de-identification issues and terminology. It summarized significant publications to date involving de-identification and re-identification. It did not make recommendations regarding the appropriateness of de-identification or specific de-identification algorithms.

Since the publication of NISTIR 8053, NIST has continued research in the area of de-identification. As part of that research NIST met with de-identification experts within and outside the United States Government, convened a Government Data De-Identification Stakeholder's Meeting in June 2016, and conducted an extensive literature review.

The result is this publication, which provides guidance to Government agencies seeking to use de-identification to make datasets containing personal data available to a broad audience without compromising the privacy of those upon whom the data are based. De-identification is one of several models for allowing the controlled sharing of sensitive data. Other models include the use of data processing enclaves and data use agreements between data producers and data consumers. For a more complete description of data sharing models, privacy preserving data publishing, and privacy preserving data mining, please see NISTIR 8053.

2.3 Terminology

While each of the de-identification traditions has developed its own terminology and mathematical models, they share many underlying goals and concepts. Where terminology differs, this document relies on the terminology developed in previous US Government and standards organization documents.

de-identification is the "general term for any process of removing the association between a set of identifying data and the data subject."[32] De-identification takes an *original dataset* and produces a *de-identified dataset*.

re-identification is the general term for any process that restores the association between a set of de-identified data and the data subject.

redaction is a kind of de-identifying technique that relies on suppression or removal of information. In general, redaction alone is not sufficient to provide formal privacy guarantees while assuring the usefulness of the remaining data.

anonymization is another term that is used for de-identification. The term is defined as "process that removes the association between the identifying dataset and the data subject." [33] Some authors use the terms "de-identification" and "anonymization" interchangeably. Others use "de-identification" to describe a process and "anonymization" to denote a specific kind of de-identification that cannot be reversed. In health care, the term anonymization is sometimes used to describe the destruction of a table that maps pseudonyms to real identifiers. However, the term

[32] ISO/TS 25237:2008(E) Health Informatics — Pseudonymization. ISO, Geneva, Switzerland. 2008, p. 3

[33] ISO/TS 25237:2008(E) Health Informatics — Pseudonymization. ISO, Geneva, Switzerland. 2008, p. 2

anonymization conveys the perception that the de-identified data *cannot* be re-identified. Absent formal methods for privacy protection, it is not possible to mathematically determine if de-identified data can be re-identified. Therefore, the word anonymization should be avoided.

In medical imaging, the term de-identification is used to denote "the process of removing real patient identifiers or the removal of all subject demographics from imaging data for anonymization," while the term *de-personalization* is taken to mean "the process of completely removing any subject-related information from an image, including clinical trial identifiers."[34] This terminology not widely used outside of the field of medical imaging and will not be used elsewhere in this document.

Because of the inconsistencies in the use and definitions of the word "anonymization," this document avoids the term except in this section and in the titles of some references. Instead, it uses the term "de-identification," with the understanding that sometimes de-identified information can sometimes be re-identified, and sometimes it cannot.

pseudonymization is a "particular type of anonymization that both removes the association with a data subject and adds an association between a particular set of characteristics relating to the data subject and one or more pseudonyms."[35] The term *coded* is frequently used in the healthcare setting to describe data that has been pseudonymized. NIST recommends that agencies treat pseudononymized data as being potentially re-identifiable.

Many government documents use the phrases *personally identifiable information* (PII) and *personal information*. PII is typically used to indicate information that contains identifiers specific to individuals, although there are a variety of definitions for PII in various laws, regulations, and agency guidance documents. Because of these differing definitions, it is possible to have information that *singles out* individuals but which does not meet a particular definition of PII. An added complication is that some documents use the phrase PII to denote any information that is attributable to individuals, or information that is uniquely attributable to a specific individual, while others use the term strictly for data that are in fact identifying.

This document avoids the term "personally identifiable information." Instead, the phrase *personal information* is used to denote information relating to individuals, and *identifying information* is used to denote information that identifies individuals. Therefore, identifying information is personal information, but personal information is not necessarily identifying information. *Private information* is used to describe information that is in a dataset that is not publicly available. Private information is not necessarily identifying.

This document envisions a *de-identification process* in which an *original dataset* containing personal information is algorithmically processed to produce a *de-identified* result. The result may be a *de-identified dataset*, or a *synthetic dataset*, in which the data were created by a model. This kind of de-identification is envisioned as a batch process. Alternatively, the de-identification process may be a system that accepts queries and returns response that do not leak

[34] Colin Miller, Joe Krasnow, Lawrence H. Schwartz, *Medical Imaging in Clinical Trials*, Springer Science & Business Media, Jan 30, 2014.

[35] ISO/TS 25237:2008(E) Health Informatics — Pseudonymization. ISO, Geneva, Switzerland. 2008, p. 5

identifying information. De-identified results may be corrected or updated and re-released on a periodic basis. Issues arising from periodic release are discussed in §3.4, "Data Release Models."

Disclosure "relates to inappropriate attribution of information to a data subject, whether an individual or an organization. Disclosure occurs when a data subject is identified from a released file (*identity disclosure*), sensitive information about a data subject is revealed through the released file (*attribute disclosure*), or the released data make it possible to determine the value of some characteristic of an individual more accurately than otherwise would have been possible (*inferential disclosure*)."[36]

Disclosure limitation is a general term for the practice of allowing summary information or queries on data within a dataset to be released without revealing information about specific individuals whose personal information is contained within the dataset. De-identification is thus a kind of disclosure limitation technique. Every disclosure limitation procedure results in some kind of *bias*, or inaccuracy, being introduced into the results.[37] One goal of disclosure limitation is to avoid the introduction of *non-ignorable biases.*[38] With respect to de-identification, a goal is that inferences learned from de-identified datasets are similar to those learned from the original dataset.

Two models for quantifying the privacy protection offered by de-identification are *k-anonymity* and *differential privacy*.

K-anonymity[39] is a framework for quantifying the amount of manipulation required of the quasi-identifiers to achieve a given desired level of privacy. The technique is based on the concept of an *equivalence class*, the set of records that have the same quasi-identifiers. A dataset is said to be *k-anonymous* if, for every specific combination of quasi-identifiers, there are at least *k* matching records. For example, if a dataset that has the quasi-identifiers (birth year) and (state) has *k*=4 anonymity, then there must be at least four records for every combination of (birth year, state). Subsequent work has refined *k-anonymity* by adding requirements for diversity of the sensitive attributes within each equivalence class (known as *l-diversity*[40] and requiring that the resulting data are statistically close to the original data (known as *t-closeness*[41]

[36] Statistical Policy Working Paper 22 (Second version, 2005), Report on Statistical Disclosure Limitation Methodology, Federal Committee on Statistical Methodology, December 2005. https://fcsm.sites.usa.gov/reports/policy-wp/

[37] For example, see Trent J. Alexander, Michael Davern and Betsy Stevenson, Inaccurate Age and Sex Data in the Census PUMS Files: Evidence and Implications, *Public Opinion Quarterly*, 74, no 3: 551-569, 2010.

[38] John M. Abowd and Ian M. Schmutte, Economic Analysis and Statistical Disclosure Limitation, *Brookings Papers on Economic Activity*, March 19, 2015. https://www.brookings.edu/bpea-articles/economic-analysis-and-statistical-disclosure-limitation/

[39] Latanya Sweeney. 2002. *k*-anonymity: a model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.* 10, 5 (October 2002), 557-570. DOI=10.1142/S0218488502001648 http://dx.doi.org/10.1142/S0218488502001648

[40] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. In Proc. 22nd Intnl. Conf. Data Engg. (ICDE), page 24, 2006.

[41] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian (2007). "t-Closeness: Privacy beyond k-anonymity and l-diversity". ICDE (Purdue University).

Differential privacy[42] is a model based on a mathematical definition of privacy that considers the risk to an individual from the release of a query on a dataset containing their personal information. Differential privacy is also a set of mathematical techniques that can achieve the differential privacy definition of privacy. Differential privacy prevents disclosure by adding non-deterministic noise (usually small random values) to the results of mathematical operations before the results are reported.[43] Differential privacy's mathematical definition holds that the result of an analysis of a dataset should be roughly the same before and after the addition or removal of the data from any individual. This works because the amount of noise added masks the contribution of any individual. The degree of sameness is defined by the parameter **ε** (epsilon). The smaller the parameter **ε**, the more noise is added, and the more difficult it is to distinguish the contribution of a single individual. The result is increased privacy for all individuals, both those in the sample and those in the population from which the sample is drawn who are not present in the dataset. Differential privacy can be implemented in an online query system or in a batch mode in which an entire dataset is de-identified at one time. In common usage, the phrase "differential privacy" is used to describe both the formal mathematical framework for evaluating privacy loss, and for algorithms that provably provide those privacy guarantees.

Every time a dataset containing private information is queried and the results of that query are released, a certain amount of privacy in the dataset is lost. Using this model, de-identifying a dataset can be viewed as subjecting the dataset to a large number of queries and presenting the results as a correlated whole. The *privacy loss budget* is the total amount of private information that can be released according to an organization's policy.

Comparing traditional disclosure limitation, *k*-anonymity and differential privacy, the first two approaches start with a mechanism and attempt to reach the goal of privacy protection, whereas the third starts with a formal definition of privacy and has attempted to evolve mechanisms that produce useful (but privacy-preserving) results. All of these techniques are currently the subject of academic research, so it is reasonable to expect new techniques to be developed in the coming years that simultaneously increase privacy protection while providing for high quality of the resulting de-identified data.

[42] Cynthia Dwork. 2006. Differential privacy. In *Proceedings of the 33rd international conference on Automata, Languages and Programming - Volume Part II* (ICALP'06), Michele Bugliesi, Bart Preneel, Vladimiro Sassone, and Ingo Wegener (Eds.), Vol. Part II. Springer-Verlag, Berlin, Heidelberg, 1-12. DOI=http://dx.doi.org/10.1007/11787006_1

[43] Cynthia Dwork, Differential Privacy, in *ICALP*, Springer, 2006

3 Governance and Management of Data De-Identification

The decisions and practices regarding the de-identification and release of government data can be integral to the mission and proper functioning of a government agency. As such, these activities should be managed by an agency's leadership in a way that assures performance and results that are consistent with the agency's mission and legal authority. As discussed above, the need for attention arises because of the conflicting goals of data transparency and privacy protection. Although many agencies once assumed that it is relatively straightforward to remove privacy sensitive data from a dataset so that the remainder could be released without restriction, experience has shown that this is not the case.[44]

Given the conflict and the history, there may be a tendence for government agencies to overprotect their data. Limiting the release of data clearly limits the risk of harm that might result from a data release. However, limiting the release of data also creates costs and risk for other government agencies (which will then not have access to the identified data), external organizations, and society as a whole. For example, absent the data release, external organizations will suffer the cost of re-collecting the data (if it is possible to do so), or the risk of incorrect decisions that might result from having insufficient information.

This section begins with a discussion of why agencies might wish to de-identify data and how agencies should balance the benefits of data release with the risks to the data subjects. It then discusses where de-identification fits within the data life cycle. Finally, it discusses options that agencies have for adopting de-identification standards.

3.1 Identifying Goals and Intended Uses of De-Identification

Before engaging in de-identification, agencies should clearly articulate their goals in performing the de-identification, the kinds of data that they intend to de-identify and the uses that they envision for the de-identified data.

In general, agencies may engage in de-identification to allow for broader access to data that previously contained privacy sensitive information. Agencies may also perform de-identification to reduce the risk associated with collecting, storing, and processing privacy sensitive data.

For example:

- **Federal Statistical Agencies** that collect, process, and publish data for use by researchers, business planners, and other well-established purposes. These agencies are likely to have in place established standards and methodologies for de-identification. As these agencies evaluate new approaches to de-identification, they should seek to document inconsistencies with previous data releases that may result.people with

- **Federal Awarding Agencies** are allowed under OMB Circular A-110 to require that institutions of higher education, hospitals, and other non-profit organizations receiving

[44] NISTIR 8053 §2.4, §3.6

federal grants provide the US Government with "the right to (1) obtain, reproduce, publish or otherwise use the data first produced under an award; and (2) authorize others to receive, reproduce, publish, or otherwise use such data for Federal Purposes."[45] Realizing this policy, awarding agencies can require that awardees establish data management plans (DMPs) for making research data publicly available. Such data are used for a variety of purposes, including transparency and reproducibility. In general, research data that contains personal information should be de-identified by the awardee prior to public release. Awarding agencies may establish de-identification standards to ensure the protection of personal information.

- **Federal Research Agencies** may wish to make de-identified data available to the general public to further the objectives of research transparency and allow others to reproduce and build upon their results. These agencies are generally prohibited from publishing research data that would contain personal information, requiring the use of de-identification.

- **All Federal Agencies** that wish to make available administrative or operational data for the purpose of transparency, accountability, or program oversight, or to enable academic research may wish to employ de-identification to avoid sharing data that contains privacy sensitive information on employees, customers, or others.

3.2 Evaluating Risks Arising from De-Identified Data Releases

Once the purpose of the data release is understood, agencies should identify the risk that might result from the data release. As part of this risk analysis, agencies should specifically evaluate the probability of re-identification, the negative actions that might result from re-identification, and strategies for remediation in the event re-identification takes place.

NIST provides detailed information on how to conduct risk assessments in NIST Special Publication 800-30*, Guide for Conducting Risk Assessments*.[46]

Risk assessments should be based on scientific, objective factors and take into account the best interests of the individuals in the dataset—it should not be based on stakeholder interest. The goal of a risk evaluation is not to eliminate risk, but to identify which risks can be reduced while still meeting the objectives of the data release, and then deciding whether or not the residual risk is justified by the goals of the data release. A stakeholder may choose to accept risk, but stakeholders should not be empowered to prevent risk from being documented and discussed.

At the present time it is difficult to have measures of risk that are both general and meaningful. This represents an important area of research in the field of risk communication.

[45] OBM Circular A110, §36 (c) (1) and (2). Revised 11/19/93, as further amended 9/30/99. https://www.whitehouse.gov/omb/circulars_a110

[46] NIST Special Publication 800-30, *Guide for Conducting Risk Assessments*, Joint Task Force Transformation Initiative, September 2012. http://dx.doi.org/10.6028/NIST.SP.800-30r1

3.2.1 Probability of Re-Identification

Potential impacts on individuals from the release and use of de-identified data include:[47]

- **Identity disclosures** — Associating a specific individual with the corresponding record(s) in the data set. Identity disclosure can result from insufficient de-identification, re-identification by linking, or pseudonym reversal.

- **Attribute disclosure** — determining that an attribute described in the dataset is held bya specific individual, even if the record(s) associated with that individual is(are) not identified. Attribute disclosure can occur without identity disclosure if the de-identified dataset contains data from a significant number of relatively homogeneous individuals.[48] In these cases, de-identification does not protect against attribute disclosure.

- **Inferential disclosure** — being able to make an inference about an individual, even if the individual was not in the dataset prior to de-identification. De-identification cannot protect against inferential disclosure.

Although these disclosures are commonly thought to be atomic events involving the release of specific data, such as a person's name matched to a record, disclosures can result from the release of data that merely changes an adversary's probabilistic belief. For example, a disclosure might change an adversary's estimate that a specific individual is present in a dataset from a 50% probability to 90%. The adversary still doesn't *know* if the individual is in the dataset or not (and the individual might not, in fact, be in the dataset), but a disclosure has still taken place. Differential privacy provides a precise mathematical formulation of how information releases affect these probabilities.

Re-identification probability[49] is the probability that an attacker will be able to use information contained in a de-identified dataset to make inferences about individuals. Different kinds of re-identification probabilities can be calculated, including:

- *Known Inclusion Re-identification Probability* (KIRP). The probably of finding the record that matches a specific individual known to be in the population corresponding to a specific record. RRPdataset. KIRP can be expressed as the probability for a specific individual, the probability averaged over the entire dataset (ARRP),AKIRP).[50]

---

[47] Li Xiong, James Gardner, Pawel Jurczyk, and James J. Lu, "Privacy-Preserving Information Discovery on EHRs," in *Information Discovery on Electronic Health Records*, edited by Vagelis Hristidis, CRC Press, 2009.

[48] NISTIR 8053 §2.4, p 13.

[49] Note that previous publications described identification probability as "re-identification risk" and used scenarios such as a journalist seeking to discredit a national statistics agency and a prosecutor seeking to find information about a suspect as the basis for probability calculations. That terminology is not presented in this document in the interest of bringing the terminology of de-identification into agreement with the terminology used in contemporary risk analyses processes. See Elliot M, Dale A. Scenarios of attack: the data intruder's perspective on statistical disclosure risk, Netherlands Official Statistics 1999;14(Spring):6-10.

[50] Some texts refer to KIRP as "prosecutor risk." The scenario is that a prosecutor is looking for records belonging to a specific, named individual.

- *Unknown Inclusion Re-identification Probability* (UIRP). The probability of finding the record that matches a specific individual, without first knowing if the individual is or the maximumis not in the dataset. UIRP can be expressed as a probability for an individual record in the dataset.probability averaged over the entire population (AUIRP).[51]
- *Recording matching probability* (RMP). The probably of finding the record that matches a specific individual chosen from the population. RMP can be expressed as the probability for a specific record (RMP), the probability averaged over the entire dataset (ARMP), or the maximum probability over the entire dataset.
- *Inclusion probability* (IP), the probability that a specific individual's presence in the dataset can be inferred.

Whether or not it is necessary to calculate these probabilities depends upon the specifics of each intended data release. For example, many cities publicly disclose whether or not the taxes have been paid on a given property. Given that this information is already public, it may not be necessary to consider inclusion probably when a dataset of property taxpayers for a specific dataset is released. Likewise, there may be some attributes in a dataset that are already public and thus do not need to be protected with disclosure limitation techniques. However, the existence of such attributes may themselves pose a re-identification risk for other information in this dataset, or in other de-identified datasets

It may be difficult to calculate specific re-identification probabilities, as the ability to re-identify depends on the original dataset, the de-identification technique, the technical skill of the attacker, the attacker's available resources, and the availability of additional data that can be linked with the de-identified data. In many cases, the probability of re-identification will increase over time as techniques improve and more contextual information become available (*e.g.*, publicly or through a purchase).

De-identification practitioners have traditionally quantified re-identification probability in part based on the skills and abilities of a potential data intruder. Datasets that were thought to have little interest or possibility for exploitation were deemed to have a lower re-identification probability than datasets containing sensitive or otherwise valuable information. Such approaches are not appropriate when attempting to evaluate the re-identification probability of government datasets:

- Although a specific de-identified dataset may not be seen as sensitive, de-identifying that dataset may be an important step in de-identifying another dataset that is sensitive. Alternatively, the adversary may merely wish to embarrass the government agency. Thus, adversaries may have a strong incentive to re-identify datasets that are seemingly innocuous.
- Although the general public may not be skilled in re-identification, many resources on the modern Internet makes it easy to acquire specialized datasets, tools, and experts for specific re-identification challenges.

[51] Some texts refer to UIRP as "journalist risk." The scenario is that a journalist has obtained the de-identified file and is trying to identify one of the data subjects, but that the journalist fundamentally does not care *who* is identified.

Instead, de-identification practitioners should assume that de-identified government datasets will be subjected to sustained, world-wide re-identification attempts, and they should gauge their de-identification requirements accordingly.

Members of vulnerable populations (e.g. prisoners, children, people with disabilities) may be more susceptible to having their identities disclosed by de-identified data than non-vulnerable populations. Likewise, residents of areas with small populations may be more susceptible to having their identities disclosed than residents of urban areas. Individuals with multiple traits will generally be more identifiable if the individual's location is geographically restricted. For example, data belonging to a person who is labeled as a pregnant, unemployed female veteran will be more identifiable if restricted to Baltimore County, MD than to North America.

3.2.2 Adverse Impacts Resulting from Re-Identification

As part of a risk analysis, agencies should attempt to enumerate specific kinds of adverse impacts that can result from the re-identification of de-identified information. These can include potential impact on individuals, the agency, and society as a whole.

Potential adverse impacts on individuals include:

- Increased availability of personal information leading to an increased risks of fraud or identity theft.
- Increased availability of an individual's location, putting that person at risk for burglary, property crime, assault, or other kinds of violence.
- Increased availability an individual's private information, exposing potentially embarrassing information or information that the individual may not otherwise choose to reveal to the public.

Potential adverse impacts to an agency resulting from a successful re-identification include:

- Embarrassment or reputational damage if it can be publicly demonstrated that de-identified data can be re-identified.
- Direct harm to the agency's operations as a result of having de-identified data re-identified.
- Financial impact resulting from the harm to the individuals (e.g. settlement of lawsuits).
- Civil or criminal sanctions against employees or contractors resulting from a data release contrary to US law.

Potential adverse impacts on society as a whole include:

- Damage to the practice of using de-identification information. De-identification is an important tool for promoting research and accountability. Poorly executed de-identification efforts may negatively impact the public's view of this technique and limit

its use as a result.

One way to calculate an upper bound on impact to an individual or the agency is to estimate the impact that would result from the inadvertent release of the original dataset. This approach will not calculate the upper bound on the societal impact, however, since that impact includes reputational damage to the practice of de-identification itself.

As part of a risk analysis process, agencies should enumerate specific measures that they will take to minimize the risk of identity successful re-identification.

3.2.3 Impacts other than re-identification

Risk assessments described in this section can also assess adverse impacts other than those that might result from re-identification. For example:

- The sharing of de-identified data might result in specific inferential disclosures which, in general, are not protected against by de-identification.
- The de-identification procedure might introduce bias or inaccuracies into the dataset that result in incorrect decisions.[52]
- Releasing a de-identified dataset might reveal non-public information about an agency's policies or practices.

3.2.4 Remediation

As part of a risk analysis process, agencies should attempt to enumerate techniques that could be used to mitigate or remediate harms that would result from a successful re-identification of de-identified information. Remediation could include victim education, the procurement of monitoring or security services, the issuance of new identifiers, or other measures.

3.3 Data Life Cycle.

NIST SP 1500-1 defines the data life cycle as "the set of processes in an application that transform raw data into actionable knowledge."[53] Currently there is no standardized model for the data life cycle.

Michener et al describe the data life cycle as a true cycle of Collect → Assure → Describe →

[52] For example, a personalized warfarin dosing model created with data that had been modified in a manner consistent with the differential privacy de-identification model produced higher mortality rates in simulation than a model created from unaltered data. See Fredrikson *et al.*, Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin Dosing, 23rd *Usenix Security Symposium*, August 20-22, 2014, San Diego, CA. Educational data de-identified according to the *k-anonymity* model can also resulte in the introduction of bias that led to spurious results. See Olivia Angiuli, Joe Blitzstein, and Jim Waldo, How to De-Identify Your Data, *Communications of the ACM*, December 2015, 58:12, pp. 48-55. DOI: 10.1145/2814340

[53] NIST Special Publication 1500-1, *NIST Big Data Interoperability Framework: Volume 1, Definitions*. NIST Big Data Public Working Group, Definitions and Taxonomies Subgroup. September 2015. http://dx.doi.org/10.6028/NIST.SP.1500-1

Deposit → Preserve → Discover → Integrate → Analyze → Collect.[54] It is unclear how de-identification fits into this life cycle, as the data owner typically retains access to the identified data.

> Chisholm and others in the business literature describe the data life cycle as a linear process that involves Data Capture → Data Maintenance → Data Synthesis → Data Usage → {Data Publication & Data Archival} → Data Purging.[55] Using this formulation, de-identification typically fits between the Data Usage and the {Data Publication & Data Archival} parts of the data life cycle. That is, fully identified data are used within the organization, but they are then de-identified prior to being published (as a dataset), shared or archived. However, de-identification could also be applied after collection, as part of the Assure (Michener) or Data Maintenance (Chisholm) steps, in the event that identified data were collected but the identifying information was not actually needed.

Indeed, applying de-identification throughout the data life cycle minimizes privacy risk and significantly easies the process of public release.

Agencies performing de-identification should document that:

- Techniques used to perform the de-identification are theoretically sound.
- Software used to perform the de-identification is reliable for the intended task.
- Individuals who performed the de-identification were suitably qualified.
- Tests were used to evaluate the effectiveness of the de-identification.
- Ongoing monitoring is in place to assure the continued effectiveness of the de-identification strategy.

No matter where de-identification is applied in the data life cycle, agencies should document the answers of these questions for each de-identified dataset:

- Are direct identifiers collected with the dataset?
- Even if direct identifiers are not collected, is it nevertheless still possible to identify the data subjects through the presence of quasi-identifiers?
- Where in the data life cycle is de-identification performed? Is it performed in only one place, or is it performed in multiple places?
- Is the original dataset retained after de-identification?
- Is there a key or map retained, so that specific data elements can be re-identified at a later time?
- How are decisions made regarding de-identification and re-identification?
- Are there specific datasets that can be used to re-identify the de-identified data? If so, what controls are in place to prevent intentional or unintentional re-identification?
- Is it a problem if a dataset is re-identified?

---

[54] Participatory design of DataONE—Enabling cyberinfrastructure for the biological and environmental sciences, *Ecological Informatics*, Vol. 11, Sept. 2012, pp. 5-15.

[55] Malcolm Chisholm, 7 Phases of a Data Life Cycle, Information Management, July 9, 2015. http://www.information-management.com/news/data-management/Data-Life-Cycle-Defined-10027232-1.html

- Is there mechanism that will inform the de-identifying agency if there is an attempt to re-identify the de-identified dataset? Is there a mechanism that will inform the agency of the attempt is successful?

3.4 Data Sharing Models

Agencies should decide the data release model that will be used to make the data available outside the agency after the data have been de-identified.[56] Options include:

- **The Release and Forget Model:**[57] The de-identified data may be released to the public, typically by being published on the Internet. It can be difficult or impossible for an organization to recall the data once released in this fashion and may limit information for future releases.

- **The Data Use Agreement (DUA) Model:** The de-identified data may be made available to under a legally binding data use agreement that details what can and cannot be done with the data. Typically, data use agreements may prohibit attempted re-identification, linking to other data, and redistribution of the data without a similarly binding DUA. A DUA will typically be negotiated between the data holder and qualified researchers (the "qualified investigator model" [58]), although they may be simply posted on the Internet with a click-through license agreement that must be agreed to before the data can be downloaded (the "click-through model" [59]).

- **The Simulated Data with Verification Model:** The original dataset is used to create a simulated dataset that contains many of the aspects of the original dataset. The simulated dataset is released, either publically or to vetted researchers. The simulated data can be used to develop queries or analytic software; these queries and/or software can then be provided to the agency, which will then apply them to the original data. The results of the queries and/or analytics processes can then be subjected to Statistical Disclosure Limitation and the results provided to the researchers.

- **The Enclave Model:**[60,61] The de-identified data may be kept in a segregated enclave that restricts the export of the original data, and instead accepts queries from qualified researchers, runs the queries on the de-identified data, and responds with results. Alternatively, vetted researchers may travel to the enclave to perform their research, as is

[56] NISTIR 8053 §2.5, p. 14

[57] Ohm, Paul, Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization. UCLA Law Review, Vol. 57, p. 1701, 2010

[58] K El Emam and B Malin, "Appendix B: Concepts and Methods for De-identifying Clinical Trial Data," in Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk, Institute of Medicine of the National Academies, The National Academies Press, Washington, DC. 2015

[59] Ibid.

[60] Ibid.

[61] O'Keefe, C. M. and Chipperfield, J. O. (2013), A Summary of Attack Methods and Confidentiality Protection Measures for Fully Automated Remote Analysis Systems. International Statistical Review, 81: 426–455. doi: 10.1111/insr.12021

done with the Federal Statistical Research Data Centers operated by US Census Bureau. Enclaves may be used to implement the verification step of the Simulated Data with Verification Model.

Sharing models should take into account the possibility of multiple or periodic releases. Just as repeated queries to the same dataset may leak personal data from the dataset, repeated de-identified releases by an agency may result in compromising the privacy of individuals unless each subsequent release is viewed in light of the previous release. Even if a contemplated release of an allegedly de-identified dataset does not directly reveal identifying information, Federal agencies should ensure that the release, combined with previous releases, will also not reveal identifying information.[62]

Instead of sharing an entire dataset, the data owner may choose to release a sample. If only a subsample is released, the probability of re-identification decreases, because an attacker will not know if a specific individual from the data universe is present in the de-identified dataset.[63] However, releasing only a subset may cause users to draw incorrect inferences on the data, and may not align with agency goals regarding transparency and accountability.

3.5 The Five Safes

The Five Safes is a popular framework created for "designing, describing and evaluating" data access systems, and especially access systems designed for the sharing of information from a national statistics institute such as the US Census Bureau or the UK Office for National Statistics, with a research community.[64] The framework proposes five "risk (or access) dimensions:"

- **Safe projects —** Is this use of the data appropriate?
- **Safe people —** Can the researchers be trusted to use it in an appropriate manner?
- **Safe data —** Is there a disclosure risk in the data itself?
- **Safe settings** — Does the access facility limit unauthorized use?
- **Safe outputs** — Are the statistical results non-disclosive?

Each of these dimensions is intended to be *independent.* That is, the legal, moral and ethical review of the research proposed by the "safe projects" dimension should be evaluated independently of the people proposing to conduct the research, and the location where the

[62] See Joel Havermann, plaintiff - Appellant, v. Carolyn W. Colvin, Acting Commissioner of the Social Security Administration, Defendant – Appellee, No. 12-2453, US Court of Appeals for the Fourth Circuit, 537 Fed. Appx. 142; 2013 US App. Aug 1, 2013. Joel Havemann v. Carolyn W. Colvin, Civil No. JFM-12-1325, US District Court for the District of Maryland, 2015 US Dist. LEXIS 27560, March 6, 2015.

[63] El Emam, Methods for the de-identification of electronic health records for genomic research, Genome Medicine 2011, 3:25 http://genomemedicine.com/content/3/4/25

[64] Desai, T., Ritchie, F. and Welpton, R. (2016) *Five Safes: Designing data access for research.* Working Paper. University of the West of England. Available from: http://eprints.uwe.ac.uk/28124

research will be conducted.

One of the positive aspects of the Five Safes framework is that it forces data owners to consider many different aspects of data release when considering or evaluating data access proposals. Frequently, the authors write, it is common for data owners to "focus on one, and only one, particular issue (such as the legal framework surrounding access to their data, or IT solutions)." With a framework such as the Five Safes, people who may be specialists in one area are focused to consider (or to explicitly not consider) a variety of different aspects of privacy protection.

The Five Safes framework can be used as a tool for designing access systems, for evaluating existing systems, for communication and for training. Agencies should consider using a framework such as The Five Safes for organizing risk analysis of data release efforts.

3.6 Disclosure Review Boards[65]

Disclosure Review Boards (DRBs), also known as Data Release Boards, are administrative bodies created within an organization that are charged with assuring that a data release meets the policy and procedural requirements of that organization. DRBs should be governed by a written *mission statement* and *charter* that are, ideally, approved by the same mechanisms that the organization uses to approve other organization-wide policies.

The DRB should have a mission statement that guides its activities. For example, the US Department of Education's DRB has the mission statement:

> "The Mission of the Department of Education Disclosure Review Board (ED-DRB) is to review proposed data releases by the Department's principal offices (POs) through a collaborate technical assistance, aiding the Department to release as much useful data as possible, while protecting the privacy of individuals and the confidentiality of their data, as required by law."[66]

The DRB charter specifies the mechanics of how the mission is implemented. A formal, written charter promotes transparency in the decision-making process, and assures consistency in the applications of its policies. It is envisioned that most DRBs will be established to weigh the interests of data release against those of individual privacy protection. However, a DRB may also be chartered to consider *group harms*[67] that can result from the release of a dataset beyond harm to individual privacy. Such considerations should be framed within existing organizational policy, regulation, and law. Some agencies may balance these concerns by employing data use models other than de-identification—for example, by establishing data enclaves where a limited number of vetted researchers can gain access to sensitive datasets in a way that provides data value while attempting to minimize the possibility for harm. In those agencies, a DRB would be

[65] Note: This section is based in part on an analysis of the Disclosure Review Board policies at the US Census Bureau, the US Department of Education, and the US Social Security Administration.

[66] The Data Disclosure Decision, Department of Education (ED) Disclosure Review Board (DRB), A Product of the Federal CIO Council Innovation Committee. Version 1.0, 2015. http://go.usa.gov/xr68F

[67] NISTIR 8053 §2.4, p. 13

empowered to approve the use of such mechanisms.

The DRB charter should specify the DRB's composition. To be effective, the DRB should include representatives from multiple groups, and should include experts in both technology and policy. It may be desired to have individuals representing the interests of potential users; such individuals need not come from outside of the organization. It may also be beneficial to include representation from among the public, specifically from groups represented in the data sets if they have a limited scope. It may be useful to have a representation from the organization's leadership team: such a representative helps establish the DRBs credibility with the rest of the organization. The DRB may also have members that are subject matter experts. The charter should establish rules for ensuring quorum, and specify if members can designate alternates on a standing or meeting-by-meeting basis. The DRB should specify the mechanism by which members are nominated and approved, their tenure, conditions for removal, and removal procedures.[68]

The charter should set policy expectations for recording keeping and reporting, including whether records and reports are considered public or restricted. The charter should indicate if it is possible to exclude sensitive decisions from these requirements and the mechanism for doing so.

To meet its requirement of evaluating data releases, the DRB should require that written applications be submitted to the DRB that specify the nature of the dataset, the de-identification methodology, and the result. An application may require that the proposer present the re-identification risk, the risk to individuals if the dataset is re-identified, and a proposed plan for detecting and mitigating successful re-identification.

DRBs may wish to institute a two-step process, in which the applicant first proposes and receives approval for a specific de-identification process that will be applied to a specific dataset, then submits and receives approval for the release of the dataset that has been de-identified according to the proposal. However, because it is theoretically impossible to predict the results of applying an arbitrary process to an arbitrary dataset,[69,70] the DRB should be empowered to reject release of a dataset even if it has been de-identified in accordance with an approved procedure, because performing the de-identification may demonstrate that the procedure was insufficient to protect privacy. The DRB may delegate the responsibility of reviewing the de-identified dataset, but it should not be delegated to the individual that performed the de-identification.

The DRB charter should specify if the Board needs to approve each data release by the organization or if it may grant blanket approval for all data of a specific type that is de-identified according to a specific methodology. The charter should specify duration of the approval. Given advances in the science and technology of de-identification, it is inadvisable that a Board be

[68] For example, in 2003 the Census Bureau had a 9-member Disclosure Review Board, with "six members representing the economic, demographic and decennial program areas that serve 6-year terms. In addition, the Board has three permanent members representing the research and policy areas." Census Confidentiality and Privacy: 1790-2002, US Census Bureau, 2003. pp. 34-35

[69] Church, A. 1936. 'A Note on the Entscheidungsproblem'. Journal of Symbolic Logic, 1, 40-41.

[70] Turing, A.M. 1936 .'On Computable Numbers, with an Application to the Entscheidungsproblem'. Proceedings of the London Mathematical Society, Series 2, 42 (1936-37), pp.230-265

empowered to grant release authority for an indefinite amount of time.

In most cases a single privacy protection methodology will be insufficient to protect the varied datasets that an agency may wish to release. That is, different techniques might best optimize the tradeoff between re-identification risk and data usability, depending on the specifics of each kind of dataset. Nevertheless, the DRB may wish to develop guidance, recommendations and training materials regarding specific de-identification techniques that are to be used. Agencies that standardize on a small number of de-identification techniques will gain familiarity with these techniques and are likely to have results that have a higher level of consistency and success than those that have no such guidance or standardization.

Although it is envisioned that DRBs will work in a cooperative, collaborative and congenial manner with those inside an agency seeking to release de-identified data, there will at times be a disagreement of opinion. For this reason, the DRB's charter should state if the DRB has the final say over disclosure matters or if the DRB's decisions can be overruled, by whom, and by what procedure. For example, an agency might give the DRB final say over disclosure matters, but allow the agency's leadership to replace members of the DRB as necessary. Alternatively, the DRB's rulings might merely be advisory, with all data releases being individually approved by agency leadership or its delegates.[71]

Finally, agencies should decide whether or not the DRB charter will include any kind of performance timetables or be bound by a service level agreement (SLA).

Key elements of a DRB:

- Written mission statement and charter.
- Members represent different groups within the organization, including leadership.
- Board receives written applications to release de-identified data.
- Board reviews *both* proposed methodology *and* the results of applying the methodology.
- Applications should identify risk associated with data release, including re-identification probability, potentially adverse events that would result if individuals are re-identified, and a mitigation strategy if re-identification takes place.
- Approvals may be valid for multiple releases, but should not be valid indefinitely.
- Mechanisms for dispute resolution.
- Timetable or service level agreement (SLA).

3.7 De-Identification Standards

Agencies can rely on de-identification standards to provide a standardized terminology, procedures, and performance criteria for de-identification efforts. Agencies can adopt existing de-identification standards or create their own. De-identification standards can be prescriptive or performance-based.

[71] At the Census Bureau, "staff members [who] are not satisfied with the DRB's decision, … may appeal to a steering committee consisting of several Census Bureau Associate Directors. Thus far, there have been few appeals, and the Steering Committee has never reversed a decision made by the Board." *Census Confidentiality and Privacy*: 1790-2002, p. 35,

3.7.1 Benefits of Standards

De-identification standards assist agencies in the process of de-identifying data prior to public release. Without standards, data owners may be unwilling to share data, as they may be unable to assess if a procedure for de-identifying data is sufficient to minimize privacy risk.

Standards can increase the availability of individuals with appropriate training by providing a specific body of knowledge and practice that training should address. Absent standards, agencies may forego opportunities to share data. De-identification standards can help practitioners to develop a community, certification and accreditation processes.

Standards decrease uncertainty and provide data owners and custodians with best practices to follow. Courts can consider standards as acceptable practices that should generally be followed. In the event of litigation, an agency can point to the standard and say that it followed good data practice.

3.7.2 Prescriptive De-Identification Standards

A prescriptive de-identification standard specifies an algorithmic procedure that, if followed, results in data that are de-identified.

The "Safe Harbor" method of the HIPAA Privacy Rule[72] is an example of a prescriptive de-identification standard. The intent of the Safe Harbor method is to "provide covered entities with a simple method to determination if [] information is adequately de-identified."[73] It does this by specifying 18 kinds of identifiers that, once removed, results in the de-identification of Protected Health Information (PHI) and the subsequent relaxing of privacy regulations. Although the Privacy Rule does state that a covered entity employing the Safe Harbor method must have no "actual knowledge" that the PHI, once de-identified, could still be used to re-identify individuals, covered entities are not obligated to employ experts or mount re-identification attacks against datasets to verify that the use of the Safe Harbor method has in fact resulted in data that cannot be re-identified.

Prescriptive standards have the advantages of being relatively easy for users to follow, but developing, testing, and validating such standards can be burdensome. Agencies creating prescriptive de-identification standards should assure that data de-identified according to the rules cannot be re-identified; such assurances frequently cannot be made unless formal privacy techniques such as *differential privacy* are employed.

Prescriptive de-identification standards carry the risk that the procedure specified in the standard may not sufficiently de-identify to avoid the risk of re-identification.

3.7.3 Performance Based De-Identification Standards

A performance based de-identification standard specifies properties that the dataset must have

[72] Health Insurance Portability and Accountability Act of 1996 (HIPAA) Privacy Rule Safe Harbor method §164.514(b)(2).

[73] *Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule,* US Department of Health and Human Services, Office for Civil Rights, 2010. http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/guidance.html#_edn32

after it is de-identified.

The "Expert Determination" method of the HIPAA Privacy Rule is an example of a performance based de-identification standard. Under the rule, a technique for de-identifying data is sufficient if an appropriate expert "determines 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."[74]

Performance based standards have the advantage of allowing users many different ways to solve a problem. As such, they leave room for innovation. Such standards also have the advantage that they can embody the desired outcome.

Performance based standards should be sufficiently detailed that they can be performed in a manner that is reliable and repeatable. For example, standards that call for the use of experts should specify how an expert's expertise is to be determined. Standards that call for the reduction of risk to an acceptable level should provide a procedure for determining that level.

3.8 Education, Training and Research

De-identifying data in a manner that preserves privacy can be a complex mathematical, statistical, and data-driven process. Frequently the opportunities for identity disclosure will vary from dataset to dataset. Privacy protecting mechanisms developed for one dataset may not be appropriate for others. For these reasons, agencies engaging in de-identification should ensure that their workers have adequate education and training in the subject domain. Agencies may wish to establish education or certification requirements for those who work directly with the datasets. Because de-identification techniques are modality dependent, agencies using de-identification may need to institute research efforts to develop and test appropriate data release methodologies.

[74] The Health Insurance Portability and Accountability Act of 1996 (HIPAA) Privacy Rule Expert Determination Method §164.514(b)(1).

4 Technical Steps for Data De-Identification

The goal of de-identification is to transform data in a way that protects privacy while preserving the validity of inferences drawn on that data. This section discusses technical options for performing de-identification and verifying the result of a de-identification procedure.

Agencies should adopt a detailed, written protocol for de-identifying data prior to commencing work on a de-identification project. The details of the protocol will depend on the particular de-identification approach that is pursued.

4.1 Determine the Privacy, Data Usability, and Access Objectives

Agencies intent on de-identifying data for release should determine the policies and standards that will be used to determine acceptable levels of data quality, de-identification, and risk of re-identification. For example:

- What is the purpose of the data release?
- What is the intended use of the data?
- What data sharing model (§3.4) will be used?
- Which standards for privacy protection or de-identification will be used?
- What is the level of risk that the project is willing to accept?
- How should compliance with that level of risk be determined?
- What are the goals for limiting re-identification? That only a few people be re-identified? That only a few people can be re-identified in theory, but no one will actually be re-identified in practice? That there will be a small percentage chance that everybody will be re-identified?
- What harm might result from re-identification, and what techniques that will be used to mitigate those harms?

Some goals and objectives are synergistic, while others are in opposition.

4.2 Data Survey

As part of the de-identification, agencies should conduct an analysis of the data that they wish to de-identify.

4.2.1 Data Modalities

Different kinds of data require different kinds of de-identification techniques.

- **Tabular numeric and categorical data** is the subject of the majority of de-identification research and practice. These datasets are most frequently de-identified by using

techniques based on the designation and removal of direct identifiers and the manipulation of quasi-identifiers. The chief criticism of de-identification based on direct and quasi-identifiers is that administrative determinations of quasi-identifiers may miss variables that can be uniquely identifying when combined and linked with external data—including data that are not available at the time the de-identification is performed, but become available in the future. De-identification can be evaluated using frameworks such as Statistical Disclosure Limitation (SDL) or k-anonymity. However, *risk determinations based on this kind of de-identification will be incorrect if direct and quasi-identifiers are not properly classified!* Tabular data may also be used to create a synthetic dataset that preserves some inference validity but does not have a 1-to-1 correspondence to the original dataset.

- **Dates and times** require special attention when de-identifying, because all dates within a dataset are inherently linked to the natural progression of time. Some dates and times are highly identifying, with others are not. Some of these linkages may be relevant to the purpose of the dataset, the identity of the data subjects, or both. Dates may also form the basis of linkages between dataset records or even within a record—for example, a record may contain the date of admission, the date of discharge, and the number of days in residence. Thus, care should be taken when de-identifying dates to locate and properly handle potential linkages and relationships: applying different techniques to different fields may result in information being left in a dataset that can be used for re-identification. Specific issues regarding date de-identification are discussed below in §4.2.2.

- **Geographic and map data** also require special attention when de-identifying, as some locations can be highly identifying, other locations are not identifying at all, and some locations are only identifying at specific times. As with dates and times, the challenge of de-identifying geographic locations comes from the fact that locations inherently link to an external reality. Identifying locations can be de-identified through the use of perturbation or generalization. The effectiveness such de-identification techniques for protecting privacy in the presence of external information has not been well characterized. [75] Specific issues regarding geographical de-identification are discussed below in §4.2.3.

- **Unstructured text** may contain direct identifiers, such as a person's name, or may contain additional information that can serve as a quasi-identifier. Finding such identifiers and distinguishing them from non-identifiers invariably requires domain-specific knowledge.[76] Note that unstructured text may be present in tabular datasets and require special attention.[77]

---

[75] NISTIR 8053, §4.5 p. 37

[76] NISTIR 8053, §4.1 p. 30

[77] For an example of how unstructured text fields can damage the policy objectives and privacy assurances of a larger structured dataset, see Andrew Peterson, *Why the names of six people who complained of sexual assault were published online by Dallas police*, The Washington Post, April 29, 2016. https://www.washingtonpost.com/news/the-

- **Photos and video** may contain identifying information such as printed names (e.g. name tags). There also exists a range of biometric techniques for matching photos of individuals against a dataset of photos and identifiers.[78]

- **Medical imagery** poses additional problems over photographs and video due to the presence of many different kinds of identifiers. For example, identifying information may be present in the image itself (e.g. a photo may show an identifying scar or tattoo), an identifier may be "burned in" to the image area, or an identifier may be present in the file metadata. The body part in the image itself may also recognized through the use of a biometric algorithm and dataset.[79]

- **Genetic sequences** and other kinds of sequence information can be identified by matching to existing databanks that match sequences and identities. There is also evidence that genetic sequences from individuals who are not in datasets can be matched through genealogical triangulation, a process that uses genetic information and other information as quasi-identifiers to single-out a specific identity.[80] At the present time there is no known method to reliably de-identify genetic sequences. Specific issues regarding the de-identification of genetic information is discussed below in §4.2.4.

An important early step in the de-identification of government data is to identify the data modalities that are present in the dataset. A dataset that is thought to contain purely tabular data may be found, upon closer examination, to include unstructured text or even photograph data.

4.2.2 De-identifying dates

Dates can exist many ways in a dataset. Dates may be in particular kinds of typed columns, such as a date of birth or the date of an encounter. Dates may be present as a number, such as the number of days since an epoch such as January 1, 1900. Dates may be present in the free text narratives. Dates may be present in photographs—for example, a photograph that shows a calendar or a picture of a computer screen that shows date information.

Several strategies have been developed for de-identifying dates:

- Under the HIPAA Privacy Rule, dates must be generalized to no greater specificity than the year (e.g. July 4, 1776 becomes 1776).

- Dates within a single person's record can be systematically adjusted by a random amount. For example, dates of a hospital admission and discharge might be systematically moved the same number of days (e.g. ±1000).[81]

---

switch/wp/2016/04/29/why-the-names-of-six-people-who-complained-of-sexual-assault-were-published-online-by-dallas-police/

[78] NISTIR 8053, §4.2 p. 32

[79] NISTIR 8053, §4.3 p. 35

[80] NISTIR 8053, §4.4 p. 36

[81] Office of Civil Rights, "Guidance Regarding Methods for De-identification of Protected Health Information in Accordance

- In addition to a systematic shift, the intervals between dates can be perturbed to protect against re-identification attacks involving identifiable intervals while still maintaining the ordering of events.

- Some dates cannot be arbitrarily changed without compromising data quality. For example, it may be necessary to preserve day-of-week or whether a day is a work day or a holiday.

- Likewise, some ages can be randomly adjusted without impacting data quality, while others cannot. For example, in many cases the age of an individual can be randomly adjusted ±2 years if the person is over the age of 25, but not if their age is between 1 and 3.

4.2.3 De-identifying geographical locations

Geographical data can exist in many ways in a dataset. Geographical locations may be indicated by map coordinates (e.g. 39.1351966, -77.2164013), street address (e.g. 100 Bureau Drive), or postal code (20899). Geographical locations can also be embedded in textual narratives.

The amount of noise required to de-identify geographical locations significantly depends on external factors. Identity may be shielded in an urban environment by adding ±100m, whereas a rural environment may require ±5Km to introduce sufficient ambiguity. A prescriptive rule, even one that accounts for varying population densities, may still not be applicable, if it fails to take into account the other quasi-identifiers in the data set. Noise should also be added with caution to avoid the creation of inconsistencies in underlying data—for example, moving the location of a residence along a coast into a body of water or across geo-political boundaries.

4.2.4 De-identifying genomic information

Deoxyribonucleic acid (DNA) is the molecule inside human cells that carries genetic instructions used for the proper functioning of living organisms. DNA present in the cell nucleus is inherited from both parents; DNA present in the mitochondria is only inherited from an organism's mother.

DNA is a repeating polymer that is made from four chemical bases: adenine (A), guanine (G), cytosine (C) and thymine (T). Human DNA consists of roughly 3 billion bases, of which 99% is the same in all people.[82] Modern technology allows the complete specific sequence of an individual's DNA to be chemically determined; it is also possible to use DNA microarray to probe for the presence or absence of specific DNA sequences at predetermined points in the genome. This approach is frequently used to determine the presence or absence of specific single nucleotide polymorphisms (SNPs).[83] DNA sequences and SNPs are the same for identical twins,

---

with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule", US Department of Health and Human Services, 2010. http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/guidance.html

[82] What is DNA, Genetics Home Reference, US National Library of Medicine. https://ghr.nlm.nih.gov/primer/basics/dna Accessed Aug 6, 2016.

[83] What are single nucleotide polymorphisms (SNPs), Genetics Home Reference, US National Library of Medicine. https://ghr.nlm.nih.gov/primer/genomicresearch/snp Accessed Aug 6, 2016

individuals resulting from divided embryos, and clones. With these exceptions, it is believed that no two humans have the same complete DNA sequence. With regards to SNPs, individual SNPs may be shared by many individuals, but it a sufficiently large number of SNPs that show sufficient variability is generally believed to produce a combination that is unique to a particular individual. Thus, there are some sections of the DNA sequence and some combinations of SNPs that have high variability within the human population as a whole and others that have significant conservation between individuals within a specific population or group.

When there is high variability, DNA sequences and SNPs can be used to match an individual with a historical sample that has been analyzed and entered into a dataset. However, the fact that genetic information is inherited has allowed researchers to determine the surnames and even the complete identities of individuals because the large number of individuals that have now been recorded allows for familial inferences to be made.[84]

Because of the high variability inherent in DNA, complete DNA sequences should be regarded as being identifiable. Likewise, biological samples for which DNA can be extracted should be considered as being identifiable. Subsections of an individual's DNA sequence and collections of highly variable SNPs should be regarded as being identifiable unless there it is known that there are many individuals that share the region of DNA or those SNPs.

4.3 A de-identification workflow

This section presents a general workflow that agencies can use to de-identify data. This workflow can be adapted as necessary.

Step 1. Identify the intended use of the released, de-identified data. This step is vital to assure that the reduction in data quality that invariably accompanies de-identification will not make the data unusable for the intended application.

Step 2. Identify the risk that would result from releasing the identified data without first de-identifying.

Step 3. Identify the data modalities that are present in the data to be de-identified. (See § 4.2.1 below.)

Step 4. Identify approaches that will be used to perform the de-identification.

Step 5. Review and remove (if appropriate) links to external files.

Step 6. Perform the de-identification using an approved method. For example, de-identification may be performed by removing identifiers and transforming quasi-identifiers (§4.4), by generating synthetic data (§4.5), or by developing an interactive query interface (§4.6).

[84] Gymrek *et al.*, Identifying Personal Genomes by Surname Inference, *Science* 18 Jan 2013, 339:6117.

Step 7. Export transformed data to a different system for testing and validation.

Step 8. Test the de-identified data quality. Perform analyses on the de-identified data to make sure that it has sufficient usefulness and data quality.

Step 9. Attempt re-identification. Examine the de-identified data to see if it can be re-identified. This step may involve the engagement of an outside tiger team.

Step 10. Document the de-identification techniques and the results in a written report.

4.4 De-identification by removing identifiers and transforming quasi-identifiers

De-identification based on the removal of identifiers and transformation of quasi-identifiers is one of the most common approaches for de-identification currently in use. This approach has the advantage of being conceptually straightforward and there being a long institutional history in using this approach within both federal statistical agencies and the healthcare industry. This approach has the disadvantage of being not based on formal methods for assuring privacy protection. The lack of formal methods does not mean that this approach cannot protect privacy, but it does mean that privacy protection is not assured.

Below is a sample protocol for de-identifying data by removing identifiers and transforming quasi-identifiers:[85]

Step 1. Determine the re-identification risk threshold. The organization determines acceptable risk for working with the dataset and possibly mitigating controls, based on strong precedents and standards (e.g., Working Paper 22: Report on Statistical Disclosure Control).

Step 2. Determine the information in the dataset that could be used to identify the data subjects. Identifying information can include:

a. **Direct identifiers**, such as names, phone numbers, and other information that unambiguously identifies an individual.
b. **Quasi-identifiers** that could be used in a linkage attack. Typically, quasi-identifiers identify multiple individuals and can be used to triangulate on a specific individual.

[85] This protocol is based on a protocol developed by Professors Khaled El Emam and Bradley Malin. See K. El Emam and B. Malin, "Appendix B: Concepts and Methods for De-Identifying Clinical Trial Data," in *Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk*, Institute of Medicine of the National Academies, The National Academies Press, Washington, DC. 2015

c. **High-dimensionality data**[86] that can be used to single out data records and thus constitute a unique pattern that could be identifying, if these values exist in a secondary source to link against.[87]

Step 3. Determine the direct identifiers in the dataset. An expert determines the elements in the dataset that serve only to identify the data subjects.

Step 4. Mask (transform) direct identifiers. The direct identifiers are either removed or replaced with pseudonyms.

Step 5. Perform threat modeling. The organization determines the additional information they might be able to use for re-identification, including both quasi-identifiers and non-identifying values that an adversary might use for re-identification.

Step 6. Determine the minimal acceptable data quality. In this step, the organization determines what uses can or will be made with the de-identified data.

Step 7. Determine the transformation process that will be used to manipulate the quasi-identifiers. Pay special attention to the data fields containing dates and geographical information, removing or recoding as necessary.

Step 8. Import (sample) data from the source dataset. Because the effort to acquire data from the source (identified) dataset may be substantial, El Emam and Malin recommend a test data import run to assist in planning.

Step 9. Review the results of the trial de-identification. Correct any coding or algorithmic errors that are detected.

Step 10. Transform the quasi-identifiers for the entire dataset.

Step 11. Evaluate the actual re-identification risk. The actual identification risk is calculated. As part of this evaluation, every aspect of the released dataset should be considered in light of the question, "can *this* information be used to identify someone?"

Step 12. Compare the actual re-identification risk with the threshold specified by the policy makers.

Step 13. If the data do not pass the actual risk threshold, adjust the procedure and Step 11. For example, additional transformations may be required. Alternatively, it may be necessary to remove outliers. Step 9: Set parameters and apply data transformations.

---

[86] Charu C. Aggarwal. 2005. On *k*-anonymity and the curse of dimensionality. In *Proceedings of the 31st international conference on Very large data bases* (VLDB '05). VLDB Endowment 901-909.

[87] For example, Narayanan and Shmatikov demonstrated that the set of movies that a person had watched could be used as an identifier, given the existence of a second dataset of movies that had been publicly rated. See Narayanan, Arvind and Shmatikov Vitaly: Robust De-anonymization of Large Sparse Datasets. IEEE Symposium on Security and Privacy 2008: 111-125

4.4.1 Removing or Transformation of Direct Identifiers

Once a determination is made regarding direct identifiers, they must be removed. Options for removal include:

- Masking with a repeating character, such as XXXXXX or 999999.
- Encryption. After encryption the cryptographic key should be discarded to prevent decryption or the possibility of a brute force attack. However, the key must not be discarded if there is a desire to employ the same transformation at a later point in time, but rather stored in a secure location separate from the de-identified dataset.
- Hashing with a keyed hash, such as an HMAC. The hash key should be have sufficient randomness to defeat a brute force attack aimed at recovering the hash key. For example, SHA-256 HMAC with a 256-bit randomly generated key. As with encryption, the key should be discarded unless there is a desire for repeatability. (Note: hash functions should not be used without a key.)
- Replacement with keywords, such as transforming "George Washington" to "PATIENT."
- Replacement by realistic surrogate values, such as transforming "George Washington" to "Abraham Polk."[88]

The technique used to remove direct identifiers should be clearly documented for users of the dataset, especially if the technique of replacement by realistic surrogate names is used.

If the agency plans to make data available for longitudinal research and contemplates multiple data releases, then the transformation process should be repeatable, and the resulting transformed identities are *pseudonyms*. Agencies should be aware that there is a significantly increased risk of re-identification if a repeatable transformation is used.

4.4.2 Pseudonymization

Pseudonymization is a way of labeling multiple de-identified records from the same individual so that they can be linked together. Pseudonymization is a form of masking identifiers; it is *not* a form of de-identification.[89]

Pseudonymization generally increases the risk that de-identified data might be re-identified. By linking together records, pseudonymization increases the opportunities of finding identified data that can be linked with the de-identified data in a record linkage attack. Pseudonymization also carries that risk that the pseudonymization technique itself might be inverted or otherwise

[88] A study by Carrell et. al found that using realistic surrogate names in the de-identified text like "John Walker" and "1600 Pennsylvania Ave" instead of generic labels like "PATIENT" and "ADDRESS" could decrease or mitigate the risk of re-identification of the few names that remained in the text, because "the reviewers were unable to distinguish the residual (leaked) identifiers from the ... surrogates." See Carrell, D., Malin, B., Aberdeen, J., Bayer, S., Clark, C., Wellner, B., & Hirschman, L. (2013). Hiding in plain sight: use of realistic surrogates to reduce exposure of protected health information in clinical text. Journal of the American Medical Informatics Association, 20(2), 342-348.

[89] For more information on pseudonymization, please see NISTIR 8053 §3.2 p. 16

reversed, directly revealing the identities of the data subjects.

4.4.3 Transforming Quasi-Identifiers

Once a determination is made regarding quasi-identifiers, they should be transformed. A variety of techniques are available to transform quasi-identifiers:

- **Top and bottom coding.** Outlier values that are above or below certain values are coded appropriately. For example, the HIPAA Privacy Rules calls for ages over 89 to be "aggregated into a single category of age 90 or older."[90]

- **Micro aggregation**, in which individual microdata are combined into small groups that preserve some data analysis capability while providing for some disclosure protection.[91]

- **Generalize categories with small values.** When preparing contingency tables, several categories with small values may be combined. For example, rather than reporting that there is 1 person with blue eyes, 2 people with green eyes, and 1 person with hazel eyes, it may be reported that there are 4 people with blue, green or hazel eyes.

- **Data suppression.** Cells in contingency tables with counts lower than a predefined threshold can be suppressed to prevent the identification of attribute combinations with small numbers.[92]

- **Blanking and imputing.** Specific values that are highly identifying can be removed and replaced with imputed values.

- **Attribute or record swapping**, in which attributes or records are swapped between records representing individuals. For example, data representing families in two similar towns within a county might be swapped with each other. "Swapping has the additional quality of removing any 100-percent assurance that a given record belongs to a given household,"[93] while preserving the accuracy of regional statistics such as sums and averages. For example, in this case the average number of children per family in the county would be unaffected by data swapping.

- **Noise infusion.** Also called "partially synthetic data," small random values may be added to attributes. For example, instead of reporting that a person is 84 years old, the person may be reported as being 79 years old. Noise infusion increases variance and leads to attenuation bias in estimated regression coefficients and correlations among attributes.[94]

---

[90] HIPAA § 164.514 (b).

[91] J. M. Mateo-Sanz, J. Domingo-Ferrer, a comparative study of microaggregation methods, *Qüestiió*, vol. 22, 3, p. 511-526, 1998.

[92] For example, see *Guidelines for Working with Small Numbers*, Washington State Department of Health, October 15, 2012. http://www.doh.wa.gov/

[93] *Census Confidentiality and Privacy*: 1790-2002, US Census Bureau, 2003, p. 31

[94] George T. Duncan, Mark Elliot, Juan-José Salazar-Gonzalez, *Statistical Confidentiality: Principles and Practice*, Springer,

The techniques are described in detail by two publications:

- *Statistical Policy Working Paper #2* (Second version, 2005) by the Federal Committee on Statistical Methodology.[95] This 137-page paper also includes worked examples of disclosure limitation, specific recommended practices for Federal agencies, profiles of federal statistical agencies conducting disclosure limitation, and an extensive bibliography.

- *The Anonymisation Decision-Making Framework*, by Mark Elliot, Elaine MacKey, Kieron O'Hara and Caroline Tudor, UKAN, University of Manchester, Manchester, UK. 2016. This 156-page book provides tutorials and worked examples for de-identifying data and calculating risk.

Swapping and noise infusion both introduce noise into the dataset, such that records literally contain incorrect data. These techniques can introduce sufficient noise to provide formal privacy guarantees.

All of these techniques impact data quality, but whether they impact data *utility* depends upon the downstream uses of the data. For example, top-coding household incomes will not impact a measurement of the 90-10 quantile ratio, but it will impact a measurement of the top 1% of household incomes.[96]

In practice, statistical agencies typically do not document in detail the specific statistical disclosure technique that they use to transform quasi-identifiers, nor do they document the parameters used in the transformations nor the amount of data that have been transformed, as documenting these techniques can allow an adversary to reverse-engineer the specific values, eliminating the privacy protection. [97] This lack of transparency can result in erroneous conclusions on the part of data users.

4.4.4 Challenges Posed by Aggregation Techniques

Aggregation does not necessarily provide privacy protection, especially when data is presented as part of multiple data releases. Consider the hypothetical example of a school uses aggregation to report the number of students performing below, at, and above grade level:

Performance	Students

2011, p. 113, cited in John M. Abowd and Ian M. Schmutte, *Economic Analysis and Statistical Disclosure Limitation*, Brookings Papers on Economic Activity, March 19, 2015. https://www.brookings.edu/bpea-articles/economic-analysis-and-statistical-disclosure-limitation/

[95] *Statistical Policy Working Paper 22 (Second version, 2005), Report on Statistical Disclosure Limitation Methodology*, Federal Committee on Statistical Methodology, Statistical and Science Policy, Office of Information and Regulatory Affairs, Office of Management and Budget, December 2005.

[96] Thomas Piketty and Emmanuel Saez, Income Inequality in the United States, 1913-1998, *Quarterly Journal of Economics* 118, no 1:1-41, 2003.

[97] John M. Abowd and Ian M. Schmutte, Economic Analysis and Statistical Disclosure Limitation, *Brookings Papers on Economic Activity*, March 19, 2015. https://www.brookings.edu/bpea-articles/economic-analysis-and-statistical-disclosure-limitation/

Below grade level	30-39
At grade level	50-59
Above grade level	20-29

The following month a new student enrolls and the school republishes the table:

Performance	Students
Below grade level	30-39
At grade level	50-59
Above grade level	30-39

By comparing the two tables, one can readily infer that the student who joined the school is performing above grade level. Because aggregation does not inherently protect privacy, its use is not sufficient to provide formal privacy guarantees.

4.4.5 Challenges posed by High-Dimensionality Data

Even after removing all of the unique identifiers and manipulating the quasi-identifiers, some data can still be identifying if it of sufficient high-dimensionality, if there exists a way to link the supposedly non-identifying values with an identity.[98]

4.4.6 Challenges Posed by Linked Data

Data can be linked in many ways. Pseudonyms allow data records from the same individual to be linked together over time. Family identifiers allow data from parents to be linked with their children. Device identifiers allow data to be linked to physical devices, and potentially link together all data coming from the same device. Data can also be linked to geographical locations.

Data linkage increases the risk of re-identification by providing more attributes that can be used to distinguish the true identity of a data record from others in the population. For example, survey responses that are linked together by household are more readily re-identified than survey responses that are not linked. For example, heart rate measurements may not be considered identifying, but given a long sequence of tests, each individual in a dataset would have a unique constellation of heart rate measurements, and thus the data set would be susceptible to being

[98] For example, consider a dataset of an anonymous survey that links together responses from parents and their children. In such a dataset, a child might be able to find their parents' confidential responses by searching for their own responses and then following the link. See also Narayanan, Arvind and Shmatikov Vitaly: Robust De-anonymization of Large Sparse Datasets. IEEE Symposium on Security and Privacy 2008: 111-125

linked with another data set that contains these same values.

Dependencies between records may result in record linkages even when there is no explicit linkage identifier. For example, it may be that an organization has new employees take a proficiency test within 7 days of being hired. This information would allow links to be drawn between an employee dataset that accurately reported an employee's start date and a training dataset that accurately reported the date that the test was administered, even if the sponsoring organization did not intend for the two datasets to be linkable.

4.4.7 Post-Release Monitoring

Following the release of a de-identified dataset, the releasing agency should monitor to assure that the assumptions made during the de-identification remain valid. This is because the identifiability of a dataset may increase over time.

For example, the de-identified dataset may contain information that can be linked to an internal dataset that is later the subject of a data breach. In such a situation, the data breach will also result in the re-identification of the de-identified dataset.

4.5 Synthetic Data

An alternative to de-identifying using the technique presented in the previous section is to use the original dataset to create a synthetic dataset.

Synthetic data can be created by two approaches: [99]

- Sampling an existing dataset and either adding noise to specific cells likely to have a high risk of disclosure, or replacing these cells with imputed values. (A "partially synthetic dataset.")

- Using the existing dataset to create a model and then using that model to create a synthetic dataset. (A "fully synthetic dataset.")

In both cases, the mathematics of differential privacy can be used to quantify the privacy protection offered by the synthetic dataset.

4.5.1 Partially Synthetic Data

A partially synthetic dataset is one in which some of the data is inconsistent with the original dataset. For example, data belonging to two families in adjoining towns may be swapped to protect the identity of the families. Alternatively, the data for an outlier variable may be removed and replaced with a range value that is incorrect (for example, replacing the value "60" with the range "30-35"). It is considered best practice that the data publisher indicate that some values have been modified or otherwise imputed, but not to reveal the specific values that have been

[99] Jörg Drechsler, Stefan Bender, Susanne Rässler, Comparing fully and partially synthetic datasets for statistical disclosure control in the German IAB Establishment Panel. 2007, United Nations, Economic Commission for Europe. Working paper, 11, New York, 8 p. http://fdz.iab.de/342/section.aspx/Publikation/k080530j05

modified.

4.5.2 Fully Synthetic Data

A fully synthetic dataset is a dataset for which there is no one-to-one mapping between data in the original dataset and in the de-identified dataset. One approach to create a fully synthetic dataset is to use the original dataset to create a high fidelity model, and then to use the model to produce individual data elements consistent with the model using a simulation.

Fully synthetic datasets cannot provide more information to the downstream user than was contained in the original model. Nevertheless, some users may prefer to work with the fully synthetic dataset instead of the model:

- Synthetic data provides users with the ability to develop queries and other techniques that can be applied to the real data, without exposing real data to users during the development process. The queries and techniques can then be provided to the data owner, which can run the queries or techniques on the real data and provide the results to the users.

- Analysts may discover things from the synthetic data that they don't see in the model, even though the model contains the information. However, such discoveries should be evaluated against the real data to assure that the things that were discovered were actually in the original data, and not an artifact of the synthetic data generation.

- Some users may place more trust in a synthetic dataset than in a model.

- When researchers form their hypotheses working with synthetic data and then verify their findings on actual data, they are protected from pretest estimation and false-discovery bias.[100]

Both high-fidelity models and synthetic data generated from models may leak personal information that is potentially re-identifiable; the amount of leakage can be controlled using formal privacy models (such as differential privacy) that typically involve the introduction of noise.

There are several advantages to agencies that chose to release de-identified data as a fully synthetic dataset:

- It can be very difficult or even impossible to map records to actual people, so fully synthetic data offers very good privacy protection.

- The privacy guarantees can be mathematically established and proven.

[100] John M. Abowd and Ian M. Schmutte, Economic Analysis and Statistical Disclosure Limitation, *Brookings Papers on Economic Activity*, March 19, 2015. p. 257. https://www.brookings.edu/bpea-articles/economic-analysis-and-statistical-disclosure-limitation/

- The privacy guarantees can remain in force even if there are future data releases.

Fully synthetic data also has these disadvantages and limitations:

- It is not possible to create pseudonyms that map back to actual people, because the records are fully synthetic.

- The data release may be less useful for accountability or transparency. For example, investigators equipped with a synthetic data release would be unable to find the actual "people" who make up the release, because they would not actually exist.

- It is impossible to find meaningful correlations or abnormalities in the synthetic data that are not represented in the model. For example, if a model is built by considering all possible functions of 1 and 2 variables, then any correlations found of 3 variables will be a spurious artifact of the way that the synthetic data were created, and not based on the underlying real data.

- Users of the data may not realize that the data are synthetic. Simply providing documentation that the data are fully synthetic may not be sufficient public notification, since the dataset may be separated from the documentation. Instead, it is best to indicate in the data itself that the values are synthetic. For example, names like "SYNTHETIC PERSON" may be placed in the data. Such names could follow the distribution of real names but obviously be not real.

4.5.3 Synthetic Data with Validation

Agencies that share or publish synthetic data can optionally make available a validation service that takes queries or algorithms developed with synthetic data and applies them to actual data. The results of these queries or algorithms can then then be compared with the results of running the same queries on the synthetic data and the researchers warned if the results are different. Alternatively, the results can be provided to the researchers after the application of statistical disclosure limitation.

4.5.4 Synthetic Data and Open Data Policy

Releases of synthetic data can be confusing to the lay public. Specifically, synthetic data may contain synthetic individuals who appear quite similar to actual individuals in the population. Furthermore, fully synthetic datasets do not have a zero disclosure risk, because they still convey some private information about individuals. The disclosure risk may be greater when synthetic data are created with traditional data imputing techniques, rather than techniques based on formal privacy models.

4.5.5 Creating a synthetic dataset with differential privacy

A growing number of mathematical algorithms have been developed for creating synthetic datasets that meet the mathematical definition of privacy provided by differential privacy. Most of these algorithms will transform a dataset containing private data into a new dataset that contains synthetic data that nevertheless provides reasonably accurate results in response to a variety of queries. However there is no algorithm or implementation currently in existence that

can be used by a person who is unskilled in the area of differential privacy.

The classic definition of differential privacy is that if results of function calculated on a dataset are indistinguishable within a certain privacy metric ϵ (epsilon) no matter whether any possible individual is included in the dataset or removed from the dataset,[101] then that function is said to provide ϵ-differential privacy.

In Dwork's mathematical formulation, the two datasets (with and without the individual) are denoted by D_1 and D_2, and the function that is said to be differential private is κ. The formal definition of differential privacy is then:

> **Definition 2.** [102] A randomized function κ gives ϵ-differential privacy if for all datasets D_1 and D_2 differing on at most one element, and all $S \subseteq \text{Range}(\kappa)$,

$$\Pr[\kappa(D_1) \in S] \leq e^{\epsilon} \times \Pr[\kappa(D_2) \in S]$$

This definition that may be easier to understand if rephrased as a dataset D with an arbitrary person p, and dataset $D - p$, the dataset without a person, and the multiplication operator replaced by a division operator, e.g.:

$$\frac{\Pr[\kappa(D - p) \in S]}{\Pr[\kappa(D) \in S]} \leq e^{\epsilon}$$

That is, the ratio between the probable outcomes of function κ operating on the datasets with and without person p should be less than e^{ϵ}. If the two probabilities are equal, then $e^{\epsilon} = 1$, and $\epsilon = 0$. If the difference between the two probabilities is potentially infinite—that is, there is no privacy—then $e^{\epsilon} = \infty$ and $\epsilon = \infty$.

What this means in practice for the creation of a synthetic dataset with differential privacy and a sufficiently large ϵ is that functions computed on the so-called "privatized" dataset will have a similar probability distribution no matter whether any person in the original data that was used to create the model is included or excluded. In practice, this similarity is provided by adding noise to the model. For datasets drawn from a population with a large number of individuals, the model (and the resulting synthetic data) will have a small amount of noise added. For models and resulting created from a small population (or for contingency tables with small cell counts), this will require the introduction of a significant amount of noise. The amount of noise added is determined by the differential privacy parameter ϵ, the number of individuals in the dataset, and the specific differential privacy mechanism that is employed.

Smaller values of ϵ provide for more privacy but decreased data quality. As stated above, the

[101] More recently, this definition has been taken to mean that any attribute of any individual within the dataset may be altered to any other value that is consistent with the other members of the dataset.

[102] From Cynthia Dwork. 2006. Differential privacy. In *Proceedings of the 33rd international conference on Automata, Languages and Programming - Volume Part II* (ICALP'06), Michele Bugliesi, Bart Preneel, Vladimiro Sassone, and Ingo Wegener (Eds.), Vol. Part II. Springer-Verlag, Berlin, Heidelberg, 1-12. DOI=http://dx.doi.org/10.1007/11787006_1. Definition 1 is not important for this publication.

value of 0 implies that the function κ provides the same answer no matter if anyone is removed or a person's attributes changed, while the value of ∞ implies that the original dataset is released with being privatized.

Many academic papers on differential privacy have assumed a value for of 1.0 or e but have not explained the rationale of the choice. Some researchers working in the field of differential privacy have just started the process of mapping existing privacy regulations to the choice of ϵ. For example, using a hypothetical example of a school that wished to release a dataset containing the school year and absence days for a number of students, the value of ϵ using one set of assumptions might be calculated to 0.3379 (producing a low degree of data quality), but this number can safely be raised to 2.776 (and correspondingly higher data quality) without significantly impacting the privacy protections.[103]

Another challenge in implementing differential privacy is the demands that the algorithms make on the correctness of implementation. For example, a Microsoft researcher discovered that four publicly available general purpose implementations of differential privacy contained a flaw that potentially leaked private information because of the binary representation of IEEE floating point numbers used by the implementations.[104]

Given the paucity of scholarly publications regarding the deployment of differential privacy in real-world situation, combined with the lack of guidance and experience in choosing appropriate values of ϵ, agencies that are interested in using differential privacy algorithms to allow querying of sensitive datasets or for the creation of synthetic data should take great care to assure that the techniques are appropriately implemented and that the privacy protections are appropriate to the desired application.

4.6 De-Identifying with an interactive query interface

Another model for granting the public access to de-identified agency information is to construct an interactive query interface that allows members of the public or qualified investigators to run queries over the agency's dataset. This option has been developed by several agencies and there are many different ways that it can be implemented.

- If the queries are run on actual data, the results can be altered through the injection of noise to protect privacy. Alternatively, the individual queries can be reviewed by agency staff to verify that privacy thresholds are maintained.

- Alternatively, the queries can be run on synthetic data. In this case, the agency can also run queries on the actual data and warn the external researchers if the queries run on

[103] Jaewoo Lee and Chris Clifton. 2011. How much is enough? choosing ε for differential privacy. In Proceedings of the 14th international conference on Information security (ISC'11), Xuejia Lai, Jianying Zhou, and Hui Li (Eds.). Springer-Verlag, Berlin, Heidelberg, 325-340.

[104] Ilya Mironov. 2012. On significance of the least significant bits for differential privacy. In Proceedings of the 2012 ACM conference on Computer and communications security (CCS '12). ACM, New York, NY, USA, 650-661. DOI: http://dx.doi.org/10.1145/2382196.2382264

synthetic data diverse from the queries run on the actual data.

- Query interfaces can be made freely available on the public internet, or they can be made available in a restricted manner to qualified researchers operating in secure locations.

4.7 Validating a de-identified dataset

Agencies should validate datasets after they are de-identified to assure that the resulting dataset meets the agency's goals in terms of both privacy protection and data usefulness.

4.7.1 Validating privacy protection with a Motivated Intruder Test

Several approaches exist for validating the privacy protection provided by de-identification, including:

- Examining the resulting data files to make sure that no identifying information is included in file data or metadata.

- Conducting a tiger-team analysis to see if outside individuals can perform re-identification using publicly available datasets or (if warranted) using confidential agency data.

4.7.2 Validating data usefulness

Several approaches exist for validating data usefulness. For example, the results of statistical calculations performed on both the original dataset and on the de-identified dataset can be compared to see if the de-identification resulted in significant changes that are unacceptable. Agencies can also hire tiger-teams to examine the de-identified dataset and see if it can be used for the intended purpose.

5 Requirements for De-Identification Tools

At the present time there are few tools available for de-identification. This section discusses tool categories and mentions several specific tools.

5.1 De-Identification Tool Features

A de-identification tool is a program that involved in the creation of de-identified datasets. De-identification tools might perform many functions, including:

- Detection of identifying information
- Calculation of re-identification risk
- Performing de-identification
- Mapping identifiers to pseudonyms
- Providing for the selective revelation of pseudonyms

De-identification tools may handle a variety of data modalities. For example, tools might be designed for tabular data or for multimedia. Particular tools might attempt to de-identify all data types, or might be developed for specific modalities. A potential risk of using de-identification tools is that a tool might be equipped to handle some but not all of the different modalities in a dataset. For example, a tool might de-identifying the categorical information in a table according to a de-identification standard, but might not detect or attempt to address the presence of identifying information in a text field.

5.2 Data Masking Tools

Data masking tools are programs that can perform removal or replacement of designated fields in a dataset while maintaining relationships between tables. These tools can be used to remove direct identifiers but generally cannot identify or modify quasi-identifiers in a manner consistent with a privacy policy or risk analysis.

Data masking tools were developed to allow software developers and testers access to datasets containing realistic data while providing minimal privacy protection. Absent additional controls or data manipulations, data masking tools should not be used for de-identification of datasets that are intended for public release.

6 Evaluation

Agencies performing de-identification should evaluate the algorithms that they intend to use, the software that implements the algorithms, and the data that results from the operation of the software.[105]

6.1 Evaluating Privacy Preserving Techniques

There has been decades of research in the field of statistical disclosure limitation and de-identification. As the understanding of statistical disclosure limitation and de-identification have evolved over time, agencies should not base their technical evaluation of a technique on the mere fact that the has been published in the peer reviewed literature or that the agency has a long history of using the technique and has not experienced any problems. Instead, it is necessary to evaluate proposed techniques in light of the totality of the scientific experience and with regards to current threats.

Traditional statistical disclosure limitation and de-identification techniques base their risk assessments, in part, on an expectation of what kinds of data are available to an attacker to conduct a linkage attack. Where possible, these assumptions should be documented and published along with a technique description of the privacy-preserving techniques that are used to transform datasets prior to release, so that they can be reviewed by external experts and the scientific community.

Because our understanding of privacy technology and the capabilities of privacy attacks are both rapidly evolving, techniques that have been previously established should be periodically reviewed. New vulnerabilities may be discovered in techniques that have been previously accepted. Alternatively, it may be that new techniques are developed that allow agencies to re-evaluate the tradeoffs that they have made with respect to privacy risk and data usability.

6.2 Evaluating De-Identification Software

Once techniques are evaluated and approved, agencies should assure that the techniques are faithfully executed by their chosen software. Privacy software evaluation should consider the tradeoff between data usability and privacy protection.

Privacy software evaluation should also seek to detect and minimize the chances of tool error and user error.

For example, agencies should verify:

- That the software properly implements the chosen algorithms.
- The software should take into account limitations regarding floating point representations.
- The software does not leak identifying information from source to destination.

[105] Please note that NIST is preparing a separate report on evaluating de-identification software and results.

- The software has sufficient usability that it can be operated in efficiently and without error.

Agencies may also wish to evaluate the performance of the de-identification software, such as:

- Efficiency. How long does it take to run on a dataset of a typical size?
- Scalability. How much does it slow down when moving from a dataset of N to 100N?
- Usability. Can users understand the user interface? Can users detect and correct their errors? Is the documentation sufficient?
- Repeatability. If the tool is run twice on the same dataset, are the results similar? If two different people run the tool, do they get similar results?

Ideally, software should be able to track the accumulated privacy leakage from multiple data releases.

6.3 Evaluating Data Quality

Finally, agencies should evaluate the quality of the de-identified data to verify that it is sufficient for the intended use. Approaches for evaluating the data quality include:

- Verifying that single variable statistics and two-variable correlations remain relatively unchanged.
- Verifying that statistical distributions do not incur undue bias as a result of the de-identification procedure.

7 Conclusion

Government agencies can use de-identification technology to make datasets available to researchers and the general public without compromising the privacy of people contained within the data.

Currently there are three primary models available for de-identification: agencies can make data available with traditional de-identification techniques relying on suppression of identifying information (direct identifiers) and manipulation of information that partially identifying (quasi-identifiers); agencies can create synthetic datasets; and agencies can make data available through a query interface. These models can be mixed within a single dataset, providing different kinds of access for different users or intended uses.

Privacy protection is strongest when agencies employ formal models for privacy protection such as differential privacy. At the present time there is a small but growing amount of experience within the government in using these systems. As a result, these systems may result in significant and at times unnecessary reduction in data quality when compared with traditional de-identification approaches that do not offer formal privacy guarantees.

Agencies that seek to use de-identification to transform privacy sensitive datasets into dataset that can be publicly released should take care to establish appropriate governance structures to support de-identification, data release, and post-release monitoring. Such structures will typically include a Disclosure Review Board as well as appropriate education, training, and research efforts.

Appendix A References

A.1 Standards

- ASTM E1869-04(2014) Standard Guide for Confidentiality, Privacy, Access, and Data Security Principles for Health Information Including Electronic Health Records
- ISO/IEC 27000:2014 Information technology -- Security techniques -- Information security management systems -- Overview and vocabulary
- ISO/IEC 24760-1:2011 Information technology -- Security techniques -- A framework for identity management -- Part 1: Terminology and concepts
- ISO/TS 25237:2008(E) Health Informatics — Pseudonymization. ISO, Geneva, Switzerland. 2008.
- ISO/IEC 20889 WORKING DRAFT 2016-05-30, Information technology – Security techniques – Privacy enhancing data de-identification techniques. 2016.

A.2 US Government Publications

- Census Confidentiality and Privacy: 1790-2002, US Census Bureau, 2003. https://www.census.gov/prod/2003pubs/conmono2.pdf
- Disclosure Avoidance Techniques at the US Census Bureau: Current Practices and Research, Research Report Series (Disclosure Avoidance #2014-02), Amy Lauger, Billy Wisniewski, and Laura McKenna, Center for Disclosure Avoidance Research, US Census. Bureau, September 26, 2014. https://www.census.gov/srd/CDAR/cdar2014-02_Discl_Avoid_Techniques.pdf
- Privacy and Confidentiality Research and the US Census Bureau, Recommendations Based on a Review of the Literature, Thomas S. Mayer, Statistical Research Division, US Bureau of the Census. February 7, 2002. https://www.census.gov/srd/papers/pdf/rsm2002-01.pdf
- Frequently Asked Questions—Disclosure Avoidance, Privacy Technical Assistance Center, US Department of Education. October 2012 (revised July 2015) http://ptac.ed.gov/sites/default/files/FAQ_Disclosure_Avoidance.pdf
- *Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule*, U.S. Department of Health & Human Services, Office for Civil Rights, November 26, 2012. http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/hhs_deid_guidance.pdf
- *OHRP-Guidance on Research Involving Private Information or Biological Specimens (2008)*, Department of Health & Human Services, Office of Human Research Protections (OHRP), August 16, 2008. http://www.hhs.gov/ohrp/policy/cdebiol.html
- *Data De-identification: An Overview of Basic Terms*, Privacy Technical Assistance Center, U.S. Department of Education. May 2013. http://ptac.ed.gov/sites/default/files/data_deidentification_terms.pdf

- *Statistical Policy Working Paper 22 (Second version, 2005),* Report on Statistical Disclosure Limitation Methodology, Federal Committee on Statistical Methodology, December 2005.
- The Data Disclosure Decision, Department of Education (ED) Disclosure Review Board (DRB), A Product of the Federal CIO Council Innovation Committee. Version 1.0, 2015. http://go.usa.gov/xr68F
- National Center for Health Statistics Policy on Micro-data Dissemination, Centers for Disease Control, July 2002. https://www.cdc.gov/nchs/data/nchs_microdata_release_policy_4-02a.pdf
- National Center for Health Statistics Data Release and Access Policy for Micro-data and Compressed Vital Statistics File, Centers for Disease Control, April 26, 2011. http://www.cdc.gov/nchs/nvss/dvs_data_release.htm

A.3 Publications by Other Governments

- *Privacy business resource 4: De-identification of data and information*, Office of the Australian Information Commissioner, Australian Government, April 2014. http://www.oaic.gov.au/images/documents/privacy/privacy-resources/privacy-business-resources/privacy_business_resource_4.pdf

- *Opinion 05/2014 on Anonymisation Techniques*, Article 29 Data Protection Working Party, 0829/14/EN WP216, Adopted on 10 April 2014

- *Anonymisation: Managing data protection risk, Code of Practice 2012,* Information Commissioner's Office. https://ico.org.uk/media/for-organisations/documents/1061/anonymisation-code.pdf. 108 pages

- *The Anonymisation Decision-Making Framework*, Mark Elliot, Elaine Mackey, Kieron O'Hara and Caroline Tudor, UKAN, University of Manchester, July 2016. http://ukanon.net/ukan-resources/ukan-decision-making-framework/

A.4 Reports and Books:

- Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics (1993), George T. Duncan, Thomas B. Jabine, and Virginia A. de Wolf, Editors; Panel on Confidentiality and Data Access; *Commission on Behavioral and Social Sciences and Education*; *Division of Behavioral and Social Sciences and Education*; National Research Council, 1993. http://dx.doi.org/10.17226/2122

- Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk, Committee on Strategies for Responsible Sharing of Clinical Trial Data, Board on Health Sciences Policy, Institute of Medicine of the National Academies, The National Academies Press, Washington, DC. 2015.

- P. Doyle and J. Lane, Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies, North-Holland Publishing, Dec 31, 2001

- George T. Duncan, Mark Elliot, Juan-José Salazar-Gonzalez, Statistical Confidentiality: Principles and Practice, Springer, 2011

- Emam, Khaled El and Luk Arbuckle, Anonymizing Health Data, O'Reilly, Cambridge, MA. 2013

- Cynthia Dwork and Aaron Roth, The Algorithmic Foundations of Differential Privacy (Foundations and Trends in Theoretical Computer Science). Now Publishers, August 11, 2014. http://www.cis.upenn.edu/~aaroth/privacybook.html

A.5 How-To Articles

- Olivia Angiuli, Joe Blitstein, and Jim Waldo, How to De-Identify Your Data, Communications of the ACM, December 2015.

- Jörg Drechsler, Stefan Bender, Susanne Rässler, Comparing fully and partially synthetic datasets for statistical disclosure control in the German IAB Establishment Panel. 2007, United Nations, Economic Commission for Europe. Working paper, 11, New York, 8 p. http://fdz.iab.de/342/section.aspx/Publikation/k080530j05

- Ebaa Fayyoumi and B. John Oommen, A survey on statistical disclosure control and micro-aggregation techniques for secure statistical databases. 2010, *Software Practice and Experience*. 40, 12 (November 2010), 1161-1188. DOI=10.1002/spe.v40:12 http://dx.doi.org/10.1002/spe.v40:12

- Jingchen Hu, Jerome P. Reiter, and Quanli Wang, Disclosure Risk Evaluation for Fully Synthetic Categorical Data, Privacy in Statistical Databases, pp. 185-199, 2014. http://link.springer.com/chapter/10.1007%2F978-3-319-11257-2_15

- Matthias Templ, Bernhard Meindl, Alexander Kowarik and Shuang Chen, Introduction to Statistical Disclosure Control (SDC), IHSN Working Paper No. 007, International Household Survey Network, August 2014. http://www.ihsn.org/home/sites/default/files/resources/ihsn-working-paper-007-Oct27.pdf

- Natalie Shlomo, Statistical Disclosure Control Methods for Census Frequency Tables, International Statistical Review (2007), 75, 2, 199-217. https://www.jstor.org/stable/41508461

Appendix B Glossary

Selected terms used in the publication are defined below. Where noted, the definition is sourced to another publication.

attribute: "inherent characteristic." (ISO 9241-302:2008)

attribute disclosure: re-identification event in which an entity learns confidential information about a data principal, without necessarily identifying the data principal (ISO/IEC 20889 WORKING DRAFT 2 2016-05-27)

anonymity: "condition in identification whereby an entity can be recognized as distinct, without sufficient identity information to establish a link to a known identity" (ISO/IEC 24760-1:2011)

attacker: person seeking to exploit potential vulnerabilities of a system

attribute: "characteristic or property of an entity that can be used to describe its state, appearance, or other aspect" (ISO/IEC 24760-1:2011)[106]

brute force attack: in cryptography, an attack that involves trying all possible combinations to find a match

coded: "1. identifying information (such as name or social security number) that would enable the investigator to readily ascertain the identity of the individual to whom the private information or specimens pertain has been replaced with a number, letter, symbol, or combination thereof (i.e., the code); and 2. a key to decipher the code exists, enabling linkage of the identifying information to the private information or specimens."[107]

control: "measure that is modifying risk. Note: controls include any process, policy, device, practice, or other actions which modify risk." (ISO/IEC 27000:2014)

covered entity: under HIPAA, a health plan, a health care clearinghouse, or a health care provider that electronically transmits protected health information (HIPAA Privacy Rule)

data subjects: "persons to whom data refer" (ISO/TS 25237:2008)

data use agreement: executed agreement between a data provider and a data recipient that specifies the terms under which the data can be used.

data universe: All possible data within a specified domain.

dataset: collection of data

---

[106] ISO/IEC 24760-1:2011, Information technology -- Security techniques -- A framework for identity management -- Part 1: Terminology and concepts

[107] *OHRP-Guidance on Research Involving Private Information or Biological Specimens*, Department of Health & Human Services, Office of Human Research Protections (OHRP), August 16, 2008. http://www.hhs.gov/ohrp/policy/cdebiol.html

dataset with identifiers: a dataset that contains information that directly identifies individuals.

dataset without identifiers: a dataset that does not contain direct identifiers

de-identification: "general term for any process of removing the association between a set of identifying data and the data subject" (ISO/TS 25237-2008)

de-identification model: approach to the application of data de-identification techniques that enables the calculation of re-identification risk (ISO/IEC 20889 WORKING DRAFT 2 2016-05-27)

de-identification process: "general term for any process of removing the association between a set of identifying data and the data principal" [ISO/TS 25237:2008]

de-identified information: "records that have had enough PII removed or obscured such that the remaining information does not identify an individual and there is no reasonable basis to believe that the information can be used to identify an individual" (SP800-122)

direct identifying data: "data that directly identifies a single individual" (ISO/TS 25237:2008)

disclosure: "divulging of, or provision of access to, data" (ISO/TS 25237:2008)

disclosure limitation: "statistical methods [] used to hinder anyone from identifying an individual respondent or establishment by analyzing published [] data, especially by manipulating mathematical and arithmetical relationships among the data."[108]

effectiveness: "extent to which planned activities are realized and planned results achieved" (ISO/IEC 27000:2014)

entity: "item inside or outside an information and communication technology system, such as a person, an organization, a device, a subsystem, or a group of such items that has recognizably distinct existence" (ISO/IEC 24760-1:2011)

Federal Committee on Statistical Methodology (FCSM): "an interagency committee dedicated to improving the quality of Federal statistics. The FCSM was created by the Office of Management and Budget (OMB) to inform and advise OMB and the Interagency Council on Statistical Policy (ICSP) on methodological and statistical issues that affect the quality of Federal data." (fscm.sites.usa.gov)

genomic information: information based on an individual's genome, such as a sequence of DNA or the results of genetic testing

[108] Definition adapted from Census Confidentiality and Privacy: 1790-2002, US Census Bureau, 2003. https://www.census.gov/prod/2003pubs/conmono2.pdf, p. 21

harm: "any adverse effects that would be experienced by an individual (i.e., that may be socially, physically, or financially damaging) or an organization if the confidentiality of PII were breached" (SP800-122)

Health Insurance Portability and Accountability Act of 1996 (HIPAA): the primary law in the United States that governs the privacy of healthcare information

HIPAA: see *Health Insurance Portability and Accountability Act of 1996*

HIPAA Privacy Rule: "establishes national standards to protect individuals' medical records and other personal health information and applies to health plans, health care clearinghouses, and those health care providers that conduct certain health care transactions electronically" (HIPAA Privacy Rule, 45 CFR 160, 162, 164)

identification: "process of using claimed or observed attributes of an entity to single out the entity among other entities in a set of identities" (ISO/TS 25237:2008)

identified information: information that explicitly identifies an individual

identifier: "information used to claim an identity, before a potential corroboration by a corresponding authenticator" (ISO/TS 25237:2008)

imputation: "a procedure for entering a value for a specific data item where the response is missing or unusable." (OECD Glossary of Statistical Terms)

inference: "refers to the ability to deduce the identity of a person associated with a set of data through "clues" contained in that information. This analysis permits determination of the individual's identity based on a combination of facts associated with that person even though specific identifiers have been removed, like name and social security number" (ASTM E1869[109])

k-anonymity: a technique "to release person-specific data such that the ability to link to other information using the quasi-identifier is limited."[110] k-anonymity achieves this through suppression of identifiers and output perturbation.

l-diversity: a refinement to the k-anonymity approach which assures that groups of records specified by the same identifiers have sufficient diversity to prevent inferential disclosure[111]

[109] ASTM E1869-04 (Reapproved 2014), Standard Guide for Confidentiality, Privacy, Access, and Data Security Principles for Health Information Including Electronic Health Records, ASTM International.

[110] L. Sweeney. k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002; 557-570.

[111] Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. In Proc. 22nd Intnl. Conf. Data Engg. (ICDE), page 24, 2006.

masking: the process of systematically removing a field or replacing it with a value in a way that does not preserve the analytic utility of the value, such as replacing a phone number with asterisks or a randomly generated pseudonym[112]

noise: "a convenient term for a series of random disturbances borrowed through communication engineering, from the theory of sound. In communication theory noise results in the possibility of a signal sent, x, being different from the signal received, y, and the latter has a probability distribution conditional upon x. If the disturbances consist of impulses at random intervals it is sometimes known as "shot noise"." (OECD Glossary of Statistical Terms)

non-deterministic noise: a random value that cannot be predicted

personal identifier: "information with the purpose of uniquely identifying a person within a given context" (ISO/TS 25237:2008)

personal data: "any information relating to an identified or identifiable natural person (*data subject*)" (ISO/TS 25237:2008)

personally identifiable information (PII): "Any information about an individual maintained by an agency, including (1) any information that can be used to distinguish or trace an individual's identity, such as name, social security number, date and place of birth, mother's maiden name, or biometric records; and (2) any other information that is linked or linkable to an individual, such as medical, educational, financial, and employment information."[113] (SP800-122)

privacy: "freedom from intrusion into the private life or affairs of an individual when that intrusion results from undue or illegal gathering and use of data about that individual" (ISO/IEC 2382-8:1998, definition 08-01-23)

protected health information (PHI): "individually identifiable health information: (1) Except as provided in paragraph (2) of this definition, that is: (i) Transmitted by electronic media; (ii) Maintained in electronic media; or (iii) Transmitted or maintained in any other form or medium. (2) *Protected health information* excludes individually identifiable health information in: (i) Education records covered by the Family Educational Rights and Privacy Act, as amended, 20 U.S.C. 1232g; (ii) Records described at 20 U.S.C. 1232g(a)(4)(B)(iv); and (iii) Employment records held by a covered entity in its role as employer." (HIPAA Privacy Rule, 45 CFR 160.103)

pseudonymization: a particular type of de-identification that both removes the association with a data subject and adds an association between a particular set of characteristics relating to the data subject and one or more pseudonyms.[114] Typically, pseudonymization is implemented by

---

[112] El Emam, Khaled and Luk Arbuckle, Anonymizing Health Data, O'Reilly, Cambridge, MA. 2013

[113] GAO Report 08-536, Privacy: Alternatives Exist for Enhancing Protection of Personally Identifiable Information, May 2008, http://www.gao.gov/new.items/d08536.pdf

[114] Note: This definition is the same as the definition in ISO/TS 25237:2008, except that the word "anonymization" is replaced with the word "de-identification."

replacing direct identifiers with a pseudonym, such as a randomly generated value.

pseudonym: "personal identifier that is different from the normally used personal identifier." (ISO/TS 25237:2008)

quasi-identifier: information that can be used to identify an individual through association with other information

recipient: "natural or legal person, public authority, agency or any other body to whom data are disclosed" (ISO/TS 25237:2008)

re-identification: general term for any process that re-establishes the relationship between identifying data and a data subject

re-identification risk: the risk that de-identified records can be re-identified. Re-identification risk is typically reported as the percentage of records in a dataset that can be re-identified.

risk: "effect of uncertainty on objectives. Note: risk is often expressed in terms of a combination of the consequences of an event (including changes in circumstances) and the associated likelihood of occurrence." (ISO/IEC 27000:2014)

synthetic data generation: a process in which seed data are used to create artificial data that has some of the statistical characteristics as the seed data

Appendix C Specific De-Identification Tools

This appendix provides a list of de-identification tools.

NOTE

Specific products and organizations identified in this report were used in order to perform the evaluations described. In no case does such identification imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that identified are necessarily the best available for the purpose.

C.1 Tabular Data

Most de-identification tools designed for tabular data implement the k-Anonymity model. Many directly implement the HIPAA Privacy Rule's Safe Harbor standard. Tools that are currently available include:

AnonTool is a German-language program that supports the k-anonymity framework. http://www.tmf-ev.de/Themen/Projekte/V08601_AnonTool.aspx

ARX is an open source data de-identification tool written in Java that implements a variety of academic de-identification models, including k-anonymity, Population uniqueness,[115] k-Map, Strict-average risk, ℓ-Diversity,[116] t-Closeness,[117] δ-Disclosure privacy,[118] and δ-presence. http://arx.deidentifier.org/

Cornell Anonymization Toolkit is an interactive tool that was developed by the Computer Science Department at Cornell University[119] for performing de-identification. It can perform data generalization, risk analysis, utility evaluation, sensitive record manipulation, and visualization functions. https://sourceforge.net/projects/anony-toolkit/

Open Anonymizer implements the k-anonymity framework. https://sourceforge.net/projects/openanonymizer/

Privacy Analytics Eclipse is a comprehensive de-identification platform that can de-identify multiple linked tabular datasets to HIPAA or other de-identification standards. The program runs

[115] Fida Kamal Dankar, Khaled El Emam, Angelica Neisa and Tyson Roffey, Estimating the re-identification risk of clinical datasets, BMC Medical Informatics and Decision Making, 2012 12:66. DOI: 10.1186/1472-6947-12-66

[116] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. 2007. *L*-diversity: Privacy beyond *k*-anonymity. *ACM Trans. Knowl. Discov. Data* 1, 1, Article 3 (March 2007). DOI=http://dx.doi.org/10.1145/1217299.1217302

[117] N. Li, T. Li and S. Venkatasubramanian, "t-Closeness: Privacy Beyond k-Anonymity and l-Diversity," *2007 IEEE 23rd International Conference on Data Engineering*, Istanbul, 2007, pp. 106-115. doi: 10.1109/ICDE.2007.367856

[118] Mehmet Ercan Nergiz, Maurizio Atzori, and Chris Clifton. 2007. Hiding the presence of individuals from shared databases. In *Proceedings of the 2007 ACM SIGMOD international conference on Management of data* (SIGMOD '07). ACM, New York, NY, USA, 665-676. DOI=http://dx.doi.org/10.1145/1247480.1247554

[119] X. Xiao, G. Wang, and J. Gehrke. Interactive anonymization of sensitive data. In SIGMOD Conference, pages 1051–1054, 2009.

on Apache SPARK to allow de-identification of massive datasets, such as those arising in medical research. http://www.privacy-analytics.com/software/privacy-analytics-core/

µ-ARGUS was developed by Statistics Netherlands for microdata release. The program was originally written in Visual Basic and was rewritten into C/C++ for an Open Source release. The program runs on Windows and Linux. http://neon.vb.cbs.nl/casc/mu.htm

sdcMicro is a package for the popular open source R statistical platform that implements a variety of statistical disclosure controls. A full tutorial is available, as are prebuilt binaries for Windows and OS X. https://cran.r-project.org/web/packages/sdcMicro/

SECRETA, a tool for evaluating and comparing anonymizations. According to the website, "SECRETA supports Incognito, Cluster, Top-down, and Full subtree bottom-up algorithms for datasets with relational attributes, and COAT, PCTA, Apriori, LRA and VPA algorithms for datasets with transaction attributes. Additionally, it supports the RMERGEr, TMERGEr, and RTMERGEr bounding methods, which enable the anonymization of RT-datasets by combining two algorithms, each designed for a different attribute type (e.g., Incognito for relational attributes and COAT for transaction attributes)." http://users.uop.gr/~poulis/SECRETA/

UTD Anonymization Toolbox is an open source tool developed by the University of Texas Dallas Data Security and Privacy Lab using funding provided by the National Institutes of Health, the National Science Foundation, and the Air Force Office of Scientific Research.

C.2 Free Text

BoB, a best-of-breed automated text de-identification system for VHA clinical documents,[120] developed by the Meystre Lab at the University of Utah School of Medicine. http://meystrelab.org/automated-ehr-text-de-identification/

MITRE Identification Scrubber Toolkit (MIST) is an open source tool for de-identifying free format text. http://mist-deid.sourceforge.net

Privacy Analytics Lexicon performs automated de-identification of unstructured data (text). http://www.privacy-analytics.com/software/privacy-analytics-lexicon/

C.3 Multimedia

DicomCleaner is an open source tool that removes identifying information from medical imagery in the DICOM format. DicomCleaner. The program can remove both metadata from the DICOM file and black out identifying information that has been "burned in" to the image area. DicomCleaner can perform redaction directly of compressed JPEG blocks so that the medical image does not need to be decompressed and re-compressed, a procedure that can introduce artifacts. http://www.dclunie.com/pixelmed/software/webstart/DicomCleanerUsage.html

[120] BoB, a best-of-breed automated text de-identification system for VHA clinical documents. Ferrández O, South BR, Shen S, Friedlin FJ, Samore MH, Meystre SM. J Am Med Inform Assoc. 2013 Jan 1;20(1):77-83. doi: 10.1136/amiajnl-2012-001020. Epub 2012 Sep 4.