NIST Special Publication 800-188 (2nd DRAFT)

De-Identifying Government Datasets

Simson L. Garfinkel

INFORMATION SECURITY

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

NIST Special Publication 800-188 (2nd DRAFT)

De-Identifying Government Datasets

Simson L. Garfinkel
Information Access Division
Information Technology Laboratory

December 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, 85 pages (December 2016)
CODEN: NSPUE2

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

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

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

Public comment period: December 15, 2016 through December 31, 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 is a process that is applied to a dataset to reduce the risk of linking information revealed in the dataset to 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. 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, or providing a query interface that incorporates de-identification of the identified data. Agencies can create 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 privacy models. People performing de-identification generally use special-purpose software tools to perform the data manipulation and calculate the likely risk of re-identification. However, not all tools that merely 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 Luk Arbuckle, Andrew Baker, Daniel Barth-Jones, Khaled El Emam, Robert Gellman, Tom Krenzke, Bradley Malin and John Moehrke for providing comments on previous drafts and valuable insight in creating this publication.

The author also wishes to thank several organizations that provided useful comments on previous drafts of this publication, the Defense Contract Management Agency; Integrating the Healthcare Enterprise (IHE), an ANSI-accredited standards organization focusing on healthcare standards; and the Privacy Tools project at Harvard University (including Stephen Chong, Kobbi Nissim, David O'Brien, Salil Vandhan and Alexandra Wood).

Audience

This document is intended for use by government engineers, data scientists, privacy officers, disclosure 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 Figures

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 research was the development of the k-anonymity model for protecting privacy,[6] which is

---

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.

6 Latanya Sweeney. 2002. *k*-anonymity: a model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.* 10, 5

reflected in the Health and Human Services guidance regarding the HIPAA Privacy Rule.[7]

- In the 2000s, researchers in computer science who were attempting to formalize the security guarantees of cryptographic protocols developed the theory of *differential privacy*,[8] which is based on a mathematical definition of the privacy loss to an individual resulting from queries on a dataset containing that individual's personal information. Starting with this definition, researchers have developed a variety of mechanisms for minimizing the amount privacy loss associated with statistical releases.

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*.[9]

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. For example, improperly de-identified data might be used to identify vulnerable individuals or groups. The release of potentially harmful information might result in reputational risk to an agency, potentially threatening its mission.

One way that agencies can manage this risk is by creating a formal Disclosure Review Board (DRB) consisting of legal and technical privacy experts as well as stakeholders within the

---

(October 2002), 557-570. DOI=http://dx.doi.org/10.1142/S0218488502001648

[7] *Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule*, Office of Civil Rights, Health and Human Services, November 26, 2012. p. 20. http://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/

[8] 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

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

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.

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

- **The Release and Forget model:**[10] 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. Under this model, the information that is present in the released data can be tailored to the specific needs, capabilities, and risk profile of the intended data users.
- **The Synthetic Data with Verification Model:** Statistical Disclosure Limitation techniques are applied to the original dataset and used to create a synthetic dataset that reflects the statistical properties of the original dataset, but which does not contain disclosing information. The synthetic dataset is released, either publicly or to vetted researchers.
- **The Enclave model:**[11,12] The de-identified data may be kept in a physical or virtual 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. Enclaves can also support features for audit and accountability.

Agencies may also choose to apply a tiered access approach that combines several of these models to address a variety of use cases and privacy threats. For example, an agency may determine it is appropriate to release a synthetic dataset to the public, while also making a second, restricted dataset that has had limited de-identification available to qualified researchers. This limited dataset might be minimally processed, such as replacing direct identifiers with pseudonyms, to allow for longitudinal analysis, better data quality, and the possibility for controlled re-identification as required by policy. This restrict dataset might be placed in an enclave for which specific uses could be assed and carried out under observation. Results derived from this second, controlled dataset might receive additional review by a Data Release Board prior to those results being allowed to leave the enclave and be distributed to a broader audience.

Agencies can create or adopt standards to guide those performing de-identification. The standards can specify 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

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

[11] Ibid.

[12] 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

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.

Agencies can also create or adopt standards regarding the quality and accuracy of de-identified data. If data accuracy cannot be well maintained along with data privacy goals, then the release of data that is inaccurate for statistical analyses could potentially result in incorrect scientific conclusions and incorrect policy decisions.

De-identification should be performed by trained individuals using software specifically designed for the purpose. Features required of this software includes detection of identifying information; calculation of re-identification probabilities; removing identifiers or mapping identifiers to pseudonyms; manipulation of quasi-identifiers; determining whether the remaining sensitive values might themselves be identifying; and providing for the selective revelation of pseudonyms.

Although it is possible to perform de-identification with off-the-shelf software such as commercial spreadsheet or financial planning program, these programs are not designed for de-identification and encourage the use of complicated de-identification methods such as deleting sensitive columns and manually searching and removing data that appears to be sensitive. While this may result in a dataset that appears to be de-identified, significant risk of disclosure may remain.

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.

Finally, different countries have different standards and policies regarding the definition and use of de-identified data. Information that is regarded as de-identified in one jurisdiction may be regarded as being identifiable in another.

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."[13]

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, thus, are only made available on a limited basis. Some datasets are so sensitive that they cannot be made publicly available at all, but can be made available on a limited basis in protected enclaves. In some cases agencies may choose to release summary statistics, or create synthetic datasets that resemble the original data but which have less[14] disclosure risk.

Government programs collect information from individuals and organization for taxation, public benefits, public health, licensing, employment, census, and the production of official statistics. And while privacy is integral to our society, data providers (individuals and organizations) typically do not have the right to opt-out of the government information requests. This can create a conflict between the conflicting goals of privacy and public benefit.

In the case of official statistical programs, this conflict is resolved by an official promise of confidentiality to individuals and organizations when they provide information to the government. [15] A bedrock principle of official statistical programs is thus that data provided to the government should generally remain confidential and not used in a way that would harm the individual or the organization providing the data. One justification for this principle is that it required for to ensure high data quality—if data providers did not feel that the information they provide would remain confidential, they might not be willing to provide information that is accurate.

Many laws, regulations and policies that govern the release of statistics and data to the public enshrine this principle of confidentiality. 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

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

[14] John M. Abowd and Lars Vilhuber, How Protective are Synthetic Data?, Privacy in Statistical Databases, Volume 5262, Lecture Notes in Computer Science, 2008, pp. 239-246,

[15] 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.

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,[16] which recognizes that disclosure of records for statistical purposes is acceptable if the data is not "individually identifiable."[17]

- Title V – Confidential Information Protection and Statistical Efficiency Act of 2002 of the E-Government Act of 2002 (CIPSEA),[18] 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."[19] The Act further 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 in substantial harm, embarrassment, inconvenience, or unfairness to any individual on whom information is maintained."[20]

- 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.[21,22] This was followed on December 8, 2009, by the Open Government Directive,[23] 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

---

[16] Public Law 93-579, 88 Stat. 1896, 5 U.S.C. § 552a.

[17] 5 USC 552a(b)(5)

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

[19] Public Law 107-347 § 512 (b)(1), Dec. 17, 2002

[20] See Title V—Confidentiality Information Protection and Statistical Efficiency, Public Law 107-347, Dec 17, 2002.

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

[22] 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

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

> 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."[24]

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.[25] Guidance from the US Department of Health and Human Services (HHS) on the HIPAA de-identification standard notes that "[b]oth methods [the safe harbor and expert determination methods for de-identification], even when properly applied, yield de-identified data that retains some risk of identification. Although the risk is very small, it is not zero, and there is a possibility that de-identified data could be linked back to the identity of the patient to which it corresponds."[26]

De-identification is a term used within the US Government to describe the removal, modification, or obfuscation of personal information from data that are collected, used, archived, and shared, with the goal of preventing or limiting informational risks to individuals, protected groups, and establishments.[27] 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 of the de-identified dataset decreases as well.

Data quality of de-identified data refers to the degree to which inferences drawn on the de-identified data will be consistent with inferences drawn on the original data. Data quality is defined as TK (ISO DEFINITION).

---

[24] 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.

[25] 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.

[26] U.S. Dep't of Health & Human Servs., Guidance Regarding Methods for De-Identification of Protected Health Information in Accordance with the Health Information Portability and Accountability Act (HIPAA) Privacy Rule 6 (2012).

[27] 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.

Utility is traditionally in economics defined as "the satisfaction derived from consumption of a good or service."[28] *Data utility* therefore refers to the value that data users can derive from data in general.

Thus, *data quality* refers to an abstract characteristic of the data, as determined by a specific, measurable statistics, whereas data utility refers to the benefit derived from the application of the data to a specific use. Although there has previously been a tendency within the official statistical organizations to conflate these two terms, it is important to keep them distinct, because they are not necessary correlated: Data may be low quality because they contain inaccuracies or substantial noise, yet they may nevertheless have high value, and thus have high utility. Likewise, data that are very close to the reality of the thing being measured may have high quality, but the may be fundamentally worthless, and thus have low utility.

In general, data quality decreases as more aggressive de-identification techniques are employed. Therefore, any effort involving the release of data that contains personal information typically involves making some kind of tradeoff between identifiability and data quality. However, increased privacy protections do not necessarily result in decreased data utility.

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, what other data may already be publicly or privately available, and the risk of re-identification. Agencies should aim to make an informed decision by systematically weighing the risks against the benefits and choosing de-identification techniques and data release models that are tailored to their analysis of the risks and benefits. In addition, when telling individuals their information will be de-identified, agencies should also disclose that privacy risks may remain despite de-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-related risks, the privacy protecting controls, the appropriateness of various privacy controls given the risks and intended uses, and the ways that those controls could fail.

De-identification can have significant cost, including 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

[28] OECD Glossary of Statistical Terms, https://stats.oecd.org/glossary/detail.asp?ID=4884, August 13, 2002.

techniques for US government datasets. It also provides a framework that can be adapted by Federal agencies to frame the governance of de-identification processes. 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 information derived from personal data remains in a de-identified dataset, there exists the possibility that the de-identified data might reveal attributes related to specific individuals, or even that specific de-identified records could be linked back to specific individuals. When this happens, the privacy protection provided by de-identification is compromised. 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. Thus, 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.

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.[29]

2.1 Historical Context

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

- For four decades, official statistical agencies have researched and investigated methods broadly termed *Statistical Disclosure Limitation* (SDL) or *Statistical Disclosure Control*[30,31,32] 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.

[29] E.g., U.S. Dep't of State v. Washington Post Co., 456 U.S. 595 (1982); U.S. Dep't of Justice v. Reporters Comm. for Freedom of the Press, 489 U.S. 749 (1989); U.S. Dep't of State v. Ray, 502 U.S. 164 (1991).

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

[31] 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/

[32] George T. Duncan, Mark Elliot, Gonzalez Juan Jose Salazar. *Statistical Confidentiality: Principles and Practice*; Springer Science 2011.

- 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 (called *quasi-identifiers* or *indirect identifiers*[33]*).* The result of this research was the identification of the k-anonymity model for protecting privacy,[34,35,36,37] which is reflected in the HIPAA Privacy Rule. Software that measures privacy risk using k-anonymity is often 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, research in theoretical computer science and cryptography developed the theory of *differential privacy,*[38] 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 model for privacy protection* because its definitions of privacy and privacy loss are based on mathematical proofs.[39] Note that this doesn't mean that algorithms implementing differential privacy cannot result in increased privacy risk. Instead, it means that the amount of privacy risk that results from the use of these algorithms can be mathematically bounded. These mathematical limits on privacy risk have created considerable interest in differential privacy in academia, commerce and business, but to date only a few systems employing differential privacy have been operationally deployed.

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

[33] Dalenius, Finding a Needle in a Haystack, or Identifying Anonymous Census Records, Journal of Official Statistics 2:3, 329-336, 1986.

[34] Pierangela Samarati and Latanya Sweeney, Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression, *Proceedings of the IEEE Symposium on Research in Security and Privacy* (S&P). May 1998, Oakland, CA

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

[36] Samarti, P. Protecting Respondents' Identities in Microdata Release, *IEEE Transactions on Knowledge and Data Engineering*, Volume 13, Issue 6, Nov. 2001, pp. 1010-1027.

[37] 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

[38] 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

[39] 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

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.[40] A year later the Office of Management and Budget issued Memorandum M-07-16[41] 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)*.[42] 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 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 while protecting 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.

Many of the techniques that are discussed in this publication (e.g. fully synthetic data and differential privacy) have limited use at the present time within the federal government due to cost, time constraints, and the sophistication required of practitioners. However, these tehcniques

[40] 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

[41] 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

[42] 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

are likely to see increased use as agencies seek to make available datasets based on administrative data that include identifying information.

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."[43] In this document we expand the definition of de-identification to include all techniques that provide researchers with access to microdata while simultaneously limiting the opportunity for disclosure. 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. Re-identification is not the only mode of failure of de-identification techniques, as information about individuals can be inferred from their data, even without restoring an association between a data subject and the de-identified data.

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, such as differential privacy. Redaction may also reduce 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." [44] 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.[45] However, the term anonymization conveys the perception that the de-identified data *cannot* be re-identified. Absent formal methods for privacy protection, it is not possible to place mathematical bounds on the amount of privacy loss that might result from the release of de-identified data. This is because techniques such as k-anonymity and traditional Statistical Disclosure Limitation based their estimates of re-identification risk on availability or lack of information that could be used to link to the de-identified dataset. Therefore, the word *anonymization* should be avoided, as it

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

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

[45] "Anonymization is a step after de-identification that involves destroying all links between the de-identified datasets and the original datasets. The key code that was used to generate the new identification code number from the original is irreversibly destroyed (ie, destroying the link between the two code numbers." TransCelerate Biopharma, Inc., *Data De-identification and Anonymization of Individual Patient Data in Clinical Studies—A Model Approach," 2013.*

makes a promise that cannot be mathematically supported.

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. So, where other references[46] might use the term *anonymized file* to describe a dataset that has been de-identified, this publication uses the terms *de-identified file* and *de-identified dataset*, in recognition that the term *de-identified* is descriptive while the term *anonymized* is aspirational.

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.”[47] The term *coded* is frequently used in the healthcare setting to describe data that has been pseudonymized. Pseudonymization is commonly used so that multiple observations of an individual over time can be matched, and so that an individual can be re-identified if there is a policy reason to do so. Although re-identification is typically done by consulting the key, which may be highly protected, the existence of the pseudonym identifiers frequently increases the risk of re-identification through other means.

Many government documents use the phrases *personally identifiable information* (PII) and *personal information*.[48,49] 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 “information that can be used to distinguish or trace an individual's identity, such as their name, social security number, biometric records, etc. alone, or when combined with other personal or identifying information which is linked or linkable to a specific individual, such as date and place of birth, mother’s maiden name, etc.”[50] Therefore, identifying information is personal information, but personal information is not necessarily identifying information. *Non-public personal information* is used to describe information that is in a dataset that is not publicly

[46] For example, see Balaji Raghunathan, *The Complete Book of Data Anonymization: From Planning to Implementation*, CRC Press, May 2013.

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

[48] OMB Memorandum M-07-16, Safeguarding Against and Responding to the Breach of Personally Identifiable Information, Clay Johnson III, Deputy Director for Management, May 22, 2007.

[49] NIST 800-188, Guide to Protecting the Confidentiality of Personally Identifiable Information (PII), Erika McCallister, Time Grance, Karen Scarfone, April 2010. http://csrc.nist.gov/publications/nistpubs/800-122/sp800-122.pdf

[50] OMB M-07-16

available. Non-public personal 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*, *aggregate statistics* such as summary tables, or a *synthetic dataset*, in which the data are 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 responses that do not leak identifying information. De-identified results may be corrected or updated and re-released on a periodic basis. The accumulated leakage of information from multiple releases may be significant, even if the leakage from a single release is small. Issues arising from multiple releases are discussed in §3.4, "Data Release Models."

Disclosure is generally the exposure of data beyond the original collection use-case. However, when the goal of de-identification is to protect privacy, disclosure "relates to inappropriate attribution of information to a data subject, whether an individual or an organization. Disclosure occurs when a specific individual can be associated with a corresponding record(s) in the released data set (*identity disclosure*), when an attribute described in a dataset is held by a specific individual, even if the record(s) associated with that individual is (are) not identified (*attribute disclosure*), or when it is possible to make an inference about an individual, even if the individual was not in the dataset prior to de-identification (*inferential disclosure*)."[51] For more information about disclosure, please see Section 3.2.1, "Probability of Re-Identification."

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 process introduces inaccuracy into the results.[52] A primary goal of disclosure limitation is to protect the privacy of individuals while avoiding the introduction of *non-ignorable biases*[53] (for example, bias that might lead a social scientist to come to the wrong conclusion) into the de-identified dataset. One way to measure the amount of bias that has been introduced is to compare statistics or models generated by analyzing the original dataset with those that are generated by analyzing the de-identified datasets.

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

[51] 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/

[52] For example, see 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] 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/

K-anonymity[54] 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 values on the quasi-identifiers. (A quasi-identifier is a variable that can be used to identify an individual through association with other information.) A dataset is said to be *k-anonymous* if, for every specific combination of quasi-identifiers, there are no fewer than *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*[55]) and requiring that the resulting data are statistically close to the original data (known as *t-closeness*).[56]

Differential privacy[57] 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 both identity and attribute disclosure by adding non-deterministic noise (usually small random values) to the results of mathematical operations before the results are reported.[58] Unlike k-anonymity and other de-identification frameworks, differential privacy is based on information theory and makes no distinction between what is private data and what is not. Differential privacy does not require that values be classified as direct identifiers, quasi-identifiers, and non-identifying values. Instead, differential privacy assumes that *all values* in a record might be identifying and therefore all must potentially be manipulated.

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

[54] 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

[55] 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.

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

[57] 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 Foundations of Differential Privacy, in Foundations and Trends in Theoretical Computer Science Vol. 9, Nos. 3–4 (2014) 211–407, https://www.cis.upenn.edu/~aaroth/Papers/privacybook.pdf; http://dx.doi.org/10.1007/11787006_1

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

and those in the population from which the sample is drawn who are not present in the dataset. The research literature describes differential privacy being used to solve a variety of tasks including statistical analysis, machine learning, and data sanitization.[59] 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.

Note that the use of differentially private algorithms does not guarantee that privacy will be preserved. Instead, the algorithms guarantee that the amount of privacy risk introduced by data processing or data release will reside within specific mathematical bounds. It is also important to remember that the impact on privacy risk is limited to

When data releases containing information about the same individual accumulate, then privacy loss accumulates. Organizations should keep this in mind and try to assess the overall accumulated risk, and differential privacy can be used to help them make this assessment.

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. 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.

Finally, privacy harms are not the only kinds of harms that can result from the release of de-identified data. Analysts working with de-identified data will often have no way of knowing how inaccurate their statistical results are due to statistical distortions introduced by the de-identification process. Thus, de-identification operations intended to shield individuals from harm could cause harm if the statistical accuracy of the data is not actively monitored and preserved, if the resulting inaccurate de-identified data are used as the basis for evidence-based policy making.

[59] 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

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.[60]

Given the conflict and the history, there may be a tendency for government agencies to either overprotect or under-protect their data. Limiting the release of data clearly limits the privacy 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 for de-identification, they should document their rationale for adopting new approaches, how successful their approaches seem to have

---

[60] NISTIR 8053 §2.4, §3.6

been over time, and inconsistencies with previous data releases.

- **Federal Awarding Agencies** are allowed under OMB Circular A-110 to require that institutions of higher education, hospitals, and other non-profit organizations receiving 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."[61] 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 contains 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 risks that might result from the data release. As part of this risk analysis, agencies should specifically evaluate the anticipated re-identification rate, 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*.[62]

Risk assessments should be based on scientific, objective factors and consider the best interests of the individuals in the dataset, the responsibilities of the agency holding the data, and the anticipated benefits to society. 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 the residual risk is justified by the goals of the data release. An agency decision making process may choose to accept or reject the risk resulting from a release of de-

[61] 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

[62] 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

identified data, but participants in the risk assessment should not be empowered to prevent risk from being documented and discussed.

At the present time it is difficult to have measures of re-identification risk that are both general and meaningful. For example, is possible to measure the similarity between individuals in the dataset under a variety of different parameters, and to model how this similarity is impacted when the larger population is considered. But such calculations may result in significantly different levels of risk for different groups. There may be some individuals in a dataset who would be significantly adversely impacted by re-identification, and for whom the likelihood of re-identification might be quite high, but these individuals might represent a tiny fraction of the entire dataset. This represents an important area for research in the field of risk communication.

3.2.1 Probability of Re-Identification

As discussed in Section 2.3, "Terminology," the potential impacts on individuals from the release and use of de-identified data include:[63]

- **Identity disclosures —** Associating a specific individual with the corresponding record(s) in the data set with high probability. 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 by a specific individual with high probability, 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.[64] In these cases, traditional de-identification does not protect against attribute disclosure, although differential privacy can.

- **Inferential disclosure —** being able to make an inference about an individual (typically a member of a group) with high probability, even if the individual was not in the dataset prior to de-identification. "Inferential disclosure is of less concern in most cases as inferences are designed to predict aggregate behavior, not individual attributes, and thus are often poor predictors of individual data values."[65] Inferential disclosure does not disclose identity, and traditional de-identification do not protect against inferential disclosure; differential privacy can only protect against it if the potential for disclosure results from the individual's presence in the dataset. Therefore, when considering inferential disclosure, it is important to distinguish between inferences about individuals that rely on the fact that the individual's data was used, and those that result from the individual's membership in a group that has been subject to data collection and analysis.

[63] 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.

[64] NISTIR 8053 §2.4, p 13.

[65] Vagelis Hristidis, *Information Discovery on Electronic Health Records*, CRC Press, Dec. 2009, p. 198. 331 pages.

Although these disclosures are commonly thought to be discrete events involving the release of specific data, such as an individual'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 probabilistic disclosure has still occurred, because the adversary's estimate of the individual has been changed by the data release.

Re-identification probability[66] is the probability that an attacker will be able to use information contained in a de-identified dataset to make identity-related inferences about individuals. Different kinds of re-identification probabilities can be calculated, including:

- *Known Inclusion Re-Identification Probability* (KIRP). The probability of finding the record that matches a specific individual known to be in the population corresponding to a specific record. KIRP can be expressed as the probability for a specific individual, or the probability averaged over the entire dataset (AKIRP).[67]

- *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 is not in the dataset. UIRP can be expressed as a probability for an individual record in the dataset averaged over the entire population (AUIRP).[68]

- *Recording matching probability* (RMP). The probability 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

[66] 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 because of possible unwanted connotations of those terms, and 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.

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

[68] 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.

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 may 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. Also, the mere fact that information is public may not negate the responsibility of an agency to provide protection for that information, as the aggregation and distribution of information may cause privacy risk that was not otherwise present.

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 (publicly available or privately held) that can be linked with the de-identified data. It is likely that the probability of re-identification increases over time as techniques improve and more contextual information becomes available to attackers.

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 that will be publicly released:

- Although a specific de-identified dataset may not be seen as sensitive, re-identifying that dataset may be an important step in re-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 public may not be skilled in re-identification in general, many resources on the Internet make it easy to acquire specialized datasets, tools, and experts for specific re-identification challenges. Also, family members, friends, colleagues, and others may possess substantial personal knowledge about individuals in the data that can be used for re-identification.

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. However, it is unreaslitic to assume that all of the world's resources will be used to attempt to re-identify every publicly released file. Therefore, it is also necessary to gauge de-identification requirements using a risk assessment. More information on conducting risk assessments can be found in NIST Special Publication 800-30, *Guide for Conducting Risk Assessments*[69].

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.

If agencies determine that the potential for harms are large in a contemplated data release, one way to manage the risk is by increasing the level of de-identification and accepting a lower data quality level. Other options include data controls, such as restricting availability of data to qualified researchers in a data enclave.

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 risk of fraud, identity theft, discrimination or abuse.
- Increased availability of an individual's location, putting that person at risk for burglary, property crime, assault, or other kinds of violence.
- Increased availability of an individual's non-public personal information, causing psychological harm by exposing potentially embarrassing information or information that the individual may not otherwise choose to reveal to the public or to family members, and potentially affecting opportunities in the economic marketplace (e.g., employment, housing, and college admission).

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

- Mandatory reporting under breach reporting laws, regulations or policies.
- Embarrassment or reputational damage if it can be publicly demonstrated that de-identified data can be re-identified.
- Harm to agency operations if some aspect of those operations required that the de-identified data remain confidential. (For example, an agency that is forced to discontinue

a scientific experiment because the data release may have biased the study participants.)

- Financial impact resulting from the harm to the individuals (e.g. 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:

- It may undermine the reputation of researchers in general and the willingness of the public to support/tolerate research and to provide accurate information to government agencies and to researchers.
- It may engender a lack of trust in government. Individuals may stop consenting to the use of their data, or even stop providing their data or provide false data.
- 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, organizations should enumerate specific measures that they will take to minimize the risk of successful re-identification. Organizations may wish to consider both the actual risk and the perceived risk on the part of those in the dataset as well as the broader community.

As part of the risk assessment, an organization may determine that there is no way to achieve the de-identification goal in terms of data quality and identifiability. In these cases, the organization will need to decide whether it should adopt additional measures to protect privacy (e.g. administrative controls or data use agreements), accept a higher level of risk, or choose not proceed with the project.

3.2.3 Impacts other than re-identification

The use of de-identified data can result in adverse impacts other than those that might result from re-identification. Risk assessments that evaluate the risks of re-identification can address these other risks as well. Such risks might include:

- The risk of inferential disclosures.
- The risk that the de-identification process might introduce bias or inaccuracies into the

dataset that result in incorrect decisions.[70]

- The risk that releasing a de-identified dataset might reveal non-public information about an agency's policies or practices.

It is preferable to use de-identification processes that provide measures of accuracy (e.g. confidence intervals) with respect to the data release. Ideally, it should be possible to reveal the de-identification process itself, so that analysts can account for potential inaccuracies. This is consistent with "Kerckhoff's principle," a widely accepted principle in cryptography that holds that the security of a system should not rely on the secrecy of the methods being used.

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

The *NIST Big Data Interoperability Framework* defines the data life cycle as "the set of processes in an application that transform raw data into actionable knowledge."[71] The data life cycle can be used to guide in the de-identification process to help analyze expected benefits and intended uses, privacy threats, and vulnerabilities of de-identified data. As such, the data life cycle concept can be used to select privacy controls that are appropriate based on a reasoned analysis of the threats. For example, privacy-by-design concepts[72] can be employed to decrease the number of identifiers collected, minimizing requirements for de-identification prior to data release. The data life cycle can also be used to design a tiered access mechanism based on this analysis.[73]

---

[70] 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. Barth-Jones, DC. *The Debate Over 'Re-Identification' Of Health Information: What Do We Risk?* Health Affairs Blog, August 10, 2012. http://healthaffairs.org/blog/2012/08/10/the-debate-over-re-identification-of-health-information-what-do-we-risk/

[71] 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

[72] Ann Cavoukian, *Privacy by Design: The 7 Foundational Principles*, Information & Privacy COmissioner, Ontario, CA. January 2011 (revised). https://www.ipc.on.ca/wp-content/uploads/Resources/7foundationalprinciples.pdf

[73] Micah Altman, Alexandra Wood, David O'Brien, Salil Vadhan, & Urs Gasser, Towards a Modern Approach to Privacy-Aware Government Data Releases, 30 Berkeley Technology Law Journal 1967 (2015), http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2779266.

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 → Deposit → Preserve → Discover → Integrate → Analyze → Collect: [74]

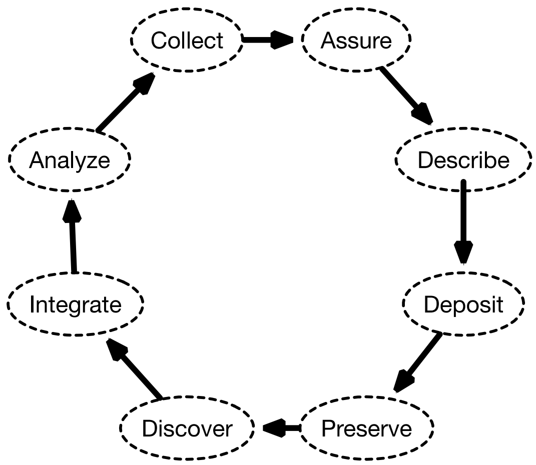

Figure 1 Michener et al.'s view of the data lifecycle is a true cycle, with analysis guiding future collection.

It is unclear how de-identification fits into a circular life cycle model, as the data owner typically retains access to the identified data. However, if the organization employs de-identification, it could be performed during the Collect, or between Collect and Assure in the event that identified data were collected but the identifying information was not actually needed. Alternatively, de-identification could be applied after Describe and prior to Deposit, to avoid archiving identifying information.

Chisholm and others 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:[75]

---

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

[75] 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

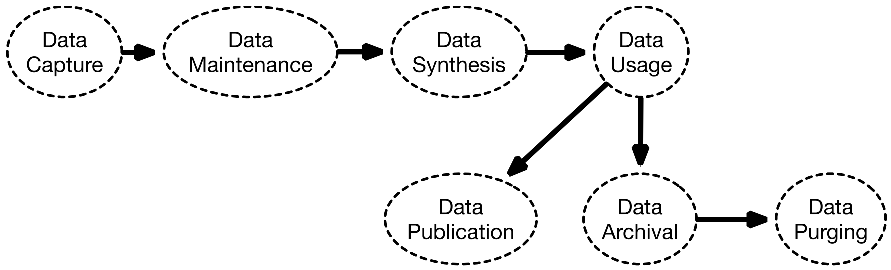

Figure 2 Chisholm's view of the data lifecycle is a linear process with a branching point after data usage.

Using this formulation, de-identification can take place either during data capture or following Data Usage. That is, identifiers that are not needed for maintenance, synthesis and usage should not be collected. If fully identified data are needed within the organization, the identifying information can be removed prior to the data being published (as a dataset), shared or archived. Indeed, applying de-identification throughout the data life cycle minimizes privacy risk and significantly eases the process of public release.

Altman et al. propose a "modern approach to privacy-aware government data releases" that incorporates progressive levels of de-identification as well different kinds of access and administrative controls in line with the sensitivity of the data.

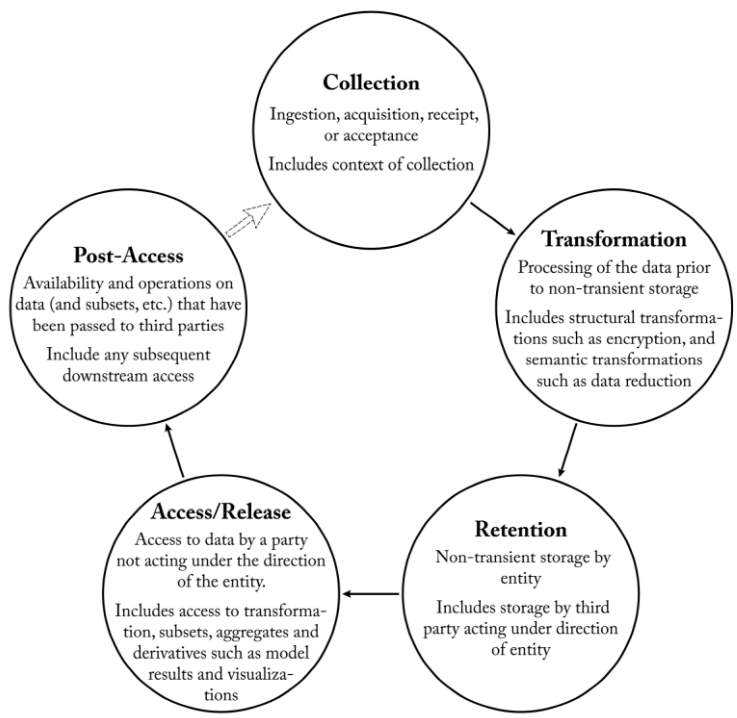

Figure 3 Lifecycle model for government data releases, from Altman et al.

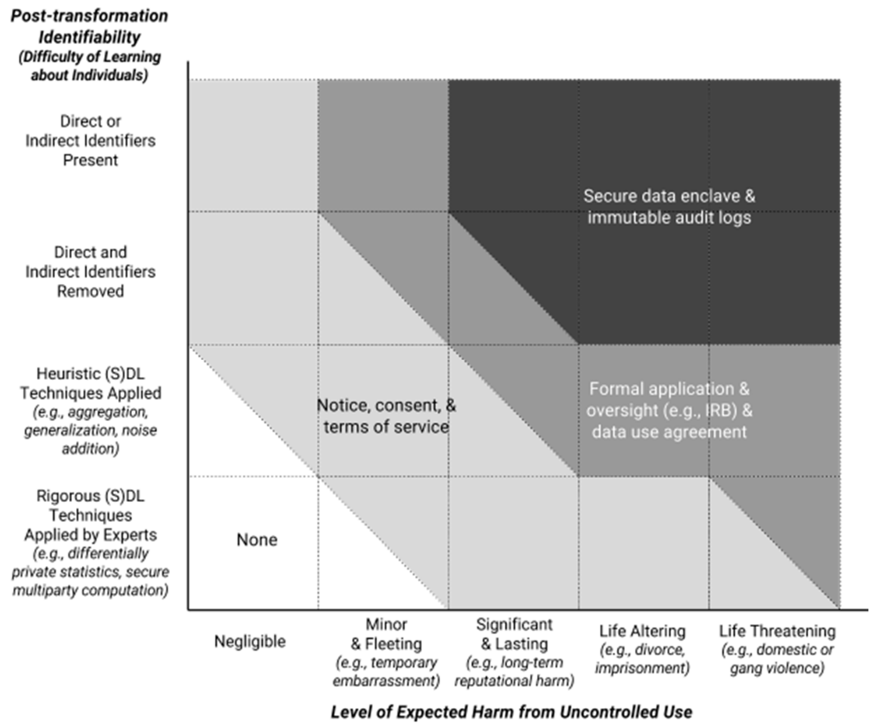

Figure 4 Conceptual diagram of the relationship between post-transformation identifiability, level of expected harm, and suitability of selected privacy controls for a data release. From Altman et al.

Agencies performing de-identification should document that:

- Techniques used to perform the de-identification are theoretically sound and generally accepted.
- 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 later?
- 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?
- Is there a 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 if 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.[76] Possible models include:

- **The Release and Forget Model:**[77] 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 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"[78]) or members of the general public (e.g. citizen scientists or the media), although they may be simply posted on the Internet with a click-through

[76] NISTIR 8053 §2.5, p. 14

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

[78] 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

license agreement that must be agreed to before the data can be downloaded (the "click-through model"[79]).

- **The Synthetic Data with Verification Model:** Statistical Disclosure Limitation techniques are applied to the original dataset and used to create a synthetic dataset that contains many of the aspects of the original dataset, but which does not contain disclosing information. The synthetic dataset is released, either publically or to vetted researchers. The synthetic dataset can then be used as a proxy for the original dataset, and if constructed well, the results of statistical analyses should be similar. If used in conjunction with an enclave model as below, researchers may use the synthetic dataset to develop queries and/or analytic software; these queries and/or software can then be taken to the enclave or provided to the agency and be applied on the original data.

- **The Enclave Model:**[80,81,82] 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. Enclaves can be physical or virtual, and can themselves operate under a variety of different models. For example, vetted researchers may travel to the enclave to perform their research, as is 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. Queries made in the enclave model may be vetted either automatically or manually (e.g., by the DRB). Vetting can try to screen for queries that might violate privacy or are inconsistent with the stated purpose of the research.

Sharing models should consider 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 (whether from the same dataset of from different datasets containing some of the same individuals) 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 a 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.[83]

---

[79] Ibid.

[80] Ibid.

[81] 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

[82] Seastrom, MM. Chapter 11, Licensing in Confidentiality, *Disclosure and Data Access: Theory and Practical Application for Statistical Agencies*. Doyle, P; Lane, JI; Theeuwes, JJM; and Zayatz, LM. (Eds) Elsevier Science, B.V. 2001

[83] See Joel Havermann, plaintiff - Appellant, v. Carolyn W. Colvin, Acting Commissioner of the Social Security Administration,

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.[84] However, releasing only a subset may decrease the statistical power of tests on the data, may cause users to draw incorrect inferences on the data if proper statistical sampling methods are not used, may obscure the ability to draw correct inferences, 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 with a research community.[85] 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 research will be conducted.

One of the positive aspects of the Five Safes framework is that it forces data controllers 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 forced to consider (or to explicitly not consider) a variety of different aspects of privacy

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.

[84] 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

[85] 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

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[86]

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."[87]

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*[88] 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 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 of privacy. Specifically, DRBs may wish to have as members:

---

[86] 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.

[87] 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

[88] NISTIR 8053 §2.4, p. 13

- Individuals representing the interests of potential users; such individuals need not come from outside of the organization.

- Representation from among the public, and specifically from groups represented in the data sets if they have a limited scope.

- Representation from the organization's leadership team. Such representation helps establish the DRB's credibility with the rest of the organization.

- A representative of the organization's senior privacy official.

- Subject matter experts.

- Outside 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.[89]

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. In addition, the DRB should require that, when individuals are informed that their information will be de-identified, that they also be informed that privacy risks may remain despite de-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,[90,91] the DRB should be empowered to reject release

[89] 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

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

[91] 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

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 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.[92]

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) that defines a level of service to which the DRB commits.

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.

[92] 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,

- 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).
- Legal and technical understanding of privacy.

Example outputs of a DRB include specifying access methods for different kinds of data releases, establishing acceptable levels of re-identification risk (e.g. values of k or $\boldsymbol{\epsilon}$), and maintaining detailed records of previous data releases—ideally including the dataset that was released and the privacy-preserving methodology that was employed.

There is some similarity between DRBs as envisioned here and the Institutional Review Board (IRBs) system created by the Common Rule[93] for regulating human subjects research in the United States. However, there are important differences:

- While the purpose of IRBs is to protect human subjects, DRBs are charged with protecting data subjects, institutions, and potentially society as a whole.
- Whereas IRBs are required to have "at least one member whose primary concerns are in nonscientific areas" and "at least one member who is not otherwise affiliated with the institution and who is not part of the immediate family of a person who is affiliated with the institution," there is no need for such members in a DRB.
- Whereas IRBs give approval for research and then typically receive reports only during an annual review or when a research project terminates, DRBs may be involved at multiple points during the process.
- Whereas approval of an IRB is required before research with human subjects can commence, DRBs are typically involved after research has taken place, prior to data or other research findings being released.
- Whereas service on an IRB requires knowledge of the Common Rule and an

[93] The Federal Policy for the Protection of Human Subjects or the "Common Rule" was published in 1991 and codified in separate regulations by 15 Federal departments and agencies. The most commonly cited reference to the Common Rule is the version in the regulations for the Department of Health and Human Services, 45 CFR part 46. The Department of Commerce references 15 CFR part 27.

understanding of ethics, service on a DRB requires knowledge of statistics, computation and public policy.

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.

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[94] 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 determine if the information is adequately de-identified."[95] 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

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

[95] *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

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. Because prescriptive de-identification standards are not changed on a case-by-case basis, there is a tendency for them to be more conservative than is necessary, resulting in the unnecessary decrease in data quality for corresponding levels of risk.

Agencies creating prescriptive de-identification standards should assure that data de-identified according to the rules have a sufficiently small risk of being re-identified that is consistent with the intended data use; such assurances frequently cannot be made unless formal privacy techniques such as *differential privacy* are employed. However, agencies may determine that public policy goals furthered by having an easy-to-use prescriptive standard outweighs the risk of a standard that does not have provable privacy guarantees.

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, especially as methodology advances and more data sources become available.

3.7.3 Performance-Based De-Identification Standards

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

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."[96] The rule does not require that experts describe the methodology used, nor does it put the expert's work under the jurisdiction of HHS.

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.

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

3.8 Education, Training and Research

De-identifying data in a manner that preserves privacy can be a complex mathematical, statistical, administrative, 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 or to adopt industry standards such as the HITrust De-Identification Framework.[97] 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.

[97] Health Information Trust Alliance, De-Identification Framework, 2016 https://hitrustalliance.net/de-identification/.

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 within the context of a target use-case. This section discusses technical options for performing de-identification and verifying the result of a de-identification procedure.

Agencies should adopt a detailed, written process for de-identifying data prior to commencing work on a de-identification project. The details of the process will depend on the particular de-identification approach that is pursued. In developing technical steps for data de-identification, agencies may wish to consider existing de-identification standards, such as the HIPAA Privacy Rule or the IHE De-Identification Handbook[98] or the HITRUST De-Identification Framework.[99]

4.1 Determine the Privacy, Data Usability, and Access Objectives

Agencies intent on de-identifying data for release should understand the nature of the data that they intended to de-identification and 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:

- Where did the data come from?
- What promises were made when the data were collected?
- What are the legal and regulatory requirements regarding privacy and release of the data?
- 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-

[98] IHE IT Infrastructure Handbook, De-Identification, Integrating the Healthcare Enterprise, June 6, 2014. http://www.ihe.net/User_Handbooks/

[99] HITRUST De-Identification Working Group (2015, March). De-Identification Framework: A Consistent, Managed Methodology for the De-Identification of Personal Data and the Sharing of Compliance and Risk Information. Frisco, TX: HITRUST. Retrieved from https://hitrustalliance.net/de-identification-license-agreement/.

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 Conducting a Data Survey

Different kinds of data require different kinds of de-identification techniques. As a result, an important early step in the de-identification of government data is to identify the data modalities that are present in the dataset and formulate a plan for de-identification that takes into account goals for data release, data quality, privacy protection, and the best available science.

For example:

- **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.* For example, if there exist quasi-identifiers that are not identified as such and not subjected to SDL, then it may be easy to re-identify records in the de-identified dataset.

  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, while others are not. Dates which refer to matters of public record (e.g., date of birth, death or home purchase) should be routinely taken as having high re-identification potential. 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.[100,101] 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.[102] Note that unstructured text may be present in tabular datasets and require special attention.[103]

- **Photos and video** may contain identifying information such as printed names (e.g. name tags), as well as metadata in the file format. There also exists a range of biometric techniques for matching photos of individuals against a dataset of photos and identifiers.[104]

- **Medical imagery** poses additional problems over photographs and video due to the presence of many 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 using a biometric algorithm and dataset.[105]

- **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

---

[100] NISTIR 8053, §4.5 p. 37

[101] The Impact of Multiple Geographies and Geographic Detail on Disclosure Risk: Interactions between Census Tract and ZIP Code Tabulation Geography" U.S. Census Bureau, 2001. https://www.census.gov/srd/sdc/steel.sperling.2001.pdf

[102] NISTIR 8053, §4.1 p. 30

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

[104] NISTIR 8053, §4.2 p. 32

[105] NISTIR 8053, §4.3 p. 35

through genealogical triangulation, a process that uses genetic information and other information as quasi-identifiers to single-out a specific identity.[106] 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.

In many cases data may be complex and contain multiple modalities. Such mixtures may complicate risk determinations.

A dataset that is thought to contain purely tabular data may be found, upon closer examination, to include unstructured text or even photograph data.

4.3 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 process for de-identifying data by removing identifiers and transforming quasi-identifiers:[107]

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.

---

[106] NISTIR 8053, §4.4 p. 36

[107] This process is based on a process 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**[108] 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.[109]

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. Options for performing this operation are discussed below in §4.3.1.

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, some researchers recommend a test data import run to assist in planning.[110]

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.

---

[108] 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.

[109] 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

[110] Khaled El Emam and Bradley Malin, Concepts and Methods for De-Identifying Clinical Trial Data, Appendix B, in Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk, National Academies Press, 2015.

Step 13. If the data do not pass the actual risk threshold, adjust the procedure and repeat Steps 11 and 12. For example, additional transformations may be required. Alternatively, it may be necessary to remove outliers.

4.3.1 Removing or Transformation of Direct Identifiers

There are many possible processes for removing direct identifiers from a dataset, including:

- Removal and replacement with the value used by the database to indicate a missing value, such as Null or NA.
- 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. Encryption used for this purpose carries special risks which need to be addressed with specific controls; see below for further information. Encryption is a pseudonymization technique.
- Hashing with a keyed hash, such as a Hash-based Message Authentication Code (HMAC)[111]. The hash key should 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. Hashing used for this purpose carries special risks which need to be addressed with specific controls; see below for further information.
- Replacement with keywords, such as transforming “George Washington” to “PATIENT.”
- Replacement by realistic surrogate values, such as transforming “George Washington” to “Abraham Polk.”[112] If the replacement by realistic surrogate values is consistent and surrogates are not reused, then replacement is a pseudonymization technique.

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

---

[111] H. Krawczyk, M. Bellare and R. Canetti, RFC 6151, “HMAC: Keyed-Hashing for Message Authentication,” February 1997. https://tools.ietf.org/html/rfc2104

[112] 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.

identities are *pseudonyms*. The mapping between the direct identifier and the pseudonym is performed using a lookup table or a repeatable transformation. In either case, the release of the lookup table or the information used for the repeatable transformation will result in the compromise identities. Thus, the lookup table or the information for the transformation must be highly protected. When using a lookup table, the pseudonym must be randomly assigned. A significant risk of using a repeatable transformation is that an attacker will be able to determine the transformation, and in so doing gain the capability to re-identify all of the records in the dataset.

SPECIAL SECURITY NOTE REGARDING
THE ENCRYPTION OR HASHING OF DIRECT IDENTIFIERS

The transformation of direct identifies through encryption or hashing carries special risks, as errors in procedure or the release of the encryption key can result in the compromise of identity. When information is protected with encryption, the security of the encrypted data depends entirely on the security of the encryption key. If a key is improperly chosen, it may be possible for an attacker to find it using a brute force search. Because there is no visual difference between data that are encrypted with a strong encryption key and data that are encrypted with a weak key, it is necessary for an organization to relies on encryption to assure through administrative controls that keys are used that are both unpredictable and suitably protected. The use of encryption or hashing to protect direct identifiers is not recommended unless justified by specific extenuating circumstances.

4.3.2 De-Identifying Numeric 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."[113]

- **Micro aggregation**, in which individual microdata are combined into small groups that preserve some data analysis capability while providing for some disclosure protection.[114]

- **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.

[113] HIPAA § 164.514 (b).

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

- **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.[115]

- **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,"[116] 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.[117]

These techniques (and others) are described in detail in several publications, including:

- *Statistical Policy Working Paper #2* (Second version, 2005) by the Federal Committee on Statistical Methodology.[118] 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.

---

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

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

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

[118] *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.

- *IHE IT Infrastructure Handbook, De-Identification,* Integrating the Healthcare Enterprise, June 6, 2014. http://www.ihe.net/User_Handbooks/

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.[119]

As currently practiced, statistical agencies typically do not document in detail the specific statistical disclosure technique that they use to transform quasi-identifiers when performing statistical disclosure limitation. Likewise, statistical agencies do not 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. [120] This lack of transparency can result in erroneous conclusions on the part of data users.

4.3.3 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 — a date of admission and discharge of July 4, 1776 and July 9, 1776 become Sept. 10, 1777 and Sept. 15, 1777[121]). However, this does not eliminate the

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

[120] 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/

[121] Office of Civil Rights, "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, 2010. http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/guidance.html

risk an attacker will make inferences based on the interval between dates.

- 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, whether a day is a work day or a holiday, or a relationship to a holiday or event.

- 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.

4.3.4 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.

Some geographical locations are not identifying (e.g. a crowded train station), while others may be highly identifying (e.g. a house in which a single person lives). Other positions may be identifying at some times but not at others. Single locations may be not identifying, but may become identifying if they represent locations linked to a single individual that are recorded over time.

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 consider 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.

De-identification of geographical data is especially challenging when location of individuals is recorded over time, because behavioral time-location patterns can act as fingerprints for re-identification purposes even with a small number of recorded locations per individual.[122]

[122] See Yves--Alexandre de Montjoye et al., Unique in the Shopping Mall: On the Reidentifiability of Credit Card Metadata, 347 Science 536 (2015); Yves--Alexandre de Montjoye et al., Unique in the Crowd: The Privacy Bounds of Human Mobility, 3 Nature Sci. Rep. 1376 (2013).

4.3.5 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.[123] Modern technology allows the complete specific sequence of an individual's DNA to be chemically determined, although this is rarely done in practice. With current technology, it is far more common to use a DNA microarray to probe for the presence or absence of specific DNA sequences at predetermined points in the genome. This approach is typically used to determine the presence or absence of specific single nucleotide polymorphisms (SNPs).[124] DNA sequences and SNPs are the same for identical twins, individuals resulting from divided embryos, and clones. With these exceptions, it is believed that no two humans have the same complete DNA sequence.

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 an individual. Thus, there are some sections of the DNA sequence and some combinations of SNPs that have high variability within the human population 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. The inheritability of genetic information has also allowed researchers to determine the surnames and even the complete identities of some individuals.[125]

As the number of individuals that have their DNA and SNPs measured increases, scientists are realizing that the characteristics of DNA and SNPs in individuals may be more complicated than the preceding paragraphs imply. DNA changes as individuals age because of senescence, transcription errors, and mutation. DNA methylation, which can impact the functioning of DNA, also changes over time.[126] Some individuals are made up with DNA from multiple individuals, typically the result of fusion of twins in early pregnancy; these people are known as *chimera* or *mosaic*. In 2015 a man in the US failed a paternity test because the genes in his saliva were

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

[124] 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

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

[126] Hans Bjornsson, Martin Sigurdsson, M. Daniele Fallin, et. Al, Intra-individual Change Over Time in DNA Methylation with Familial Clustering, JAMA. 2008;299(24):2877-2833. Doi:10.1001/jama.299.24.2877

different from those in his sperm.[127] A human chimera was identified in 1953 as a result of having a blood that a mixture of two blood types, A and O.[128] The incidence of human chimeras is unknown.

Because of the high variability inherent in DNA, complete DNA sequences may be identifiable. Likewise, biological samples for which DNA can be extracted may be identifiable. Subsections of an individual's DNA sequence and collections of highly variable SNPs may be identifiable unless there it is known that there are many individuals that share the region of DNA or those SNPs. Furthermore, genetic information may not only identify an individual, but may also be able to identify an individual's ancestors, siblings, and descendants.

4.3.6 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
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,

[127] Shehab Khan, 'Human chimera': Man fails paternity test because genes in his saliva are different to those in sperm, The Independent, October 24, 2015.

[128] Bowley, C. C.; Ann M. Hutchison; Joan S. Thompson; Ruth Sanger (July 11, 1953). "A human blood-group chimera." British Medical Journal: 81. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2028470/

aggregation alone is not sufficient to provide formal privacy guarantees. However, the differential privacy literature provides many methods for performing aggregation that are both formally private and highly accurate on large datasets. These methods work through the additional of carefully calibrated “random noise.”

4.3.7 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 is of sufficient high-dimensionality, and if there exists a way to link the supposedly non-identifying values with an identity.[129]

4.3.8 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 could be susceptible to being linked with another data set that contains these same values. (Note that this is different than characterizing an individual’s heartbeat pattern so that it could be used as a biometric. In this case, it is a specific sequence of heartbeats that is recognized.) Geographical location data can, when linked over time create individual behavioral time-location patterns can act as fingerprints for re-identification purposes even with a small number of recorded locations per individual.[130]

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.

[129] 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

[130] See Yves--Alexandre de Montjoye et al., Unique in the Shopping Mall: On the Reidentifiability of Credit Card Metadata, 347 Science 536 (2015); Yves--Alexandre de Montjoye et al., Unique in the Crowd: The Privacy Bounds of Human Mobility, 3 Nature Sci. Rep. 1376 (2013).

4.3.9 Challenges Posed by Composition

In computer science, the term *composition* refers to combining multiple functions together to create more complicated ones. One of the defining characteristics of complex systems is that complicated functions created by composition can have unpredictable results, even when they are composed from very simple components. The challenge of composition is to develop approaches for understanding when composition will have unpredictable results and to address those results proactively.

When de-identifying, it is important to understand if the techniques that are used will retain their privacy guarantees when they are subject to composition. For example, if the same dataset is made available through two different de-identification regimes, attention must be paid to whether the privacy guarantees will remain if the two downstream datasets are re-combined.

Composition concerns can arise when the same dataset is provided to multiple downstream users, when the dataset is published on a periodic basis, or when changes in computer technology result in new aspects of a dataset being made available. Privacy risk can result from unanticipated composition, which is one of the reasons that released datasets should be subjected to periodic review and reconsideration.

4.3.10 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 could also result in the re-identification of the de-identified dataset.

4.4 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: [131]

- 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

[131] 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

synthetic dataset. (A "fully synthetic dataset.")

In both cases, formal privacy techniques can be used to quantify the privacy protection offered by the synthetic dataset.

It is also possible to create *test data* that is syntactically valid but which does not convey accurate information when analyzed. Such data can be used for software development. When creating test data, it is useful for the names, addresses and other information in the data to be conspicuously non-natural, so that the test data is not inadvertently confused with actual data.

Other terms have been used to describe synthetic data; Table 1 presents terms that have been collected from the academic literature.

Data Adjective	**Meaning**
Fully Synthetic	Data for which there is no one-to-one mapping between any record in the original dataset and in the synthetic dataset.
Partially Synthetic	Data for which there may be one-to-one mappings between records in the original dataset and in the synthetic dataset, but for which attributes have been altered or swapped between records. This approach is sometimes called *blank-and-impute*.
Test	Data which resemble the original dataset in terms of structure and the range of values, but for which there is no attempt to assure that inferences drawn on the test data will be similar to those drawn on the original data. Test data may also include extreme values that are not in the original data, but which are present for the purpose of testing software.
Realistic	Data that have a characteristic that is similar to the original data, but which is not developed by modifying original data.

Table 1 Terms for Synthetic Data

4.4.1 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 modified.

4.4.2 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. Special efforts must be taken to maintain marginal and join probabilities when creating fully synthetic data.

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.[132]

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

[132] 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/

synthetic dataset:

- It can be very difficult or even impossible to map records to actual people.
- The privacy guarantees can potentially be mathematically established and proven (cf. the section below on "Creating a synthetic dataset with differential privacy").
- 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.

4.4.3 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.

4.4.4 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 non-public personal information about individuals. The disclosure risk may be greater when synthetic data are created with traditional data imputing techniques, rather than those based on formal privacy models such as differential privacy, as the formal models have provisions for tracking the accumulated privacy loss budget resulting from multiple data operations.

One of the advantages of synthetic data is that the privacy loss budget can be spent in creating the synthetic dataset, rather than in responding to interactive queries. The danger in using the privacy loss budget to respond to interactive queries is that each query decreases the budget. As the number of queries continues, the data controller needs to respond by increasing the amount of noise, by accepting a higher level of privacy risk, or by ceasing to answer questions. This can result in equity issues, if the first users to query the dataset are able to obtain better answers than later users.[133]

4.4.5 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.[134] 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,[135] then that function is said to provide ε-differential privacy.

In the mathematical formulation of differential privacy, 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. [136] A randomized function κ gives ε-differential privacy if for all datasets D_1

[133] "If we're going to move to a new model we may say we're going to have to limit these people from doing analysis because these people got there first. That's something we have to think about." Testimony of Cavan Capps, U.S. Department of the Census, before the Department of Health and Human Services, Subcommittee on Privacy, Confidentiality & Security, National Committee on Vital and Health Statistics, Hearing of: "De-Identification and the Health Insurance Portability and Accountability Act (HIPAA)," May 25, 2016. http://www.ncvhs.hhs.gov/transcripts-minutes/transcript-of-the-may-25-2016-ncvhs-subcommittee-on-privacy-confidentiality-security-hearing/

[134] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In Proceedings of the 3rd Theory of Cryptography Conference, pages 265–284, 2006.

[135] 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.

[136] 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.

and D_2 differing on at most one element, and all $S \subseteq \text{Range}(\kappa)$,

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

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.:

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

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 $\epsilon = \infty$.

For values of epsilon that are small, e^{ϵ} is approximately equal to $1 + \epsilon$. Intuitively, this means that small values of ϵ result in high privacy outcomes, while large values of ϵ result in low privacy outcomes.

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 results 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 value of 0 implies that the function κ provides the same distribution of answers no matter if anyone is removed or a person's attributes changed, while the value of ∞ allows the original dataset to be released without being subject to disclosure limitation.

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 ⍰. 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 undermining the privacy protections.[137]

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 non-public personal information because of the binary representation of IEEE floating point numbers used by the implementations.[138]

Since there are relatively few 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.

4.5 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, potentially satisfying a formal privacy model such as differential 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 synthetic data deviate significantly from the queries run on the actual data (taking care to ensure that the warning itself does not compromise privacy of some individual).
- 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.

Care must be taken in implementing interactive query interfaces, as it is possible to reconstruct

[137] 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.

[138] 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

private microdata from a query interface that does not incorporate sufficient noise infusion.[139] For this reason, query interfaces should also log both queries and query results in order to deter and detect malicious use.

4.6 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 data usefulness and privacy protection.

4.6.1 Validating data usefulness

De-identification decreases data quality and the usefulness of the resulting dataset. It is therefore important to assure that the de-identified dataset is still useful for the intended application—otherwise there is no reason to go through the expense and added risk of de-identification.

Several approaches exist for validating data usefulness. For example, insiders can perform statistical calculations on both the original dataset and on the de-identified dataset and compare the results to see if the de-identification resulted in changes that are unacceptable. Agencies can engage trusted outsiders to examine the de-identified dataset and determine if the data can be used for the intended purpose.

4.6.2 Validating privacy protection

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.
- Examining the resulting data files to make sure that the resulting data meet stated goals for ambiguity under a k-anonymity model, if such a standard is desired.
- Critically evaluating all default assumptions used by software that performs data modification or modeling.
- Conducting a *motivated intruder test* to see if reasonably competent outside individuals can perform re-identification using publicly available datasets. Motivations for a motivated intruder can include prurient interest; the goal of causing embarrassment or harm; revealing private facts about public figures; or engaging in a reputation attack. Details of the motivated intruder test can be found in *Anonymisation: code of practice, managing data protection risk*, published by the United Kingdom's Information

[139] Dinur, Irit and Kobbi Nissim, *Revealing Information while Preserving Privacy*, Proceedings of the 22nd Symposium on Principles of Database Construction (SIGMOD-SIGACT-SIGART), pp. 202-210, 2003. DOI:10.1145/773153.773173.

Commissioner's Office.[140]

- Providing the team conducting the motivated intruder test with using confidential agency data, to simulated what might happen in the result of a breach or a hostile insider.

These approaches do not provide provable guarantees on the protection offered by de-identification, but they may be useful as part of an overall agency risk assessment.[141] Applications that require provable privacy guarantees should rely on formal privacy methods such as differential privacy when planning their data releases.

Validating the privacy protection of de-identified data is greatly simplified by using validated de-identification software, as discussed in Section 6, "Evaluation."

[140] Anonymisation: code of practice, managing data protection risk. Information Commissioner's Office. 2012. https://ico.org.uk/media/1061/anonymisation-code.pdf

[141] Note: Although there exist other documents discussing de-identification use the term *risk assessment* to refer to a specific calculation of ambiguity using the k-anonyminity de-identification model, this document uses the term *risk assessment* to refer to a much broader process. Specifically risk assessment is defined as: "The process of identifying, estimating, and prioritizing risks to organizational operations (including mission, functions, image, reputation), organizational assets, individuals, other organizations, and the Nation, resulting from the operation of an information system. Part of risk management, incorporates threat and vulnerability analyses, and considers mitigations provided by security controls planned or in place. Synonymous with risk analysis." [NIST SP 800-39]

5 Software Requirements, Evaluation and Validation

Agencies performing de-identification should clearly define the requirements for de-identification algorithms and the software that implements those algorithms. They should be sure that the algorithms that they intend to use are validated, that the software that implements the algorithms as expected, and the data that results from the operation of the software are correct.[142]

5.1 Evaluating Privacy Preserving Techniques

There have 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 technique 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 considering 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.

5.2 De-Identification Tools

A de-identification tool is a program that is involved in the creation of de-identified datasets.

5.2.1 De-Identification Tool Features

De-identification tools might perform many functions, including:

- Detection of identifying information
- Calculation of re-identification risk
- Performing de-identification

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

- 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. For this reason, de-identification tools should be validated. For further information, see Section 6, "Software Requirements, Evaluation."

Appendix 8.7, "Specific De-Identification Tools," provides a listing of some de-identification tools that were known at the time of this publication.

5.2.2 Data Provenance and File Formats

Output files created by de-identification tools and data masking tools can record provenance information, such as metadata regarding input datasets, the de-identification methods used, and the resulting decrease in data quality. Output files can also be explicitly marked to indicate that they have been de-identified. For example, de-identification profiles that are part of the Digital Imaging and Communications in Medicine (DICOM) specification indicate which elements are direct vs quasi identifiers, and which de-identification algorithms have been employed.[143]

5.2.3 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, and data masking tools should not be used as the sole mechanism to assure confidentiality in non-public data sharing.

5.3 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

[143] See Appendix E, "Attribute Confidentiality Profiles," DICOM Standards Committee, DICOM PS3.15 2016e — Security and System Management Profiles, 2016 National Electrical Manufacturers Association (NEMA). http://dicom.nema.org/medical/dicom/current/output/html/part15.html#chapter_E

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 does not leak identifying information including in unexpected ways such as through the inaccuracies of floating-point arithmetic or the differences in execution time (if observable to an adversary).
- The software has sufficient usability that it can be operated 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.

5.4 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.

Agencies can create or adopt standards regarding the quality and accuracy of de-identified data. If data accuracy cannot be well maintained along with data privacy goals, then the release of data that is inaccurate for statistical analyses could potentially result in incorrect scientific conclusions and incorrect policy decisions.

6 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 strengthened when agencies employ formal models for privacy protection such as differential privacy, because the mathematical models that these systems use are designed to assure privacy protection irrespective of future data releases or developments in re-identification technology. However, the mathematics underlying these systems is very new, and there is little experience within the government in using these systems. Thus, 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.

Finally, different countries have different standards and policies regarding the definition and use of de-identified data. Information that is regarded as de-identified in one jurisdiction may be regarded as being identifiable in another.

7 References

7.1 Standards

- ASTM E1869-04(2014) Standard Guide for Confidentiality, Privacy, Access, and Data Security Principles for Health Information Including Electronic Health Records

- DICOM PS3.15 2016d – Security and System Management Profiles Chapter E Attribute Confidentiality Profiles, DICOM Standards Committee, NEMA 2016. http://dicom.nema.org/medical/dicom/current/output/html/part15.html#chapter_E

- HITRUST De-Identification Working Group (2015, March). De-Identification Framework: A Consistent, Managed Methodology for the De-Identification of Personal Data and the Sharing of Compliance and Risk Information. Frisco, TX: HITRUST. Retrieved from https://hitrustalliance.net/de-identification-license-agreement/

- 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.

- IHE IT Infrastructure Handbook, De-Identification, Integrating the Healthcare Enterprise, June 6, 2014. http://www.ihe.net/User_Handbooks/

7.2 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

- *Linking Data for Health Services Research: A Framework and Instructional Guide., Dusetzina SB, Tyree S, Meyer AM, Meyer A, Green L, Carpenter WR. (Prepared by the University of North Carolina at Chapel Hill under Contract No. 290-2010-000141.) AHRQ Publication No. 14-EHC033-EF. Rockville, MD: Agency for Healthcare Research and Quality; September 2014.*

7.3 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/

7.4 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

- 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

- Khaled El Emam, Guide to the De-Identificaiton of Personal Health Information, CRC Press, 2013

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

- 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

- Anco Hundepool, Josep Domingo-Ferrer, Luisa Franconi, Sarah Giessing, Eric Schulte Nordholt, Keith Spicer, Peter-Paul de Wolf, *Statistical Disclosure Control*, Wiley, September 2012.

7.5 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

7.6 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)

anticipated re-identification rate: when an organization contemplates performing re-identification, the re-identification rate that the resulting de-identified data are likely to have.

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)[144]

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."[145]

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)

---

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

[145] *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

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

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."[146]

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)

expert determination: within the context of de-identification, expert determination refers to the Expert Determination method for de-identifying protected health information in accordance with

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

the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule.

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

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)

identifying information: information that can be used to distinguish or trace an individual's identity, such as their name, social security number, biometric records, etc. alone, or when combined with other personal or identifying information which is linked or linkable to a specific individual, such as date and place of birth, mother’s maiden name, etc. (OMB M-07-16)

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[147])

k-anonymity: a technique "to release person-specific data such that the ability to link to other information using the quasi-identifier is limited."[148] 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[149]

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[150]

motivated intruder test: "The 'motivated intruder' is taken to be a person who starts without any prior knowledge but who wishes to identify the individual from whose personal data the anonymised data has been derived. This test is meant to assess whether the motivated intruder would be successful." [151]

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

non-public personal information: information about a person that is not publicly known; called "private information" in some other publications.

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*

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

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

[149] 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.

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

[151] Anonymisation: code of practice, managing data protection risk. Information Commissioner's Office. 2012. https://ico.org.uk/media/1061/anonymisation-code.pdf

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."[152] (SP800-122)

perturbation-based methods: “Perturbation-based methods falsify the data before publication by introducing an element of error purposely for confidentiality reasons. This error can be inserted in the cell values after the table is created, which means the error is introduced to the output of the data and will therefore be referred to as output perturbation, or the error can be inserted in the original data on the microdata level, which is the input of the tables one wants to create; the method with then be referred to as data perturbation—input perturbation being the better but uncommonly used expression. Possible methods are: - rounding; - random perturbation; - disclosure control methods for microstatistics applied to macrostatistics.” (OECD Glossary of Statistical Terms)

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.[153] Typically, pseudonymization is implemented by 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)

---

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

[153] 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.”

quasi-identifier: a variable that can be used to identify an individual through association with another variable

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 restores the association between a set of de-identified data and a data subject

re-identification risk: a measure of the extent to which an entity is threated by the re-identification of records within a dataset, typically a function of: (i) the adverse impacts that would arise if the re-identification would occur; and (ii) the likelihood of occurrence.

re-identification rate: the percentage of records in a dataset that can be re-identified.

risk: "A measure of the extent to which an entity is threatened by a potential circumstance or event, and typically a function of: (i) the adverse impacts that would arise if the circumstance or event occurs; and (ii) the likelihood of occurrence." (CNSSI No. 4009)

risk assessment: "The process of identifying, estimating, and prioritizing risks to organizational operations (including mission, functions, image, reputation), organizational assets, individuals, other organizations, and the Nation, resulting from the operation of an information system. Part of risk management, incorporates threat and vulnerability analyses, and considers mitigations provided by security controls planned or in place. Synonymous with risk analysis." (NIST SP 800-39)

safe harbor: within the context of de-identification, safe harbor refers to the Safe Harbor method for de-identifying protected health information in accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule.

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

7.7 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.

7.7.1 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,[154] k-Map, Strict-average risk, ℓ-Diversity,[155] t-Closeness,[156] δ-Disclosure privacy,[157] 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[158] 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 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

---

[154] 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

[155] 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

[156] 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

[157] 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

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

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.

7.7.2 Free Text

BoB, a best-of-breed automated text de-identification system for VHA clinical documents,[159] 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/

7.7.3 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

[159] 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.