

DRAFT NIST Special Publication 800-188

De-Identifying Government Datasets

Simson L. Garfinkel

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

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

DRAFT NIST Special Publication 800-188

De-Identifying Government Datasets

Simson L. Garfinkel
Information Access Division
Information Technology Laboratory

August 2016



U.S. Department of Commerce
Penny Pritzker, Secretary

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

Authority

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

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

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

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: August 25, 2016 through September 26, 2016

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

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

Reports on Computer Systems Technology

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

Abstract

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

Keywords

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

Acknowledgements

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

Audience

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

Table of Contents

Executive Summary	vi
1 Introduction	1
1.1 Document Purpose and Scope	3
1.2 Intended Audience	3
1.3 Organization	3
2 Introducing De-Identification	5
2.1 Historical Context	5
2.2 NISTIR 8053.....	6
2.3 Terminology.....	7
3 Governance and Management of Data De-Identification	11
3.1 Identifying Goals and Intended Uses of De-Identification.....	11
3.2 Evaluating Risks Arising from De-Identified Data Releases	12
3.2.1 Probability of Re-Identification.....	13
3.2.2 Adverse Impacts Resulting from Re-Identification	15
3.2.3 Impacts other than re-identification	16
3.2.4 Remediation	16
3.3 Data Life Cycle.....	16
3.4 Data Sharing Models.....	18
3.5 The Five Safes	19
3.6 Disclosure Review Boards	20
3.7 De-Identification Standards.....	22
3.7.1 Benefits of Standards	23
3.7.2 Prescriptive De-Identification Standards	23
3.7.3 Performance Based De-Identification Standards	23
3.8 Education, Training and Research.....	24
4 Technical Steps for Data De-Identification	25
4.1 Determine the Privacy, Data Usability, and Access Objectives.....	25
4.2 Data Survey.....	25
4.2.1 Data Modalities.....	25
4.2.2 De-identifying dates.....	27
4.2.3 De-identifying geographical locations.....	28
4.2.4 De-identifying genomic information	28
4.3 A de-identification workflow.....	29
4.4 De-identification by removing identifiers and transforming quasi-identifiers.....	30
4.4.1 Removing or Transformation of Direct Identifiers.....	32

- 4.4.2 Pseudonymization 32
- 4.4.3 Transforming Quasi-Identifiers 33
- 4.4.4 Challenges Posed by Aggregation Techniques 34
- 4.4.5 Challenges posed by High-Dimensionality Data 35
- 4.4.6 Challenges Posed by Linked Data 35
- 4.4.7 Post-Release Monitoring 36
- 4.5 Synthetic Data 36
 - 4.5.1 Partially Synthetic Data 36
 - 4.5.2 Fully Synthetic Data 37
 - 4.5.3 Synthetic Data with Validation 38
 - 4.5.4 Synthetic Data and Open Data Policy 38
 - 4.5.5 Creating a synthetic dataset with differential privacy 38
- 4.6 De-Identifying with an interactive query interface 40
- 4.7 Validating a de-identified dataset 41
 - 4.7.1 Validating privacy protection with a Motivated Intruder Test 41
 - 4.7.2 Validating data usefulness 41
- 5 Requirements for De-Identification Tools 42**
 - 5.1 De-Identification Tool Features 42
 - 5.2 Data Masking Tools 42
- 6 Evaluation 43**
 - 6.1 Evaluating Privacy Preserving Techniques 43
 - 6.2 Evaluating De-Identification Software 43
 - 6.3 Evaluating Data Quality 44
- 7 Conclusion 45**

List of Appendices

- Appendix A References 46**
 - A.1 Standards 46
 - A.2 US Government Publications 46
 - A.3 Publications by Other Governments 47
 - A.4 Reports and Books: 47
 - A.5 How-To Articles 48
- Appendix B Glossary 49**
- Appendix C Specific De-Identification Tools 54**
 - C.1 Tabular Data 54
 - C.2 Free Text 55
 - C.3 Multimedia 55

1 Executive Summary

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

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

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

- 17 • For four decades, official statistical agencies have researched and investigated methods
18 broadly termed *Statistical Disclosure Limitation (SDL)* or *Statistical Disclosure*
19 *Control*^{3,4}
- 20 • In the 1990s there was an increase in the unrestricted release of microdata, or individual
21 responses from surveys or administrative records. Initially these releases merely stripped
22 obviously identifying information such as names and social security numbers (what are
23 now called direct identifiers). Following some releases, researchers discovered that it was
24 possible to re-identify individual data by triangulating with some of the remaining
25 identifiers (now called quasi-identifiers or indirect identifiers).⁵ The result of this

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

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

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

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

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

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

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

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

39 Since the publication of NISTIR 8053, NIST has continued research in the area of de-
40 identification. NIST met with de-identification experts within and outside the United States
41 Government, convened a Government Data De-Identification Stakeholder's Meeting in June
42 2016, and conducted an extensive literature review.

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

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

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

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

⁷ Cynthia Dwork. 2006. Differential 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

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

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

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

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

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

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

¹⁰ Ibid.

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

88 1 Introduction

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

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

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

- 106 • US Code Title 13, Section 9 which governs confidentiality of information provided to the
107 Census Bureau, prohibits “any publication whereby the data furnished by any particular
108 establishment or individual under this title can be identified.”
- 109 • The release of personal information by the government is generally covered by the
110 Privacy Act of 1974¹⁴ and the E-Government Act of 2002.¹⁵ Specifically, the E-
111 Government Act states that “[d]ata or information acquired by an agency under a pledge
112 of confidentiality for exclusively statistical purposes shall not be disclosed by an agency
113 in identifiable form, for any use other than an exclusively statistical purpose, except with
114 the informed consent of the respondent.”¹⁶
- 115 • The Confidentiality Information Protection and Statistical Efficiency Act of 2002
116 requires that federal statistical agencies “establish appropriate administrative, technical,
117 and physical safeguards to insure the security and confidentiality of records and to protect
118 against any anticipated threats or hazards to their security or integrity which could result

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

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

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

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

¹⁶ Pub.L. 107-347 § 512 (b)(1).

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

167 **1.1 Document Purpose and Scope**

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

172 **1.2 Intended Audience**

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

178 **1.3 Organization**

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

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

190 This publication also includes three appendices: “References,” “Glossary,” and “Specific De-
191 Identification Tools.”

192 2 Introducing De-Identification

193 This document presents recommendations for de-identifying government datasets.

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

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

210 2.1 Historical Context

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

- 212 • For four decades, official statistical agencies have researched and investigated methods
 213 broadly termed *Statistical Disclosure Limitation (SDL)* or *Statistical Disclosure*
 214 *Control*^{23,24} Most of these methods were created to allow the release of statistical tables
 215 and *public use files (PUF)* that allow users to learn factual information or perform
 216 original research, while protecting the privacy of the individuals in the dataset. SDL is
 217 widely used in contemporary statistical reporting.
- 218 • In the 1990s, there was an increase in the release of *microdata* files for public use, with
 219 individual responses from surveys or administrative records. Initially these releases
 220 merely stripped obviously identifying information such as names and social security
 221 numbers (what are now called *direct identifiers*). Following some releases, researchers
 222 discovered that it was possible to re-identify individuals' data by triangulating with some
 223 of the remaining identifiers (now called *quasi-identifiers* or *indirect identifiers*).²⁵ The

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

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

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

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

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

238 Separately, during the first decade of the 21st century there was a growing awareness within the
 239 US Government about the risks that could result from the improper handling and inadvertent
 240 release of personal identifying and financial information. This realization, combined with a
 241 growing number of inadvertent data disclosures within the US government, resulted in President
 242 George Bush signing Executive Order 13402 establishing an Identity Theft Task Force on May
 243 10, 2006.²⁹ A year later the Office of Management and Budget issued Memorandum M-07-16³⁰
 244 which required Federal agencies to develop and implement breach notification policies. As part
 245 of this effort, NIST issued Special Publication 800-122, *Guide to Protecting the Confidentiality*
 246 *of Personally Identifiable Information (PII)*.³¹ These policies and documents had the specific
 247 goal of limiting the accessibility of information that could be directly used for identity theft, but
 248 did not create a framework for processing government datasets so that they could be released
 249 without impacting the privacy of the data subjects.

250 2.2 NISTIR 8053

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

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

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

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

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

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

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

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

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

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

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

269 **2.3 Terminology**

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

274 *de-identification* is the “general term for any process of removing the association between a set
275 of identifying data and the data subject.”³² De-identification takes an *original dataset* and
276 produces a *de-identified dataset*.

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

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

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

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

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

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

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

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

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

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

314 This document avoids the term “personally identifiable information.” Instead, the phrase
315 *personal information* is used to denote information relating to individuals, and *identifying*
316 *information* is used to denote information that identifies individuals. Therefore, identifying
317 information is personal information, but personal information is not necessarily identifying
318 information. *Private information* is used to describe information that is in a dataset that is not
319 publicly available. Private information is not necessarily identifying.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

352 Differential privacy⁴² is a model based on a mathematical definition of privacy that considers the
353 risk to an individual from the release of a query on a dataset containing their personal
354 information. Differential privacy is also a set of mathematical techniques that can achieve the
355 differential privacy definition of privacy. Differential privacy prevents disclosure by adding non-
356 deterministic noise (usually small random values) to the results of mathematical operations
357 before the results are reported.⁴³ Differential privacy's mathematical definition holds that the
358 result of an analysis of a dataset should be roughly the same before and after the addition or
359 removal of the data from any individual. This works because the amount of noise added masks
360 the contribution of any individual. The degree of sameness is defined by the parameter ϵ
361 (epsilon). The smaller the parameter ϵ , the more noise is added, and the more difficult it is to
362 distinguish the contribution of a single individual. The result is increased privacy for all
363 individuals, both those in the sample and those in the population from which the sample is drawn
364 who are not present in the dataset. Differential privacy can be implemented in an online query
365 system or in a batch mode in which an entire dataset is de-identified at one time. In common
366 usage, the phrase "differential privacy" is used to describe both the formal mathematical
367 framework for evaluating privacy loss, and for algorithms that provably provide those privacy
368 guarantees.

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

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

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

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

3 Governance and Management of Data De-Identification

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

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

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

3.1 Identifying Goals and Intended Uses of De-Identification

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

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

For example:

- **Federal Statistical Agencies** that collect, process, and publish data for use by researchers, business planners, and other well-established purposes. These agencies are likely to have in place established standards and methodologies for de-identification. As these agencies evaluate new approaches to de-identification, they should seek to document inconsistencies with previous data releases that may result in people with
- **Federal Awarding Agencies** are allowed under OMB Circular A-110 to require that institutions of higher education, hospitals, and other non-profit organizations receiving

⁴⁴ NISTIR 8053 §2.4, §3.6

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

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

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

434 3.2 Evaluating Risks Arising from De-Identified Data Releases

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

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

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

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

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

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

449 3.2.1 Probability of Re-Identification

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

- 451 • **Identity disclosures** — Associating a specific individual with the corresponding
452 record(s) in the data set. Identity disclosure can result from insufficient de-identification,
453 re-identification by linking, or pseudonym reversal.
- 454 • **Attribute disclosure** — determining that an attribute described in the dataset is held by a
455 specific individual, even if the record(s) associated with that individual is(are) not
456 identified. Attribute disclosure can occur without identity disclosure if the de-identified
457 dataset contains data from a significant number of relatively homogeneous individuals.⁴⁸
458 In these cases, de-identification does not protect against attribute disclosure.
- 459 • **Inferential disclosure** — being able to make an inference about an individual, even if
460 the individual was not in the dataset prior to de-identification. De-identification cannot
461 protect against inferential disclosure.

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

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

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

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

⁴⁸ NISTIR 8053 §2.4, p 13.

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

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

- 477 • *Unknown Inclusion Re-identification Probability (UIRP)*. The probability of finding the
478 record that matches a specific individual, without first knowing if the individual is or the
479 maximum is not in the dataset. UIRP can be expressed as a probability for an individual
480 record in the dataset, probability averaged over the entire population (AUIRP).⁵¹
481 • *Recording matching probability (RMP)*. The probability of finding the record that matches
482 a specific individual chosen from the population. RMP can be expressed as the
483 probability for a specific record (RMP), the probability averaged over the entire dataset
484 (ARMP), or the maximum probability over the entire dataset.
485 • *Inclusion probability (IP)*, the probability that a specific individual's presence in the
486 dataset can be inferred.

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

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

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

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

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

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

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

525 **3.2.2 Adverse Impacts Resulting from Re-Identification**

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

529 Potential adverse impacts on individuals include:

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

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

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

545 Potential adverse impacts on society as a whole include:

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

549 its use as a result.

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

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

556 3.2.3 Impacts other than re-identification

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

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

565 3.2.4 Remediation

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

570 3.3 Data Life Cycle.

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

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

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

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

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

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

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

589 Agencies performing de-identification should document that:

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

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

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

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

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

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

613 3.4 Data Sharing Models

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

- 616 • **The Release and Forget Model:**⁵⁷ The de-identified data may be released to the public,
 617 typically by being published on the Internet. It can be difficult or impossible for an
 618 organization to recall the data once released in this fashion and may limit information for
 619 future releases.
- 620 • **The Data Use Agreement (DUA) Model:** The de-identified data may be made available
 621 to under a legally binding data use agreement that details what can and cannot be done
 622 with the data. Typically, data use agreements may prohibit attempted re-identification,
 623 linking to other data, and redistribution of the data without a similarly binding DUA. A
 624 DUA will typically be negotiated between the data holder and qualified researchers (the
 625 “qualified investigator model”⁵⁸), although they may be simply posted on the Internet
 626 with a click-through license agreement that must be agreed to before the data can be
 627 downloaded (the “click-through model”⁵⁹).
- 628 • **The Simulated Data with Verification Model:** The original dataset is used to create a
 629 simulated dataset that contains many of the aspects of the original dataset. The simulated
 630 dataset is released, either publically or to vetted researchers. The simulated data can be
 631 used to develop queries or analytic software; these queries and/or software can then be
 632 provided to the agency, which will then apply them to the original data. The results of the
 633 queries and/or analytics processes can then be subjected to Statistical Disclosure
 634 Limitation and the results provided to the researchers.
- 635 • **The Enclave Model:**^{60,61} The de-identified data may be kept in a segregated enclave that
 636 restricts the export of the original data, and instead accepts queries from qualified
 637 researchers, runs the queries on the de-identified data, and responds with results.
 638 Alternatively, vetted researchers may travel to the enclave to perform their research, as is

⁵⁶ NISTIR 8053 §2.5, p. 14

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

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

⁵⁹ *Ibid.*

⁶⁰ *Ibid.*

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

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

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

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

654 3.5 The Five Safes

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

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

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

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

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

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

668 research will be conducted.

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

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

678 **3.6 Disclosure Review Boards⁶⁵**

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

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

686 “The Mission of the Department of Education Disclosure Review Board (ED-DRB) is to
687 review proposed data releases by the Department’s principal offices (POs) through a
688 collaborate technical assistance, aiding the Department to release as much useful data as
689 possible, while protecting the privacy of individuals and the confidentiality of their data, as
690 required by law.”⁶⁶

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

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

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

⁶⁷ NISTIR 8053 §2.4, p. 13

701 empowered to approve the use of such mechanisms.

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

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

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

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

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

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

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

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

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

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

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

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

754 Key elements of a DRB:

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

765 **3.7 De-Identification Standards**

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

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

770 3.7.1 Benefits of Standards

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

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

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

782 3.7.2 Prescriptive De-Identification Standards

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

785 The “Safe Harbor” method of the HIPAA Privacy Rule⁷² is an example of a prescriptive de-
786 identification standard. The intent of the Safe Harbor method is to “provide covered entities with
787 a simple method to determination if [] information is adequately de-identified.”⁷³ It does this by
788 specifying 18 kinds of identifiers that, once removed, results in the de-identification of Protected
789 Health Information (PHI) and the subsequent relaxing of privacy regulations. Although the
790 Privacy Rule does state that a covered entity employing the Safe Harbor method must have no
791 “actual knowledge” that the PHI, once de-identified, could still be used to re-identify individuals,
792 covered entities are not obligated to employ experts or mount re-identification attacks against
793 datasets to verify that the use of the Safe Harbor method has in fact resulted in data that cannot
794 be re-identified.

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

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

802 3.7.3 Performance Based De-Identification Standards

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

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

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

804 after it is de-identified.

805 The “Expert Determination” method of the HIPAA Privacy Rule is an example of a performance
806 based de-identification standard. Under the rule, a technique for de-identifying data is sufficient
807 if an appropriate expert “determines that the risk is very small that the information could be used,
808 alone or in combination with other reasonably available information, by an anticipated recipient
809 to identify an individual who is a subject of the information.”⁷⁴

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

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

817 **3.8 Education, Training and Research**

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

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

827 **4 Technical Steps for Data De-Identification**

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

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

834 **4.1 Determine the Privacy, Data Usability, and Access Objectives**

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

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

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

851 **4.2 Data Survey**

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

854 **4.2.1 Data Modalities**

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

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

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

869 • **Dates and times** require special attention when de-identifying, because all dates within a
870 dataset are inherently linked to the natural progression of time. Some dates and times are
871 highly identifying, with others are not. Some of these linkages may be relevant to the
872 purpose of the dataset, the identity of the data subjects, or both. Dates may also form the
873 basis of linkages between dataset records or even within a record—for example, a record
874 may contain the date of admission, the date of discharge, and the number of days in
875 residence. Thus, care should be taken when de-identifying dates to locate and properly
876 handle potential linkages and relationships: applying different techniques to different
877 fields may result in information being left in a dataset that can be used for re-
878 identification. Specific issues regarding date de-identification are discussed below in
879 §4.2.2.

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

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

⁷⁵ NISTIR 8053, §4.5 p. 37

⁷⁶ NISTIR 8053, §4.1 p. 30

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

- 894 • **Photos and video** may contain identifying information such as printed names (e.g. name
895 tags). There also exists a range of biometric techniques for matching photos of
896 individuals against a dataset of photos and identifiers.⁷⁸
- 897 • **Medical imagery** poses additional problems over photographs and video due to the
898 presence of many different kinds of identifiers. For example, identifying information may
899 be present in the image itself (e.g. a photo may show an identifying scar or tattoo), an
900 identifier may be “burned in” to the image area, or an identifier may be present in the file
901 metadata. The body part in the image itself may also recognized through the use of a
902 biometric algorithm and dataset.⁷⁹
- 903 • **Genetic sequences** and other kinds of sequence information can be identified by
904 matching to existing databanks that match sequences and identities. There is also
905 evidence that genetic sequences from individuals who are not in datasets can be matched
906 through genealogical triangulation, a process that uses genetic information and other
907 information as quasi-identifiers to single-out a specific identity.⁸⁰ At the present time
908 there is no known method to reliably de-identify genetic sequences. Specific issues
909 regarding the de-identification of genetic information is discussed below in §4.2.4.

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

913 **4.2.2 De-identifying dates**

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

919 Several strategies have been developed for de-identifying dates:

- 920 • Under the HIPAA Privacy Rule, dates must be generalized to no greater specificity than
921 the year (e.g. July 4, 1776 becomes 1776).
- 922 • Dates within a single person’s record can be systematically adjusted by a random amount.
923 For example, dates of a hospital admission and discharge might be systematically moved
924 the same number of days (e.g. ± 1000).⁸¹

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

⁷⁸ NISTIR 8053, §4.2 p. 32

⁷⁹ NISTIR 8053, §4.3 p. 35

⁸⁰ NISTIR 8053, §4.4 p. 36

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

- 925 • In addition to a systematic shift, the intervals between dates can be perturbed to protect
926 against re-identification attacks involving identifiable intervals while still maintaining the
927 ordering of events.
- 928 • Some dates cannot be arbitrarily changed without compromising data quality. For
929 example, it may be necessary to preserve day-of-week or whether a day is a work day or
930 a holiday.
- 931 • Likewise, some ages can be randomly adjusted without impacting data quality, while
932 others cannot. For example, in many cases the age of an individual can be randomly
933 adjusted ± 2 years if the person is over the age of 25, but not if their age is between 1 and
934 3.

935 4.2.3 De-identifying geographical locations

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

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

946 4.2.4 De-identifying genomic information

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

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

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

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

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

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

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

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

975 **4.3 A de-identification workflow**

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

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

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

983 Step 3. Identify the data modalities that are present in the data to be de-identified. (See §
984 4.2.1 below.)

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

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

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

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

- 991 Step 7. Export transformed data to a different system for testing and validation.
- 992 Step 8. Test the de-identified data quality. Perform analyses on the de-identified data to
993 make sure that it has sufficient usefulness and data quality.
- 994 Step 9. Attempt re-identification. Examine the de-identified data to see if it can be re-
995 identified. This step may involve the engagement of an outside tiger team.
- 996 Step 10. Document the de-identification techniques and the results in a written report.
- 997

998 **4.4 De-identification by removing identifiers and transforming quasi-** 999 **identifiers**

1000 De-identification based on the removal of identifiers and transformation of quasi-identifiers is
1001 one of the most common approaches for de-identification currently in use. This approach has the
1002 advantage of being conceptually straightforward and there being a long institutional history in
1003 using this approach within both federal statistical agencies and the healthcare industry. This
1004 approach has the disadvantage of being not based on formal methods for assuring privacy
1005 protection. The lack of formal methods does not mean that this approach cannot protect privacy,
1006 but it does mean that privacy protection is not assured.

1007 Below is a sample protocol for de-identifying data by removing identifiers and transforming
1008 quasi-identifiers:⁸⁵

- 1009 Step 1. Determine the re-identification risk threshold. The organization determines
1010 acceptable risk for working with the dataset and possibly mitigating controls, based on
1011 strong precedents and standards (e.g., Working Paper 22: Report on Statistical Disclosure
1012 Control).
- 1013 Step 2. Determine the information in the dataset that could be used to identify the data
1014 subjects. Identifying information can include:
- 1015 a. **Direct identifiers**, such as names, phone numbers, and other information that
1016 unambiguously identifies an individual.
 - 1017 b. **Quasi-identifiers** that could be used in a linkage attack. Typically, quasi-
1018 identifiers identify multiple individuals and can be used to triangulate on a
1019 specific individual.

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

- 1020 c. **High-dimensionality data**⁸⁶ that can be used to single out data records and thus
 1021 constitute a unique pattern that could be identifying, if these values exist in a
 1022 secondary source to link against.⁸⁷
- 1023 Step 3. Determine the direct identifiers in the dataset. An expert determines the elements
 1024 in the dataset that serve only to identify the data subjects.
- 1025 Step 4. Mask (transform) direct identifiers. The direct identifiers are either removed or
 1026 replaced with pseudonyms.
- 1027 Step 5. Perform threat modeling. The organization determines the additional information
 1028 they might be able to use for re-identification, including both quasi-identifiers and non-
 1029 identifying values that an adversary might use for re-identification.
- 1030 Step 6. Determine the minimal acceptable data quality. In this step, the organization
 1031 determines what uses can or will be made with the de-identified data.
- 1032 Step 7. Determine the transformation process that will be used to manipulate the quasi-
 1033 identifiers. Pay special attention to the data fields containing dates and geographical
 1034 information, removing or recoding as necessary.
- 1035 Step 8. Import (sample) data from the source dataset. Because the effort to acquire data
 1036 from the source (identified) dataset may be substantial, El Emam and Malin recommend a
 1037 test data import run to assist in planning.
- 1038 Step 9. Review the results of the trial de-identification. Correct any coding or algorithmic
 1039 errors that are detected.
- 1040 Step 10. Transform the quasi-identifiers for the entire dataset.
- 1041 Step 11. Evaluate the actual re-identification risk. The actual identification risk is
 1042 calculated. As part of this evaluation, every aspect of the released dataset should be
 1043 considered in light of the question, “can *this* information be used to identify someone?”
- 1044 Step 12. Compare the actual re-identification risk with the threshold specified by the
 1045 policy makers.
- 1046 Step 13. If the data do not pass the actual risk threshold, adjust the procedure and Step 11.
 1047 For example, additional transformations may be required. Alternatively, it may be
 1048 necessary to remove outliers. Step 9: Set parameters and apply data transformations.

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

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

1049 **4.4.1 Removing or Transformation of Direct Identifiers**

1050 Once a determination is made regarding direct identifiers, they must be removed. Options for
1051 removal include:

- 1052 • Masking with a repeating character, such as XXXXXX or 999999.
- 1053 • Encryption. After encryption the cryptographic key should be discarded to prevent
1054 decryption or the possibility of a brute force attack. However, the key must not be
1055 discarded if there is a desire to employ the same transformation at a later point in time,
1056 but rather stored in a secure location separate from the de-identified dataset.
- 1057 • Hashing with a keyed hash, such as an HMAC. The hash key should be have sufficient
1058 randomness to defeat a brute force attack aimed at recovering the hash key. For example,
1059 SHA-256 HMAC with a 256-bit randomly generated key. As with encryption, the key
1060 should be discarded unless there is a desire for repeatability. (Note: hash functions should
1061 not be used without a key.)
- 1062 • Replacement with keywords, such as transforming “George Washington” to “PATIENT.”
- 1063 • Replacement by realistic surrogate values, such as transforming “George Washington” to
1064 “Abraham Polk.”⁸⁸

1065 The technique used to remove direct identifiers should be clearly documented for users of the
1066 dataset, especially if the technique of replacement by realistic surrogate names is used.

1067 If the agency plans to make data available for longitudinal research and contemplates multiple
1068 data releases, then the transformation process should be repeatable, and the resulting transformed
1069 identities are *pseudonyms*. Agencies should be aware that there is a significantly increased risk of
1070 re-identification if a repeatable transformation is used.

1071 **4.4.2 Pseudonymization**

1072 Pseudonymization is a way of labeling multiple de-identified records from the same individual
1073 so that they can be linked together. Pseudonymization is a form of masking identifiers; it is *not* a
1074 form of de-identification.⁸⁹

1075 Pseudonymization generally increases the risk that de-identified data might be re-identified. By
1076 linking together records, pseudonymization increases the opportunities of finding identified data
1077 that can be linked with the de-identified data in a record linkage attack. Pseudonymization also
1078 carries that risk that the pseudonymization technique itself might be inverted or otherwise

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

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

1079 reversed, directly revealing the identities of the data subjects.

1080 4.4.3 Transforming Quasi-Identifiers

1081 Once a determination is made regarding quasi-identifiers, they should be transformed. A variety
1082 of techniques are available to transform quasi-identifiers:

- 1083 • **Top and bottom coding.** Outlier values that are above or below certain values are coded
1084 appropriately. For example, the HIPAA Privacy Rules calls for ages over 89 to be
1085 “aggregated into a single category of age 90 or older.”⁹⁰
- 1086 • **Micro aggregation,** in which individual microdata are combined into small groups that
1087 preserve some data analysis capability while providing for some disclosure protection.⁹¹
- 1088 • **Generalize categories with small values.** When preparing contingency tables, several
1089 categories with small values may be combined. For example, rather than reporting that
1090 there is 1 person with blue eyes, 2 people with green eyes, and 1 person with hazel eyes,
1091 it may be reported that there are 4 people with blue, green or hazel eyes.
- 1092 • **Data suppression.** Cells in contingency tables with counts lower than a predefined
1093 threshold can be suppressed to prevent the identification of attribute combinations with
1094 small numbers.⁹²
- 1095 • **Blanking and imputing.** Specific values that are highly identifying can be removed and
1096 replaced with imputed values.
- 1097 • **Attribute or record swapping,** in which attributes or records are swapped between
1098 records representing individuals. For example, data representing families in two similar
1099 towns within a county might be swapped with each other. “Swapping has the additional
1100 quality of removing any 100-percent assurance that a given record belongs to a given
1101 household,”⁹³ while preserving the accuracy of regional statistics such as sums and
1102 averages. For example, in this case the average number of children per family in the
1103 county would be unaffected by data swapping.
- 1104 • **Noise infusion.** Also called “partially synthetic data,” small random values may be added
1105 to attributes. For example, instead of reporting that a person is 84 years old, the person
1106 may be reported as being 79 years old. Noise infusion increases variance and leads to
1107 attenuation bias in estimated regression coefficients and correlations among attributes.⁹⁴

⁹⁰ HIPAA § 164.514 (b).

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

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

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

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

1108 The techniques are described in detail by two publications:

- 1109 • *Statistical Policy Working Paper #2* (Second version, 2005) by the Federal Committee on
1110 Statistical Methodology.⁹⁵ This 137-page paper also includes worked examples of
1111 disclosure limitation, specific recommended practices for Federal agencies, profiles of
1112 federal statistical agencies conducting disclosure limitation, and an extensive
1113 bibliography.
- 1114 • *The Anonymisation Decision-Making Framework*, by Mark Elliot, Elaine MacKey,
1115 Kieron O’Hara and Caroline Tudor, UKAN, University of Manchester, Manchester, UK.
1116 2016. This 156-page book provides tutorials and worked examples for de-identifying data
1117 and calculating risk.

1118 Swapping and noise infusion both introduce noise into the dataset, such that records literally
1119 contain incorrect data. These techniques can introduce sufficient noise to provide formal privacy
1120 guarantees.

1121 All of these techniques impact data quality, but whether they impact data *utility* depends upon
1122 the downstream uses of the data. For example, top-coding household incomes will not impact a
1123 measurement of the 90-10 quantile ratio, but it will impact a measurement of the top 1% of
1124 household incomes.⁹⁶

1125 In practice, statistical agencies typically do not document in detail the specific statistical
1126 disclosure technique that they use to transform quasi-identifiers, nor do they document the
1127 parameters used in the transformations nor the amount of data that have been transformed, as
1128 documenting these techniques can allow an adversary to reverse-engineer the specific values,
1129 eliminating the privacy protection.⁹⁷ This lack of transparency can result in erroneous
1130 conclusions on the part of data users.

1131 **4.4.4 Challenges Posed by Aggregation Techniques**

1132 Aggregation does not necessarily provide privacy protection, especially when data is presented
1133 as part of multiple data releases. Consider the hypothetical example of a school uses aggregation
1134 to report the number of students performing below, at, and above grade level:

Performance	Students
-------------	----------

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

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

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

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

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

1135

1136 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

1137

1138 By comparing the two tables, one can readily infer that the student who joined the school is
 1139 performing above grade level. Because aggregation does not inherently protect privacy, its use is
 1140 not sufficient to provide formal privacy guarantees.

1141 **4.4.5 Challenges posed by High-Dimensionality Data**

1142 Even after removing all of the unique identifiers and manipulating the quasi-identifiers, some
 1143 data can still be identifying if it of sufficient high-dimensionality, if there exists a way to link the
 1144 supposedly non-identifying values with an identity.⁹⁸

1145 **4.4.6 Challenges Posed by Linked Data**

1146 Data can be linked in many ways. Pseudonyms allow data records from the same individual to be
 1147 linked together over time. Family identifiers allow data from parents to be linked with their
 1148 children. Device identifiers allow data to be linked to physical devices, and potentially link
 1149 together all data coming from the same device. Data can also be linked to geographical locations.

1150 Data linkage increases the risk of re-identification by providing more attributes that can be used
 1151 to distinguish the true identity of a data record from others in the population. For example,
 1152 survey responses that are linked together by household are more readily re-identified than survey
 1153 responses that are not linked. For example, heart rate measurements may not be considered
 1154 identifying, but given a long sequence of tests, each individual in a dataset would have a unique
 1155 constellation of heart rate measurements, and thus the data set would be susceptible to being

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

1156 linked with another data set that contains these same values.

1157 Dependencies between records may result in record linkages even when there is no explicit
1158 linkage identifier. For example, it may be that an organization has new employees take a
1159 proficiency test within 7 days of being hired. This information would allow links to be drawn
1160 between an employee dataset that accurately reported an employee's start date and a training
1161 dataset that accurately reported the date that the test was administered, even if the sponsoring
1162 organization did not intend for the two datasets to be linkable.

1163 **4.4.7 Post-Release Monitoring**

1164 Following the release of a de-identified dataset, the releasing agency should monitor to assure
1165 that the assumptions made during the de-identification remain valid. This is because the
1166 identifiability of a dataset may increase over time.

1167 For example, the de-identified dataset may contain information that can be linked to an internal
1168 dataset that is later the subject of a data breach. In such a situation, the data breach will also
1169 result in the re-identification of the de-identified dataset.

1170 **4.5 Synthetic Data**

1171 An alternative to de-identifying using the technique presented in the previous section is to use
1172 the original dataset to create a synthetic dataset.

1173 Synthetic data can be created by two approaches:⁹⁹

- 1174 • Sampling an existing dataset and either adding noise to specific cells likely to have a high
1175 risk of disclosure, or replacing these cells with imputed values. (A “partially synthetic
1176 dataset.”)
- 1177 • Using the existing dataset to create a model and then using that model to create a
1178 synthetic dataset. (A “fully synthetic dataset.”)

1179 In both cases, the mathematics of differential privacy can be used to quantify the privacy
1180 protection offered by the synthetic dataset.

1181 **4.5.1 Partially Synthetic Data**

1182 A partially synthetic dataset is one in which some of the data is inconsistent with the original
1183 dataset. For example, data belonging to two families in adjoining towns may be swapped to
1184 protect the identity of the families. Alternatively, the data for an outlier variable may be removed
1185 and replaced with a range value that is incorrect (for example, replacing the value “60” with the
1186 range “30-35”). It is considered best practice that the data publisher indicate that some values
1187 have been modified or otherwise imputed, but not to reveal the specific values that have been

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

1188 modified.

1189 **4.5.2 Fully Synthetic Data**

1190 A fully synthetic dataset is a dataset for which there is no one-to-one mapping between data in
1191 the original dataset and in the de-identified dataset. One approach to create a fully synthetic
1192 dataset is to use the original dataset to create a high fidelity model, and then to use the model to
1193 produce individual data elements consistent with the model using a simulation.

1194 Fully synthetic datasets cannot provide more information to the downstream user than was
1195 contained in the original model. Nevertheless, some users may prefer to work with the fully
1196 synthetic dataset instead of the model:

- 1197 • Synthetic data provides users with the ability to develop queries and other techniques that
1198 can be applied to the real data, without exposing real data to users during the
1199 development process. The queries and techniques can then be provided to the data owner,
1200 which can run the queries or techniques on the real data and provide the results to the
1201 users.
- 1202 • Analysts may discover things from the synthetic data that they don't see in the model,
1203 even though the model contains the information. However, such discoveries should be
1204 evaluated against the real data to assure that the things that were discovered were actually
1205 in the original data, and not an artifact of the synthetic data generation.
- 1206 • Some users may place more trust in a synthetic dataset than in a model.
- 1207 • When researchers form their hypotheses working with synthetic data and then verify their
1208 findings on actual data, they are protected from pretest estimation and false-discovery
1209 bias.¹⁰⁰

1210 Both high-fidelity models and synthetic data generated from models may leak personal
1211 information that is potentially re-identifiable; the amount of leakage can be controlled using
1212 formal privacy models (such as differential privacy) that typically involve the introduction of
1213 noise.

1214 There are several advantages to agencies that chose to release de-identified data as a fully
1215 synthetic dataset:

- 1216 • It can be very difficult or even impossible to map records to actual people, so fully
1217 synthetic data offers very good privacy protection.
- 1218 • The privacy guarantees can be mathematically established and proven.

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

- 1219 • The privacy guarantees can remain in force even if there are future data releases.

1220 Fully synthetic data also has these disadvantages and limitations:

- 1221 • It is not possible to create pseudonyms that map back to actual people, because the
1222 records are fully synthetic.

- 1223 • The data release may be less useful for accountability or transparency. For example,
1224 investigators equipped with a synthetic data release would be unable to find the actual
1225 “people” who make up the release, because they would not actually exist.

- 1226 • It is impossible to find meaningful correlations or abnormalities in the synthetic data that
1227 are not represented in the model. For example, if a model is built by considering all
1228 possible functions of 1 and 2 variables, then any correlations found of 3 variables will be
1229 a spurious artifact of the way that the synthetic data were created, and not based on the
1230 underlying real data.

- 1231 • Users of the data may not realize that the data are synthetic. Simply providing
1232 documentation that the data are fully synthetic may not be sufficient public notification,
1233 since the dataset may be separated from the documentation. Instead, it is best to indicate
1234 in the data itself that the values are synthetic. For example, names like “SYNTHETIC
1235 PERSON” may be placed in the data. Such names could follow the distribution of real
1236 names but obviously be not real.

1237 **4.5.3 Synthetic Data with Validation**

1238 Agencies that share or publish synthetic data can optionally make available a validation service
1239 that takes queries or algorithms developed with synthetic data and applies them to actual data.
1240 The results of these queries or algorithms can then then be compared with the results of running
1241 the same queries on the synthetic data and the researchers warned if the results are different.
1242 Alternatively, the results can be provided to the researchers after the application of statistical
1243 disclosure limitation.

1244 **4.5.4 Synthetic Data and Open Data Policy**

1245 Releases of synthetic data can be confusing to the lay public. Specifically, synthetic data may
1246 contain synthetic individuals who appear quite similar to actual individuals in the population.
1247 Furthermore, fully synthetic datasets do not have a zero disclosure risk, because they still convey
1248 some private information about individuals. The disclosure risk may be greater when synthetic
1249 data are created with traditional data imputing techniques, rather than techniques based on formal
1250 privacy models.

1251 **4.5.5 Creating a synthetic dataset with differential privacy**

1252 A growing number of mathematical algorithms have been developed for creating synthetic
1253 datasets that meet the mathematical definition of privacy provided by differential privacy. Most
1254 of these algorithms will transform a dataset containing private data into a new dataset that
1255 contains synthetic data that nevertheless provides reasonably accurate results in response to a
1256 variety of queries. However there is no algorithm or implementation currently in existence that

1257 can be used by a person who is unskilled in the area of differential privacy.

1258 The classic definition of differential privacy is that if results of function calculated on a dataset
 1259 are indistinguishable within a certain privacy metric ϵ (epsilon) no matter whether any
 1260 possible individual is included in the dataset or removed from the dataset,¹⁰¹ then that
 1261 function is said to provide ϵ -differential privacy.

1262 In Dwork's mathematical formulation, the two datasets (with and without the individual) are
 1263 denoted by D_1 and D_2 , and the function that is said to be differential private is κ . The formal
 1264 definition of differential privacy is then:

1265 **Definition 2.**¹⁰² A randomized function κ gives ϵ -differential privacy if for all datasets D_1
 1266 and D_2 differing on at most one element, and all $S \subseteq \text{Range}(\kappa)$,

$$1267 \quad \Pr[\kappa(D_1) \in S] \leq e^\epsilon \times \Pr[\kappa(D_2) \in S]$$

1268 This definition that may be easier to understand if rephrased as a dataset D with an arbitrary
 1269 person p , and dataset $D - p$, the dataset without a person, and the multiplication operator
 1270 replaced by a division operator, e.g.:

$$1271 \quad \frac{\Pr[\kappa(D - p) \in S]}{\Pr[\kappa(D) \in S]} \leq e^\epsilon$$

1272 That is, the ratio between the probable outcomes of function κ operating on the datasets with and
 1273 without person p should be less than e^ϵ . If the two probabilities are equal, then $e^\epsilon = 1$, and $\epsilon =$
 1274 0. If the difference between the two probabilities is potentially infinite—that is, there is no
 1275 privacy—then $e^\epsilon = \infty$ and $\epsilon = \infty$.

1276 What this means in practice for the creation of a synthetic dataset with differential privacy and a
 1277 sufficiently large ϵ is that functions computed on the so-called “privatized” dataset will have a
 1278 similar probability distribution no matter whether any person in the original data that was used to
 1279 create the model is included or excluded. In practice, this similarity is provided by adding noise
 1280 to the model. For datasets drawn from a population with a large number of individuals, the model
 1281 (and the resulting synthetic data) will have a small amount of noise added. For models and
 1282 resulting created from a small population (or for contingency tables with small cell counts), this
 1283 will require the introduction of a significant amount of noise. The amount of noise added is
 1284 determined by the differential privacy parameter ϵ , the number of individuals in the dataset, and
 1285 the specific differential privacy mechanism that is employed.

1286 Smaller values of ϵ provide for more privacy but decreased data quality. As stated above, the

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

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

1287 value of 0 implies that the function κ provides the same answer no matter if anyone is removed
 1288 or a person's attributes changed, while the value of ∞ implies that the original dataset is released
 1289 with being privatized.

1290 Many academic papers on differential privacy have assumed a value for ϵ of 1.0 or e but have not
 1291 explained the rationale of the choice. Some researchers working in the field of differential
 1292 privacy have just started the process of mapping existing privacy regulations to the choice of ϵ .
 1293 For example, using a hypothetical example of a school that wished to release a dataset containing
 1294 the school year and absence days for a number of students, the value of ϵ using one set of
 1295 assumptions might be calculated to 0.3379 (producing a low degree of data quality), but this
 1296 number can safely be raised to 2.776 (and correspondingly higher data quality) without
 1297 significantly impacting the privacy protections.¹⁰³

1298 Another challenge in implementing differential privacy is the demands that the algorithms make
 1299 on the correctness of implementation. For example, a Microsoft researcher discovered that four
 1300 publicly available general purpose implementations of differential privacy contained a flaw that
 1301 potentially leaked private information because of the binary representation of IEEE floating point
 1302 numbers used by the implementations.¹⁰⁴

1303 Given the paucity of scholarly publications regarding the deployment of differential privacy in
 1304 real-world situation, combined with the lack of guidance and experience in choosing appropriate
 1305 values of ϵ , agencies that are interested in using differential privacy algorithms to allow
 1306 querying of sensitive datasets or for the creation of synthetic data should take great care to
 1307 assure that the techniques are appropriately implemented and that the privacy protections
 1308 are appropriate to the desired application.

1309 **4.6 De-Identifying with an interactive query interface**

1310 Another model for granting the public access to de-identified agency information is to construct
 1311 an interactive query interface that allows members of the public or qualified investigators to run
 1312 queries over the agency's dataset. This option has been developed by several agencies and there
 1313 are many different ways that it can be implemented.

- 1314 • If the queries are run on actual data, the results can be altered through the injection of
 1315 noise to protect privacy. Alternatively, the individual queries can be reviewed by agency
 1316 staff to verify that privacy thresholds are maintained.
- 1317 • Alternatively, the queries can be run on synthetic data. In this case, the agency can also
 1318 run queries on the actual data and warn the external researchers if the queries run on

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

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

1319 synthetic data diverse from the queries run on the actual data.

- 1320 • Query interfaces can be made freely available on the public internet, or they can be made
1321 available in a restricted manner to qualified researchers operating in secure locations.

1322 **4.7 Validating a de-identified dataset**

1323 Agencies should validate datasets after they are de-identified to assure that the resulting dataset
1324 meets the agency's goals in terms of both privacy protection and data usefulness.

1325 **4.7.1 Validating privacy protection with a Motivated Intruder Test**

1326 Several approaches exist for validating the privacy protection provided by de-identification,
1327 including:

- 1328 • Examining the resulting data files to make sure that no identifying information is
1329 included in file data or metadata.
- 1330 • Conducting a tiger-team analysis to see if outside individuals can perform re-
1331 identification using publicly available datasets or (if warranted) using confidential agency
1332 data.

1333 **4.7.2 Validating data usefulness**

1334 Several approaches exist for validating data usefulness. For example, the results of statistical
1335 calculations performed on both the original dataset and on the de-identified dataset can be
1336 compared to see if the de-identification resulted in significant changes that are unacceptable.
1337 Agencies can also hire tiger-teams to examine the de-identified dataset and see if it can be used
1338 for the intended purpose.

1339 **5 Requirements for De-Identification Tools**

1340 At the present time there are few tools available for de-identification. This section discusses tool
1341 categories and mentions several specific tools.

1342 **5.1 De-Identification Tool Features**

1343 A de-identification tool is a program that involved in the creation of de-identified datasets. De-
1344 identification tools might perform many functions, including:

- 1345 • Detection of identifying information
- 1346 • Calculation of re-identification risk
- 1347 • Performing de-identification
- 1348 • Mapping identifiers to pseudonyms
- 1349 • Providing for the selective revelation of pseudonyms

1350 De-identification tools may handle a variety of data modalities. For example, tools might be
1351 designed for tabular data or for multimedia. Particular tools might attempt to de-identify all data
1352 types, or might be developed for specific modalities. A potential risk of using de-identification
1353 tools is that a tool might be equipped to handle some but not all of the different modalities in a
1354 dataset. For example, a tool might de-identifying the categorical information in a table according
1355 to a de-identification standard, but might not detect or attempt to address the presence of
1356 identifying information in a text field.

1357 **5.2 Data Masking Tools**

1358 Data masking tools are programs that can perform removal or replacement of designated fields in
1359 a dataset while maintaining relationships between tables. These tools can be used to remove
1360 direct identifiers but generally cannot identify or modify quasi-identifiers in a manner consistent
1361 with a privacy policy or risk analysis.

1362 Data masking tools were developed to allow software developers and testers access to datasets
1363 containing realistic data while providing minimal privacy protection. Absent additional controls
1364 or data manipulations, data masking tools should not be used for de-identification of datasets that
1365 are intended for public release.

1366 **6 Evaluation**

1367 Agencies performing de-identification should evaluate the algorithms that they intend to use, the
1368 software that implements the algorithms, and the data that results from the operation of the
1369 software.¹⁰⁵

1370 **6.1 Evaluating Privacy Preserving Techniques**

1371 There has been decades of research in the field of statistical disclosure limitation and de-
1372 identification. As the understanding of statistical disclosure limitation and de-identification have
1373 evolved over time, agencies should not base their technical evaluation of a technique on the mere
1374 fact that the has been published in the peer reviewed literature or that the agency has a long
1375 history of using the technique and has not experienced any problems. Instead, it is necessary to
1376 evaluate proposed techniques in light of the totality of the scientific experience and with regards
1377 to current threats.

1378 Traditional statistical disclosure limitation and de-identification techniques base their risk
1379 assessments, in part, on an expectation of what kinds of data are available to an attacker to
1380 conduct a linkage attack. Where possible, these assumptions should be documented and
1381 published along with a technique description of the privacy-preserving techniques that are used
1382 to transform datasets prior to release, so that they can be reviewed by external experts and the
1383 scientific community.

1384 Because our understanding of privacy technology and the capabilities of privacy attacks are both
1385 rapidly evolving, techniques that have been previously established should be periodically
1386 reviewed. New vulnerabilities may be discovered in techniques that have been previously
1387 accepted. Alternatively, it may be that new techniques are developed that allow agencies to re-
1388 evaluate the tradeoffs that they have made with respect to privacy risk and data usability.

1389 **6.2 Evaluating De-Identification Software**

1390 Once techniques are evaluated and approved, agencies should assure that the techniques are
1391 faithfully executed by their chosen software. Privacy software evaluation should consider the
1392 tradeoff between data usability and privacy protection.

1393 Privacy software evaluation should also seek to detect and minimize the chances of tool error
1394 and user error.

1395 For example, agencies should verify:

- 1396 • That the software properly implements the chosen algorithms.
- 1397 • The software should take into account limitations regarding floating point
1398 representations.
- 1399 • The software does not leak identifying information from source to destination.

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

- 1400 • The software has sufficient usability that it can be operated in efficiently and without
1401 error.

1402 Agencies may also wish to evaluate the performance of the de-identification software, such as:

- 1403 • Efficiency. How long does it take to run on a dataset of a typical size?
1404 • Scalability. How much does it slow down when moving from a dataset of N to 100N?
1405 • Usability. Can users understand the user interface? Can users detect and correct their
1406 errors? Is the documentation sufficient?
1407 • Repeatability. If the tool is run twice on the same dataset, are the results similar? If two
1408 different people run the tool, do they get similar results?

1409 Ideally, software should be able to track the accumulated privacy leakage from multiple data
1410 releases.

1411 **6.3 Evaluating Data Quality**

1412 Finally, agencies should evaluate the quality of the de-identified data to verify that it is sufficient
1413 for the intended use. Approaches for evaluating the data quality include:

- 1414 • Verifying that single variable statistics and two-variable correlations remain relatively
1415 unchanged.
1416 • Verifying that statistical distributions do not incur undue bias as a result of the de-
1417 identification procedure.

1418 **7 Conclusion**

1419 Government agencies can use de-identification technology to make datasets available to
1420 researchers and the general public without compromising the privacy of people contained within
1421 the data.

1422 Currently there are three primary models available for de-identification: agencies can make data
1423 available with traditional de-identification techniques relying on suppression of identifying
1424 information (direct identifiers) and manipulation of information that partially identifying (quasi-
1425 identifiers); agencies can create synthetic datasets; and agencies can make data available through
1426 a query interface. These models can be mixed within a single dataset, providing different kinds
1427 of access for different users or intended uses.

1428 Privacy protection is strongest when agencies employ formal models for privacy protection such
1429 as differential privacy. At the present time there is a small but growing amount of experience
1430 within the government in using these systems. As a result, these systems may result in significant
1431 and at times unnecessary reduction in data quality when compared with traditional de-
1432 identification approaches that do not offer formal privacy guarantees.

1433 Agencies that seek to use de-identification to transform privacy sensitive datasets into dataset
1434 that can be publicly released should take care to establish appropriate governance structures to
1435 support de-identification, data release, and post-release monitoring. Such structures will typically
1436 include a Disclosure Review Board as well as appropriate education, training, and research
1437 efforts.

1438

1439 Appendix A References

1440 A.1 Standards

- 1441 • ASTM E1869-04(2014) Standard Guide for Confidentiality, Privacy, Access, and Data
1442 Security Principles for Health Information Including Electronic Health Records
- 1443 • ISO/IEC 27000:2014 Information technology -- Security techniques -- Information
1444 security management systems -- Overview and vocabulary
- 1445 • ISO/IEC 24760-1:2011 Information technology -- Security techniques -- A framework
1446 for identity management -- Part 1: Terminology and concepts
- 1447 • ISO/TS 25237:2008(E) Health Informatics — Pseudonymization. ISO, Geneva,
1448 Switzerland. 2008.
- 1449 • ISO/IEC 20889 WORKING DRAFT 2016-05-30, Information technology – Security
1450 techniques – Privacy enhancing data de-identification techniques. 2016.

1451 A.2 US Government Publications

- 1452 • Census Confidentiality and Privacy: 1790-2002, US Census Bureau, 2003.
1453 <https://www.census.gov/prod/2003pubs/conmono2.pdf>
- 1454 • Disclosure Avoidance Techniques at the US Census Bureau: Current Practices and
1455 Research, Research Report Series (Disclosure Avoidance #2014-02), Amy Lauger, Billy
1456 Wisniewski, and Laura McKenna, Center for Disclosure Avoidance Research, US
1457 Census. Bureau, September 26, 2014. [https://www.census.gov/srd/CDAR/cdar2014-](https://www.census.gov/srd/CDAR/cdar2014-02_Disc_Avoid_Techniques.pdf)
1458 [02_Disc_Avoid_Techniques.pdf](https://www.census.gov/srd/CDAR/cdar2014-02_Disc_Avoid_Techniques.pdf)
- 1459 • Privacy and Confidentiality Research and the US Census Bureau, Recommendations
1460 Based on a Review of the Literature, Thomas S. Mayer, Statistical Research Division, US
1461 Bureau of the Census. February 7, 2002.
1462 <https://www.census.gov/srd/papers/pdf/rsm2002-01.pdf>
- 1463 • Frequently Asked Questions—Disclosure Avoidance, Privacy Technical Assistance
1464 Center, US Department of Education. October 2012 (revised July 2015)
1465 http://ptac.ed.gov/sites/default/files/FAQ_Disclosure_Avoidance.pdf
- 1466 • *Guidance Regarding Methods for De-identification of Protected Health Information in*
1467 *Accordance with the Health Insurance Portability and Accountability Act (HIPAA)*
1468 *Privacy Rule*, U.S. Department of Health & Human Services, Office for Civil Rights,
1469 November 26, 2012.
1470 [http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-](http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/hhs_deid_guidance.pdf)
1471 [identification/hhs_deid_guidance.pdf](http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/hhs_deid_guidance.pdf)
- 1472 • *OHRP-Guidance on Research Involving Private Information or Biological Specimens*
1473 *(2008)*, Department of Health & Human Services, Office of Human Research Protections
1474 (OHRP), August 16, 2008. <http://www.hhs.gov/ohrp/policy/cdebiol.html>
- 1475 • *Data De-identification: An Overview of Basic Terms*, Privacy Technical Assistance
1476 Center, U.S. Department of Education. May 2013.
1477 http://ptac.ed.gov/sites/default/files/data_deidentification_terms.pdf

- 1478 • *Statistical Policy Working Paper 22 (Second version, 2005)*, Report on Statistical
1479 Disclosure Limitation Methodology, Federal Committee on Statistical Methodology,
1480 December 2005.
- 1481 • [The Data Disclosure Decision, Department of Education](#) (ED) Disclosure Review Board
1482 (DRB), A Product of the Federal CIO Council Innovation Committee. Version 1.0, 2015.
1483 <http://go.usa.gov/xr68F>
- 1484 • National Center for Health Statistics Policy on Micro-data Dissemination, Centers for
1485 Disease Control, July 2002.
1486 https://www.cdc.gov/nchs/data/nchs_microdata_release_policy_4-02a.pdf
- 1487 • National Center for Health Statistics Data Release and Access Policy for Micro-data and
1488 Compressed Vital Statistics File, Centers for Disease Control, April 26, 2011.
1489 http://www.cdc.gov/nchs/nvss/dvs_data_release.htm

1490 A.3 Publications by Other Governments

- 1491 • *Privacy business resource 4: De-identification of data and information*, Office of the
1492 Australian Information Commissioner, Australian Government, April 2014.
1493 [http://www.oaic.gov.au/images/documents/privacy/privacy-resources/privacy-business-
resources/privacy_business_resource_4.pdf](http://www.oaic.gov.au/images/documents/privacy/privacy-resources/privacy-business-
1494 resources/privacy_business_resource_4.pdf)
- 1495 • *Opinion 05/2014 on Anonymisation Techniques*, Article 29 Data Protection Working
1496 Party, 0829/14/EN WP216, Adopted on 10 April 2014
- 1497 • *Anonymisation: Managing data protection risk, Code of Practice 2012*, Information
1498 Commissioner's Office. [https://ico.org.uk/media/for-
organisations/documents/1061/anonymisation-code.pdf](https://ico.org.uk/media/for-
1499 organisations/documents/1061/anonymisation-code.pdf). 108 pages
- 1500 • *The Anonymisation Decision-Making Framework*, Mark Elliot, Elaine Mackey, Kieron
1501 O'Hara and Caroline Tudor, UKAN, University of Manchester, July 2016.
1502 <http://ukanon.net/ukan-resources/ukan-decision-making-framework/>

1503 A.4 Reports and Books:

- 1504 • *Private Lives and Public Policies: Confidentiality and Accessibility of Government*
1505 *Statistics* (1993), George T. Duncan, Thomas B. Jabine, and Virginia A. de Wolf,
1506 Editors; Panel on Confidentiality and Data Access; [Commission on Behavioral and](#)
1507 [Social Sciences and Education](#); [Division of Behavioral and Social Sciences and](#)
1508 [Education](#); National Research Council, 1993. <http://dx.doi.org/10.17226/2122>
- 1509 • *Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk*, Committee on
1510 *Strategies for Responsible Sharing of Clinical Trial Data*, Board on Health Sciences
1511 Policy, Institute of Medicine of the National Academies, The National Academies Press,
1512 Washington, DC. 2015.
- 1513 • P. Doyle and J. Lane, *Confidentiality, Disclosure and Data Access: Theory and Practical*
1514 *Applications for Statistical Agencies*, North-Holland Publishing, Dec 31, 2001

- 1515 • George T. Duncan, Mark Elliot, Juan-José Salazar-Gonzalez, *Statistical Confidentiality: Principles and Practice*, Springer, 2011
1516
- 1517 • Emam, Khaled El and Luk Arbuckle, *Anonymizing Health Data*, O'Reilly, Cambridge,
1518 MA. 2013
- 1519 • Cynthia Dwork and Aaron Roth, *The Algorithmic Foundations of Differential Privacy*
1520 (Foundations and Trends in Theoretical Computer Science). Now Publishers, August 11,
1521 2014. <http://www.cis.upenn.edu/~aaroht/privacybook.html>

1522 **A.5 How-To Articles**

- 1523 • Olivia Angiuli, Joe Blitstein, and Jim Waldo, *How to De-Identify Your Data*,
1524 *Communications of the ACM*, December 2015.
- 1525 • Jörg Drechsler, Stefan Bender, Susanne Rässler, Comparing fully and partially synthetic
1526 datasets for statistical disclosure control in the German IAB Establishment Panel. 2007,
1527 United Nations, Economic Commission for Europe. Working paper, 11, New York, 8 p.
1528 <http://fdz.iab.de/342/section.aspx/Publikation/k080530j05>
- 1529 • Ebaa Fayyumi and B. John Oommen, A survey on statistical disclosure control and
1530 micro-aggregation techniques for secure statistical databases. 2010, *Software Practice*
1531 *and Experience*. 40, 12 (November 2010), 1161-1188. DOI=10.1002/spe.v40:12
1532 <http://dx.doi.org/10.1002/spe.v40:12>
- 1533 • Jingchen Hu, Jerome P. Reiter, and Quanli Wang, Disclosure Risk Evaluation for Fully
1534 Synthetic Categorical Data, *Privacy in Statistical Databases*, pp. 185-199, 2014.
1535 http://link.springer.com/chapter/10.1007%2F978-3-319-11257-2_15
- 1536 • Matthias Templ, Bernhard Meindl, Alexander Kowarik and Shuang Chen, Introduction to
1537 Statistical Disclosure Control (SDC), IHSN Working Paper No. 007, International
1538 Household Survey Network, August 2014.
1539 [http://www.ihsn.org/home/sites/default/files/resources/ihsn-working-paper-007-](http://www.ihsn.org/home/sites/default/files/resources/ihsn-working-paper-007-Oct27.pdf)
1540 [Oct27.pdf](http://www.ihsn.org/home/sites/default/files/resources/ihsn-working-paper-007-Oct27.pdf)
- 1541 • Natalie Shlomo, Statistical Disclosure Control Methods for Census Frequency Tables,
1542 *International Statistical Review* (2007), 75, 2, 199-217.
1543 <https://www.jstor.org/stable/41508461>

1544

1545 **Appendix B Glossary**

1546 Selected terms used in the publication are defined below. Where noted, the definition is sourced
1547 to another publication.

1548 **attribute:** “inherent characteristic.” (ISO 9241-302:2008)

1549 **attribute disclosure:** re-identification event in which an entity learns confidential information
1550 about a data principal, without necessarily identifying the data principal (ISO/IEC 20889
1551 WORKING DRAFT 2 2016-05-27)

1552 **anonymity:** “condition in identification whereby an entity can be recognized as distinct, without
1553 sufficient identity information to establish a link to a known identity” (ISO/IEC 24760-1:2011)

1554 **attacker:** person seeking to exploit potential vulnerabilities of a system

1555 **attribute:** “characteristic or property of an entity that can be used to describe its state,
1556 appearance, or other aspect” (ISO/IEC 24760-1:2011)¹⁰⁶

1557 **brute force attack:** in cryptography, an attack that involves trying all possible combinations to
1558 find a match

1559 **coded:** “1. identifying information (such as name or social security number) that would enable
1560 the investigator to readily ascertain the identity of the individual to whom the private information
1561 or specimens pertain has been replaced with a number, letter, symbol, or combination thereof
1562 (i.e., the code); and 2. a key to decipher the code exists, enabling linkage of the identifying
1563 information to the private information or specimens.”¹⁰⁷

1564 **control:** “measure that is modifying risk. Note: controls include any process, policy, device,
1565 practice, or other actions which modify risk.” (ISO/IEC 27000:2014)

1566 **covered entity:** under HIPAA, a health plan, a health care clearinghouse, or a health care
1567 provider that electronically transmits protected health information (HIPAA Privacy Rule)

1568 **data subjects:** “persons to whom data refer” (ISO/TS 25237:2008)

1569 **data use agreement:** executed agreement between a data provider and a data recipient that
1570 specifies the terms under which the data can be used.

1571 **data universe:** All possible data within a specified domain.

1572 **dataset:** collection of data

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

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

- 1573 **dataset with identifiers:** a dataset that contains information that directly identifies individuals.
- 1574 **dataset without identifiers:** a dataset that does not contain direct identifiers
- 1575 **de-identification:** “general term for any process of removing the association between a set of
1576 identifying data and the data subject” (ISO/TS 25237-2008)
- 1577 **de-identification model:** approach to the application of data de-identification techniques that
1578 enables the calculation of re-identification risk (ISO/IEC 20889 WORKING DRAFT 2 2016-05-
1579 27)
- 1580 **de-identification process:** “general term for any process of removing the association between a
1581 set of identifying data and the data principal” [ISO/TS 25237:2008]
- 1582 **de-identified information:** “records that have had enough PII removed or obscured such that the
1583 remaining information does not identify an individual and there is no reasonable basis to believe
1584 that the information can be used to identify an individual” (SP800-122)
- 1585 **direct identifying data:** “data that directly identifies a single individual” (ISO/TS 25237:2008)
- 1586 **disclosure:** “divulging of, or provision of access to, data” (ISO/TS 25237:2008)
- 1587 **disclosure limitation:** “statistical methods [] used to hinder anyone from identifying an
1588 individual respondent or establishment by analyzing published [] data, especially by
1589 manipulating mathematical and arithmetical relationships among the data.”¹⁰⁸
- 1590 **effectiveness:** “extent to which planned activities are realized and planned results achieved”
1591 (ISO/IEC 27000:2014)
- 1592 **entity:** “item inside or outside an information and communication technology system, such as a
1593 person, an organization, a device, a subsystem, or a group of such items that has recognizably
1594 distinct existence” (ISO/IEC 24760-1:2011)
- 1595 **Federal Committee on Statistical Methodology (FCSM):** “an interagency committee
1596 dedicated to improving the quality of Federal statistics. The FCSM was created by the Office of
1597 Management and Budget (OMB) to inform and advise OMB and the Interagency Council on
1598 Statistical Policy (ICSP) on methodological and statistical issues that affect the quality of Federal
1599 data.” (fscm.sites.usa.gov)
- 1600 **genomic information:** information based on an individual’s genome, such as a sequence of
1601 DNA or the results of genetic testing

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

- 1602 **harm:** “any adverse effects that would be experienced by an individual (i.e., that may be
1603 socially, physically, or financially damaging) or an organization if the confidentiality of PII were
1604 breached” (SP800-122)
- 1605 **Health Insurance Portability and Accountability Act of 1996 (HIPAA):** the primary law in
1606 the United States that governs the privacy of healthcare information
- 1607 **HIPAA:** see *Health Insurance Portability and Accountability Act of 1996*
- 1608 **HIPAA Privacy Rule:** “establishes national standards to protect individuals’ medical records
1609 and other personal health information and applies to health plans, health care clearinghouses, and
1610 those health care providers that conduct certain health care transactions electronically” (HIPAA
1611 Privacy Rule, 45 CFR 160, 162, 164)
- 1612 **identification:** “process of using claimed or observed attributes of an entity to single out the
1613 entity among other entities in a set of identities” (ISO/TS 25237:2008)
- 1614 **identified information:** information that explicitly identifies an individual
- 1615 **identifier:** “information used to claim an identity, before a potential corroboration by a
1616 corresponding authenticator” (ISO/TS 25237:2008)
- 1617 **imputation:** “a procedure for entering a value for a specific data item where the response is
1618 missing or unusable.” (OECD Glossary of Statistical Terms)
- 1619 **inference:** “refers to the ability to deduce the identity of a person associated with a set of data
1620 through “clues” contained in that information. This analysis permits determination of the
1621 individual’s identity based on a combination of facts associated with that person even though
1622 specific identifiers have been removed, like name and social security number” (ASTM E1869¹⁰⁹)
- 1623 **k-anonymity:** a technique “to release person-specific data such that the ability to link to other
1624 information using the quasi-identifier is limited.”¹¹⁰ k-anonymity achieves this through
1625 suppression of identifiers and output perturbation.
- 1626 **l-diversity:** a refinement to the k-anonymity approach which assures that groups of records
1627 specified by the same identifiers have sufficient diversity to prevent inferential disclosure¹¹¹

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

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

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

1628 **masking:** the process of systematically removing a field or replacing it with a value in a way that
 1629 does not preserve the analytic utility of the value, such as replacing a phone number with
 1630 asterisks or a randomly generated pseudonym¹¹²

1631 **noise:** “a convenient term for a series of random disturbances borrowed through communication
 1632 engineering, from the theory of sound. In communication theory noise results in the possibility of
 1633 a signal sent, x , being different from the signal received, y , and the latter has a probability
 1634 distribution conditional upon x . If the disturbances consist of impulses at random intervals it is
 1635 sometimes known as “shot noise.” (OECD Glossary of Statistical Terms)

1636 **non-deterministic noise:** a random value that cannot be predicted

1637 **personal identifier:** “information with the purpose of uniquely identifying a person within a
 1638 given context” (ISO/TS 25237:2008)

1639 **personal data:** “any information relating to an identified or identifiable natural person (*data*
 1640 *subject*)” (ISO/TS 25237:2008)

1641 **personally identifiable information (PII):** “Any information about an individual maintained by
 1642 an agency, including (1) any information that can be used to distinguish or trace an individual’s
 1643 identity, such as name, social security number, date and place of birth, mother’s maiden name, or
 1644 biometric records; and (2) any other information that is linked or linkable to an individual, such
 1645 as medical, educational, financial, and employment information.”¹¹³ (SP800-122)

1646 **privacy:** “freedom from intrusion into the private life or affairs of an individual when that
 1647 intrusion results from undue or illegal gathering and use of data about that individual” (ISO/IEC
 1648 2382-8:1998, definition 08-01-23)

1649 **protected health information (PHI):** “individually identifiable health information: (1) Except
 1650 as provided in paragraph (2) of this definition, that is: (i) Transmitted by electronic media;
 1651 (ii) Maintained in electronic media; or (iii) Transmitted or maintained in any other form or
 1652 medium. (2) *Protected health information* excludes individually identifiable health information
 1653 in: (i) Education records covered by the Family Educational Rights and Privacy Act, as
 1654 amended, [20 U.S.C. 1232g](#); (ii) Records described at [20 U.S.C. 1232g\(a\)\(4\)\(B\)\(iv\)](#); and
 1655 (iii) Employment records held by a covered entity in its role as employer.” (HIPAA Privacy
 1656 Rule, 45 CFR 160.103)

1657 **pseudonymization:** a particular type of de-identification that both removes the association with
 1658 a data subject and adds an association between a particular set of characteristics relating to the
 1659 data subject and one or more pseudonyms.¹¹⁴ Typically, pseudonymization is implemented by

¹¹² El Emam, Khaled and Luk Arbuckle, *Anonymizing Health Data*, O’Reilly, Cambridge, MA. 2013

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

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

- 1660 replacing direct identifiers with a pseudonym, such as a randomly generated value.
- 1661 **pseudonym:** “personal identifier that is different from the normally used personal identifier.”
1662 (ISO/TS 25237:2008)
- 1663 **quasi-identifier:** information that can be used to identify an individual through association with
1664 other information
- 1665 **recipient:** “natural or legal person, public authority, agency or any other body to whom data are
1666 disclosed” (ISO/TS 25237:2008)
- 1667 **re-identification:** general term for any process that re-establishes the relationship between
1668 identifying data and a data subject
- 1669 **re-identification risk:** the risk that de-identified records can be re-identified. Re-identification
1670 risk is typically reported as the percentage of records in a dataset that can be re-identified.
- 1671 **risk:** “effect of uncertainty on objectives. Note: risk is often expressed in terms of a combination
1672 of the consequences of an event (including changes in circumstances) and the associated
1673 likelihood of occurrence.” (ISO/IEC 27000:2014)
- 1674 **synthetic data generation:** a process in which seed data are used to create artificial data that has
1675 some of the statistical characteristics as the seed data
- 1676

1677 Appendix C Specific De-Identification Tools

1678 This appendix provides a list of de-identification tools.

1679 NOTE

1680 Specific products and organizations identified in this report were used in order to perform the
 1681 evaluations described. In no case does such identification imply recommendation or
 1682 endorsement by the National Institute of Standards and Technology, nor does it imply that
 1683 identified are necessarily the best available for the purpose.

1684 C.1 Tabular Data

1685 Most de-identification tools designed for tabular data implement the k-Anonymity model. Many
 1686 directly implement the HIPAA Privacy Rule's Safe Harbor standard. Tools that are currently
 1687 available include:

1688 **AnonTool** is a German-language program that supports the k-anonymity framework.
 1689 http://www.tmf-ev.de/Themen/Projekte/V08601_AnonTool.aspx

1690 **ARX** is an open source data de-identification tool written in Java that implements a variety of
 1691 academic de-identification models, including k-anonymity, Population uniqueness,¹¹⁵ k-Map,
 1692 Strict-average risk, ℓ -Diversity,¹¹⁶ t-Closeness,¹¹⁷ δ -Disclosure privacy,¹¹⁸ and δ -presence.
 1693 <http://arx.deidentifier.org/>

1694 **Cornell Anonymization Toolkit** is an interactive tool that was developed by the Computer
 1695 Science Department at Cornell University¹¹⁹ for performing de-identification. It can perform data
 1696 generalization, risk analysis, utility evaluation, sensitive record manipulation, and visualization
 1697 functions. <https://sourceforge.net/projects/anony-toolkit/>

1698 **Open Anonymizer** implements the k-anonymity framework.
 1699 <https://sourceforge.net/projects/openanonymizer/>

1700 **Privacy Analytics Eclipse** is a comprehensive de-identification platform that can de-identify
 1701 multiple linked tabular datasets to HIPAA or other de-identification standards. The program runs

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

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

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

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

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

1702 on Apache SPARK to allow de-identification of massive datasets, such as those arising in
1703 medical research. <http://www.privacy-analytics.com/software/privacy-analytics-core/>

1704 **μ-ARGUS** was developed by Statistics Netherlands for microdata release. The program was
1705 originally written in Visual Basic and was rewritten into C/C++ for an Open Source release. The
1706 program runs on Windows and Linux. <http://neon.vb.cbs.nl/casc/mu.htm>

1707 **sdcMicro** is a package for the popular open source R statistical platform that implements a
1708 variety of statistical disclosure controls. A full tutorial is available, as are prebuilt binaries for
1709 Windows and OS X. <https://cran.r-project.org/web/packages/sdcMicro/>

1710 **SECRET**A, a tool for evaluating and comparing anonymizations. According to the website,
1711 “SECRETA supports Incognito, Cluster, Top-down, and Full subtree bottom-up algorithms for
1712 datasets with relational attributes, and COAT, PCTA, Apriori, LRA and VPA algorithms for
1713 datasets with transaction attributes. Additionally, it supports the RMERGER, TMERGER, and
1714 RTMERGER bounding methods, which enable the anonymization of RT-datasets by combining
1715 two algorithms, each designed for a different attribute type (e.g., Incognito for relational
1716 attributes and COAT for transaction attributes).” <http://users.uop.gr/~poulis/SECRET>A/

1717 **UTD Anonymization Toolbox** is an open source tool developed by the University of Texas
1718 Dallas Data Security and Privacy Lab using funding provided by the National Institutes of
1719 Health, the National Science Foundation, and the Air Force Office of Scientific Research.

1720 **C.2 Free Text**

1721 **BoB, a best-of-breed automated text de-identification system for VHA clinical**
1722 **documents**,¹²⁰ developed by the Meystre Lab at the University of Utah School of Medicine.
1723 <http://meystrelab.org/automated-ehr-text-de-identification/>

1724 **MITRE Identification Scrubber Toolkit (MIST)** is an open source tool for de-identifying free
1725 format text. <http://mist-deid.sourceforge.net>

1726 **Privacy Analytics Lexicon** performs automated de-identification of unstructured data (text).
1727 <http://www.privacy-analytics.com/software/privacy-analytics-lexicon/>

1728 **C.3 Multimedia**

1729 **DicomCleaner** is an open source tool that removes identifying information from medical
1730 imagery in the DICOM format. DicomCleaner. The program can remove both metadata from the
1731 DICOM file and black out identifying information that has been “burned in” to the image area.
1732 DicomCleaner can perform redaction directly of compressed JPEG blocks so that the medical
1733 image does not need to be decompressed and re-compressed, a procedure that can introduce
1734 artifacts. <http://www.dclunie.com/pixelmed/software/webstart/DicomCleanerUsage.html>

¹²⁰ [BoB, a best-of-breed automated text de-identification system for VHA clinical documents](http://www.dclunie.com/pixelmed/software/webstart/DicomCleanerUsage.html). 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.