

# Withdrawn Draft

## Warning Notice

The attached draft document has been withdrawn and is provided solely for historical purposes. It has been superseded by the document identified below.

**Withdrawal Date** November 15, 2022

**Original Release Date** December 15, 2016

## Superseding Document

**Status** 3<sup>rd</sup> Public Draft (3PD)

**Series/Number** NIST Special Publication 800-188

**Title** De-Identifying Government Data Sets

**Publication Date** November 2022

**DOI** <https://doi.org/10.6028/NIST.SP.800-188.3pd>

**CSRC URL** <https://csrc.nist.gov/publications/detail/sp/800-188/draft>

## Additional Information

---

---

4 **De-Identifying Government Datasets**

---

5  
6  
7

8 Simson L. Garfinkel  
9  
10  
11  
12  
13  
14  
15  
16

---

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

---

20 **De-Identifying Government Datasets**

21  
22  
23 **Simson L. Garfinkel**  
24 *Information Access Division*  
25 *Information Technology Laboratory*  
26  
27  
28

29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
  
45  
46  
47  
48  
49  
50  
51  
52  
  
December 2016



U.S. Department of Commerce  
*Penny Pritzker, Secretary*

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

53

## Authority

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

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

67 National Institute of Standards and Technology Special Publication 800-188  
68 Natl. Inst. Stand. Technol. Spec. Publ. 800-188, 85 pages (December 2016)  
69 CODEN: NSPUE2

70

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

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

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

84

85 ***Public comment period: December 15, 2016 through December 31, 2016***

86 National Institute of Standards and Technology  
87 Attn: Information Access Division, Information Technology Laboratory  
88 100 Bureau Drive (Mail Stop 8940) Gaithersburg, MD 20899-8940  
89 Email: sp800-188-draft@nist.gov

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

91

92  
93

## Reports on Computer Systems Technology

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

102

### Abstract

103 De-identification is a process that is applied to a dataset to reduce the risk of linking information  
104 revealed in the dataset to specific individuals. Government agencies can use de-identification to  
105 reduce the privacy risk associated with collecting, processing, archiving, distributing or publishing  
106 government data. Previously NIST published NISTIR 8053, *De-Identification of Personal*  
107 *Information*, which provided a survey of de-identification and re-identification techniques. This  
108 document provides specific guidance to government agencies that wish to use de-identification.  
109 Before using de-identification, agencies should evaluate their goals in using de-identification and  
110 the potential risks that de-identification might create. Agencies should decide upon a de-  
111 identification release model, such as publishing de-identified data, publishing synthetic data based  
112 on identified data, or providing a query interface that incorporates de-identification of the  
113 identified data. Agencies can create a Disclosure Review Board to oversee the process of de-  
114 identification; they can also adopt a de-identification standard with measurable performance  
115 levels. Several specific techniques for de-identification are available, including de-identification  
116 by removing identifiers and transforming quasi-identifiers and the use of formal privacy models.  
117 People performing de-identification generally use special-purpose software tools to perform the  
118 data manipulation and calculate the likely risk of re-identification. However, not all tools that  
119 merely mask personal information provide sufficient functionality for performing de-  
120 identification. This document also includes an extensive list of references, a glossary, and a list of  
121 specific de-identification tools, although the mention of these tools is only to be used to convey  
122 the range of tools currently available, and is not intended to imply recommendation or endorsement  
123 by NIST.

124

### Keywords

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

128

129

## Acknowledgements

130 The author wishes to thank the US Census Bureau for its help in researching and preparing this  
131 publication, with specific thanks to John Abowd, Ron Jarmin, Christa Jones, and Laura  
132 McKenna. The author would also like to thank Luk Arbuckle, Andrew Baker, Daniel Barth-  
133 Jones, Khaled El Emam, Robert Gellman, Tom Krenzke, Bradley Malin and John Moehrke for  
134 providing comments on previous drafts and valuable insight in creating this publication.

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

140

## Audience

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

146  
147  
148  
149  
150  
151  
152  
153  
154  
155  
156  
157  
158  
159  
160  
161  
162  
163  
164  
165  
166  
167  
168  
169  
170  
171  
172  
173  
174  
175  
176  
177  
178  
179  
180  
181

**Table of Contents**

**Executive Summary ..... vii**

**1 Introduction ..... 1**

    1.1 Document Purpose and Scope ..... 4

    1.2 Intended Audience ..... 5

    1.3 Organization ..... 5

**2 Introducing De-Identification ..... 6**

    2.1 Historical Context ..... 6

    2.2 NISTIR 8053..... 8

    2.3 Terminology..... 9

**3 Governance and Management of Data De-Identification ..... 14**

    3.1 Identifying Goals and Intended Uses of De-Identification ..... 14

    3.2 Evaluating Risks Arising from De-Identified Data Releases ..... 15

        3.2.1 Probability of Re-Identification..... 16

        3.2.2 Adverse Impacts Resulting from Re-Identification ..... 19

        3.2.3 Impacts other than re-identification ..... 20

        3.2.4 Remediation ..... 21

    3.3 Data Life Cycle ..... 21

    3.4 Data Sharing Models..... 25

    3.5 The Five Safes ..... 27

    3.6 Disclosure Review Boards ..... 28

    3.7 De-Identification Standards..... 32

        3.7.1 Benefits of Standards ..... 32

        3.7.2 Prescriptive De-Identification Standards ..... 32

        3.7.3 Performance-Based De-Identification Standards..... 33

    3.8 Education, Training and Research..... 34

**4 Technical Steps for Data De-Identification ..... 35**

    4.1 Determine the Privacy, Data Usability, and Access Objectives..... 35

    4.2 Conducting a Data Survey ..... 36

    4.3 De-identification by removing identifiers and transforming quasi-identifiers..... 38

        4.3.1 Removing or Transformation of Direct Identifiers..... 40

        4.3.2 De-Identifying Numeric Quasi-Identifiers..... 41

        4.3.3 De-identifying dates..... 43

        4.3.4 De-identifying geographical locations..... 44

        4.3.5 De-identifying genomic information ..... 45

182 4.3.6 Challenges Posed by Aggregation Techniques ..... 46

183 4.3.7 Challenges posed by High-Dimensionality Data ..... 47

184 4.3.8 Challenges Posed by Linked Data ..... 47

185 4.3.9 Challenges Posed by Composition ..... 48

186 4.3.10 Post-Release Monitoring ..... 48

187 4.4 Synthetic Data ..... 48

188 4.4.1 Partially Synthetic Data ..... 49

189 4.4.2 Fully Synthetic Data ..... 50

190 4.4.3 Synthetic Data with Validation ..... 51

191 4.4.4 Synthetic Data and Open Data Policy ..... 51

192 4.4.5 Creating a synthetic dataset with differential privacy ..... 52

193 4.5 De-Identifying with an interactive query interface ..... 54

194 4.6 Validating a de-identified dataset ..... 55

195 4.6.1 Validating data usefulness ..... 55

196 4.6.2 Validating privacy protection ..... 55

197 **5 Software Requirements, Evaluation and Validation..... 57**

198 5.1 Evaluating Privacy Preserving Techniques ..... 57

199 5.2 De-Identification Tools ..... 57

200 5.2.1 De-Identification Tool Features ..... 57

201 5.2.2 Data Provenance and File Formats ..... 58

202 5.2.3 Data Masking Tools ..... 58

203 5.3 Evaluating De-Identification Software ..... 58

204 5.4 Evaluating Data Quality ..... 59

205 **6 Conclusion ..... 60**

206 **7 References ..... 61**

207 7.1 Standards ..... 61

208 7.2 US Government Publications ..... 61

209 7.3 Publications by Other Governments ..... 63

210 7.4 Reports and Books: ..... 63

211 7.5 How-To Articles ..... 64

212 7.6 Glossary ..... 65

213 7.7 Specific De-Identification Tools ..... 70

214 7.7.1 Tabular Data ..... 71

215 7.7.2 Free Text ..... 72

216 7.7.3 Multimedia ..... 72

217



218  
219  
220  
221  
222  
223  
224  
225  
226  
227  
228  
229

**List of Figures**

Figure 1 Michener et al.'s view of the data lifecycle is a true cycle, with analysis guiding future collection. .... 22

Figure 2 Chisholm's view of the data lifecycle is a linear process with a branching point after data usage. .... 23

Figure 3 Lifecycle model for government data releases, from Altman et al. .... 23

Figure 4 Conceptual diagram of the relationship between post-transformation identifiability, level of expected harm, and suitability of selected privacy controls for a data release. From Altman et al. .... 24

## 230 **Executive Summary**

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

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

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

- 246 ● For four decades, official statistical agencies have researched and investigated methods  
 247 broadly termed *Statistical Disclosure Limitation (SDL)* or *Statistical Disclosure*  
 248 *Control*<sup>3,4</sup>
- 249 ● In the 1990s there was an increase in the unrestricted release of microdata, or individual  
 250 responses from surveys or administrative records. Initially these releases merely stripped  
 251 obviously identifying information such as names and social security numbers (what are  
 252 now called direct identifiers). Following some releases, researchers discovered that it was  
 253 possible to re-identify individual data by triangulating with some of the remaining  
 254 identifiers (now called quasi-identifiers or indirect identifiers).<sup>5</sup> The result of this  
 255 research was the development of the k-anonymity model for protecting privacy,<sup>6</sup> which is

<sup>1</sup> 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.

<sup>2</sup> 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.

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

<sup>4</sup> 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/>

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

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

256 reflected in the Health and Human Services guidance regarding the HIPAA Privacy  
257 Rule.<sup>7</sup>

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

264 In recognition of both the growing importance of de-identification within the US Government  
265 and the paucity of efforts addressing de-identification as a holistic field, NIST began research in  
266 this area in 2015. As part of that investigation, NIST researched and published NIST Interagency  
267 Report 8053, *De-Identification of Personal Information*.<sup>9</sup>

268 Since the publication of NISTIR 8053, NIST has continued research in the area of de-  
269 identification. NIST met with de-identification experts within and outside the United States  
270 Government, convened a Government Data De-Identification Stakeholder's Meeting in June  
271 2016, and conducted an extensive literature review.

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

276 Before engaging in de-identification, agencies should clearly articulate their goals in performing  
277 the de-identification, the kinds of data that they intend to de-identify and the uses that they  
278 envision for the de-identified data. Agencies should also conduct a risk assessment that takes into  
279 account the potential adverse actions that might result from the release of the de-identified data;  
280 this risk assessment should include analysis of risk that might result from the data being re-  
281 identified and risk that might result from the mere release of the de-identified data itself. For  
282 example, improperly de-identified data might be used to identify vulnerable individuals or  
283 groups. The release of potentially harmful information might result in reputational risk to an  
284 agency, potentially threatening its mission.

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

---

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

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

<sup>8</sup> 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](http://dx.doi.org/10.1007/11787006_1)

<sup>9</sup> NISTIR 8053, *De-Identification of Personal Data*, Simson Garfinkel, September 2015, National Institute of Standards and Technology, Gaithersburg, MD

287 organization and representatives of the organization’s leadership. The DRB should evaluate  
288 applications for de-identification that describe the data to be released, the techniques that will be  
289 used to minimize the risk of disclosure, and how the effectiveness of those techniques will be  
290 evaluated.

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

- 292 ● **The Release and Forget model:**<sup>10</sup> The de-identified data may be released to the public,  
293 typically by being published on the Internet.
- 294 ● **The Data Use Agreement (DUA) model:** The de-identified data may be made available  
295 to qualified users under a legally binding data use agreement that details what can and  
296 cannot be done with the data. Under this model, the information that is present in the  
297 released data can be tailored to the specific needs, capabilities, and risk profile of the  
298 intended data users.
- 299 ● **The Synthetic Data with Verification Model:** Statistical Disclosure Limitation  
300 techniques are applied to the original dataset and used to create a synthetic dataset that  
301 reflects the statistical properties of the original dataset, but which does not contain  
302 disclosing information. The synthetic dataset is released, either publicly or to vetted  
303 researchers.
- 304 ● **The Enclave model:**<sup>11,12</sup> The de-identified data may be kept in a physical or virtual  
305 segregated enclave that restricts the export of the original data, and instead accepts  
306 queries from qualified researchers, runs the queries on the de-identified data, and  
307 responds with results. Enclaves can also support features for audit and accountability.

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

318 Agencies can create or adopt standards to guide those performing de-identification. The  
319 standards can specify disclosure techniques, or they can specify privacy guarantees that the de-  
320 identified data must uphold. There are many techniques available for de-identifying data; most of

---

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

<sup>11</sup> Ibid.

<sup>12</sup> 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

321 these techniques are specific to a particular modality. Some techniques are based on ad-hoc  
322 procedures, while others are based on formal privacy models that make it possible to rigorously  
323 calculate the amount of data manipulation required of the data to assure a particular level of  
324 privacy protection.

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

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

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

341 Today there are several non-commercial, open source programs for performing de-identification  
342 but only a few commercial products. Currently there are no performance standards, certification,  
343 or third-party testing programs available for de-identification software.

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

## 347 **1 Introduction**

348 The US Government collects, maintains, and uses many kinds of datasets. Every federal agency  
349 creates and maintains internal datasets that are vital for fulfilling its mission, such as delivering  
350 services to taxpayers or ensuring regulatory compliance. Additionally, there are 13 Federal  
351 statistical agencies whose primary passion is the collection, compilation, processing or analysis  
352 of information for statistical purposes.”<sup>13</sup>

353 Increasingly these datasets are being made available to the public. Many of these datasets are  
354 openly published to promote commerce, support scientific research, and generally promote the  
355 public good. Other datasets contain sensitive data elements and, thus, are only made available on  
356 a limited basis. Some datasets are so sensitive that they cannot be made publicly available at all,  
357 but can be made available on a limited basis in protected enclaves. In some cases agencies may  
358 choose to release summary statistics, or create synthetic datasets that resemble the original data  
359 but which have less<sup>14</sup> disclosure risk.

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

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

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

- 375 ● US Code Title 13, Section 9, which governs confidentiality of information provided to  
376 the Census Bureau, prohibits “any publication whereby the data furnished by any

---

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

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

<sup>15</sup> 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.

- 377 particular establishment or individual under this title can be identified.”
- 378 ● The release of personal information by the government is generally covered by the  
379 Privacy Act of 1974,<sup>16</sup> which recognizes that disclosure of records for statistical purposes  
380 is acceptable if the data is not “individually identifiable.”<sup>17</sup>
- 381 ● Title V – Confidential Information Protection and Statistical Efficiency Act of 2002 of  
382 the E-Government Act of 2002 (CIPSEA),<sup>18</sup> states that “[d]ata or information acquired  
383 by an agency under a pledge of confidentiality for exclusively statistical purposes shall  
384 not be disclosed by an agency in identifiable form, for any use other than an exclusively  
385 statistical purpose, except with the informed consent of the respondent.”<sup>19</sup> The Act  
386 further requires that federal statistical agencies “establish appropriate administrative,  
387 technical, and physical safeguards to insure the security and confidentiality of records  
388 and to protect against any anticipated threats or hazards to their security or integrity  
389 which could result in substantial harm, embarrassment, inconvenience, or unfairness to  
390 any individual on whom information is maintained.”<sup>20</sup>
- 391 ● On January 21, 2009, President Obama issued a memorandum to the heads of executive  
392 departments and agencies calling for US government to be transparent, participatory and  
393 collaborative.<sup>21,22</sup> This was followed on December 8, 2009, by the Open Government  
394 Directive,<sup>23</sup> which called on the executive departments and agencies “to expand access to  
395 information by making it available online in open formats. With respect to information,  
396 the presumption shall be in favor of openness (to the extent permitted by law and subject  
397 to valid privacy, confidentiality, security, or other restrictions).”
- 398 ● On February 22, 2013, the White House Office of Science and Technology Policy  
399 (OSTP) directed Federal agencies with over \$100 million in annual research and  
400 development expenditures to develop plans to provide for increased public access to  
401 digital scientific data. Agencies were instructed to “[m]aximize access, by the general  
402 public and without charge, to digitally formatted scientific data created with Federal  
403 funds, while: i) protecting confidentiality and personal privacy, ii) recognizing

---

<sup>16</sup> Public Law 93-579, 88 Stat. 1896, 5 U.S.C. § 552a.

<sup>17</sup> 5 USC 552a(b)(5)

<sup>18</sup> Pub.L. 107-347, 116 Stat. 2899, 44 U.S.C. § 101, H.R. 2458/S. 803

<sup>19</sup> Public Law 107-347 § 512 (b)(1), Dec. 17, 2002

<sup>20</sup> See Title V—Confidentiality Information Protection and Statistical Efficiency, Public Law 107-347, Dec 17, 2002.

<sup>21</sup> Barack Obama, *Transparency and Open Government*, The White House, January 21, 2009.

<sup>22</sup> 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](https://www.whitehouse.gov/sites/default/files/omb/assets/memoranda_fy2009/m09-12.pdf)

<sup>23</sup> OMB Memorandum M-10-06, *Open Government Directive*, December 8, 2009, M-10-06.

404 proprietary interests, business confidential information, and intellectual property rights  
 405 and avoiding significant negative impact on intellectual property rights, innovation, and  
 406 U.S. competitiveness, and iii) preserving the balance between the relative value of long-  
 407 term preservation and access and the associated cost and administrative burden.”<sup>24</sup>

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

410 Minimizing privacy risk is not an absolute goal of Federal laws and regulations. Instead, privacy  
 411 risk is weighed against other factors, such as transparency, accountability, and the opportunity  
 412 for public good. This is why, for example, personally identifiable information collected by the  
 413 Census Bureau remains confidential for 72 years, and is then transferred to the National Archives  
 414 and Records Administration where it is released to the public.<sup>25</sup> Guidance from the US  
 415 Department of Health and Human Services (HHS) on the HIPAA de-identification standard  
 416 notes that “[b]oth methods [the safe harbor and expert determination methods for de-  
 417 identification], even when properly applied, yield de-identified data that retains some risk of  
 418 identification. Although the risk is very small, it is not zero, and there is a possibility that de-  
 419 identified data could be linked back to the identity of the patient to which it corresponds.”<sup>26</sup>

420 *De-identification* is a term used within the US Government to describe the removal,  
 421 modification, or obfuscation of personal information from data that are collected, used, archived,  
 422 and shared, with the goal of preventing or limiting informational risks to individuals, protected  
 423 groups, and establishments.<sup>27</sup> De-identification is not a single technique, but a collection of  
 424 approaches, algorithms, and tools that can be applied to different kinds of data with differing  
 425 levels of effectiveness. In general, the potential risk to privacy posed by a dataset’s release  
 426 decreases as more aggressive de-identification techniques are employed, but data quality of the  
 427 de-identified dataset decreases as well.

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

---

<sup>24</sup> 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.

<sup>25</sup> The “72-Year Rule,” US Census Bureau, [https://www.census.gov/history/www/genealogy/decennial\\_census\\_records/the\\_72\\_year\\_rule\\_1.html](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.

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

<sup>27</sup> 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.



431 *Utility* is traditionally in economics defined as “the satisfaction derived from consumption of a  
432 good or service.”<sup>28</sup> *Data utility* therefore refers to the value that data users can derive from data  
433 in general.

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

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

446 Some users of de-identified data may be able to use the data to make inferences about private  
447 facts regarding the data subjects; they may even be able to re-identify the data subjects—that is,  
448 to undo the privacy guarantees of de-identification. Agencies that release data should understand  
449 what data they are releasing, what other data may already be publicly or privately available, and  
450 the risk of re-identification. Agencies should aim to make an informed decision by systematically  
451 weighing the risks against the benefits and choosing de-identification techniques and data release  
452 models that are tailored to their analysis of the risks and benefits. In addition, when telling  
453 individuals their information will be de-identified, agencies should also disclose that privacy  
454 risks may remain despite de-identification.

455 Planning is essential for successful de-identification and data release. Data management and  
456 privacy protection should be an integrated part of scientific research. This planning will include  
457 research design, data collection, protection of identifiers, disclosure analysis, and data sharing  
458 strategy. In an operational environment, this planning includes a comprehensive analysis of the  
459 purpose of the data release and the expected use of the released data, the privacy-related risks,  
460 the privacy protecting controls, the appropriateness of various privacy controls given the risks  
461 and intended uses, and the ways that those controls could fail.

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

## 465 **1.1 Document Purpose and Scope**

466 This document provides guidance regarding the selection, use and evaluation of de-identification

---

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

467 techniques for US government datasets. It also provides a framework that can be adapted by  
468 Federal agencies to frame the governance of de-identification processes. The ultimate goal of  
469 this document is to reduce disclosure risk that might result from an intentional data release.

## 470 **1.2 Intended Audience**

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

## 476 **1.3 Organization**

477 The remainder of this publication is organized as follows: Section 2, “Introducing De-  
478 Identification”, presents a background on the science and terminology of de-identification.  
479 Section 3, “Governance and Management of Data De-Identification,” provides guidance to  
480 agencies on the establishment or improvement to a program that makes privacy-sensitive data  
481 available to researchers and the general public. Section 4, “Technical Steps for Data De-  
482 Identification,” provides specific technical guidance for performing de-identification using a  
483 variety of mathematical approaches. Section 5, “Requirements for De-Identification Tools,”  
484 provides a recommended set of features that should be in de-identification tools; this information  
485 may be useful for potential purchasers or developers of such software. Section 6, “Evaluation,”  
486 provides information for evaluating both de-identification tools and de-identified datasets. This  
487 publication concludes with Section 7, “Conclusion.”

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

490

## 491 2 Introducing De-Identification

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

493 As long as information derived from personal data remains in a de-identified dataset, there exists  
494 the possibility that the de-identified data might reveal attributes related to specific individuals, or  
495 even that specific de-identified records could be linked back to specific individuals. When this  
496 happens, the privacy protection provided by de-identification is compromised. Even if a specific  
497 individual cannot be matched to a specific data record, de-identified data can be used to improve  
498 the accuracy of inferences regarding individuals whose de-identified data are in the dataset. This  
499 so-called *inference risk* cannot be eliminated if there is any information content in the de-  
500 identified data, but it can be minimized. Thus, the decision of how or if to de-identify data should  
501 thus be made in conjunction with decisions of how the de-identified data will be used, shared or  
502 released.

503 De-identification is especially important for government agencies, businesses, and other  
504 organizations that seek to make data available to outsiders. For example, significant medical  
505 research resulting in societal benefit is made possible by the sharing of de-identified patient  
506 information under the framework established by the Health Insurance Portability and  
507 Accountability Act (HIPAA) Privacy Rule, the primary US regulation providing for privacy of  
508 medical records. Agencies may also be required to de-identify records as part of responding to a  
509 Freedom of Information Act (FOIA) request.<sup>29</sup>

### 510 2.1 Historical Context

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

- 512 • For four decades, official statistical agencies have researched and investigated methods  
513 broadly termed *Statistical Disclosure Limitation (SDL)* or *Statistical Disclosure*  
514 *Control*<sup>30,31,32</sup> Most of these methods were created to allow the release of statistical tables  
515 and *public use files (PUF)* that allow users to learn factual information or perform  
516 original research, while protecting the privacy of the individuals in the dataset. SDL is  
517 widely used in contemporary statistical reporting.

---

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

<sup>30</sup> T. Dalenius, Towards a methodology for statistical disclosure control. *Statistik Tidskrift* 15, pp. 429-222, 1977

<sup>31</sup> 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/>

<sup>32</sup> George T. Duncan, Mark Elliot, Gonzalez Juan Jose Salazar. *Statistical Confidentiality: Principles and Practice*; Springer Science 2011.

- 518       ● In the 1990s, there was an increase in the release of *microdata* files for public use, with  
 519 individual responses from surveys or administrative records. Initially these releases  
 520 merely stripped obviously identifying information such as names and social security  
 521 numbers (what are now called *direct identifiers*). Following some releases, researchers  
 522 discovered that it was possible to re-identify individuals' data by triangulating with some  
 523 of the remaining identifiers (called *quasi-identifiers* or *indirect identifiers*<sup>33</sup>). The result  
 524 of this research was the identification of the k-anonymity model for protecting  
 525 privacy,<sup>34,35, 36, 37</sup> which is reflected in the HIPAA Privacy Rule. Software that measures  
 526 privacy risk using k-anonymity is often used to allow the sharing of medical microdata.  
 527 This intellectual tradition is typically called *de-identification*, although this document  
 528 uses the word de-identification to describe all three intellectual traditions.
- 529       ● In the 2000s, research in theoretical computer science and cryptography developed the  
 530 theory of *differential privacy*,<sup>38</sup> which is based on a mathematical definition of the  
 531 privacy loss to an individual resulting from queries on a database containing that  
 532 individual's personal information. Differential privacy is termed a *formal model for*  
 533 *privacy protection* because its definitions of privacy and privacy loss are based on  
 534 mathematical proofs.<sup>39</sup> Note that this doesn't mean that algorithms implementing  
 535 differential privacy cannot result in increased privacy risk. Instead, it means that the  
 536 amount of privacy risk that results from the use of these algorithms can be  
 537 mathematically bounded. These mathematical limits on privacy risk have created  
 538 considerable interest in differential privacy in academia, commerce and business, but to  
 539 date only a few systems employing differential privacy have been operationally deployed.
- 540 Separately, during the first decade of the 21<sup>st</sup> century there was a growing awareness within the  
 541 US Government about the risks that could result from the improper handling and inadvertent

---

<sup>33</sup> Dalenius, Finding a Needle in a Haystack, or Identifying Anonymous Census Records, *Journal of Official Statistics* 2:3, 329-336, 1986.

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

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

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

<sup>37</sup> 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>

<sup>38</sup> 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](http://dx.doi.org/10.1007/11787006_1)

<sup>39</sup> 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>

542 release of personal identifying and financial information. This realization, combined with a  
543 growing number of inadvertent data disclosures within the US government, resulted in President  
544 George Bush signing Executive Order 13402 establishing an Identity Theft Task Force on May  
545 10, 2006.<sup>40</sup> A year later the Office of Management and Budget issued Memorandum M-07-16<sup>41</sup>  
546 which required Federal agencies to develop and implement breach notification policies. As part  
547 of this effort, NIST issued Special Publication 800-122, *Guide to Protecting the Confidentiality*  
548 *of Personally Identifiable Information (PII)*.<sup>42</sup> These policies and documents had the specific  
549 goal of limiting the accessibility of information that could be directly used for identity theft, but  
550 did not create a framework for processing government datasets so that they could be released  
551 without impacting the privacy of the data subjects.

## 552 **2.2 NISTIR 8053**

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

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

564 The result is this publication, which provides guidance to Government agencies seeking to use  
565 de-identification to make datasets containing personal data available to a broad audience while  
566 protecting the privacy of those upon whom the data are based.

567 De-identification is one of several models for allowing the controlled sharing of sensitive data.  
568 Other models include the use of data processing enclaves and data use agreements between data  
569 producers and data consumers. For a more complete description of data sharing models, privacy  
570 preserving data publishing, and privacy preserving data mining, please see NISTIR 8053.

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

---

<sup>40</sup> 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>

<sup>41</sup> 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>

<sup>42</sup> 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>

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

## 576 **2.3 Terminology**

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

581 *de-identification* is the “general term for any process of removing the association between a set  
582 of identifying data and the data subject.”<sup>43</sup> In this document we expand the definition of de-  
583 identification to include all techniques that provide researchers with access to microdata while  
584 simultaneously limiting the opportunity for disclosure. De-identification takes an *original*  
585 *dataset* and produces a *de-identified dataset*.

586 *re-identification* is the general term for any process that restores the association between a set of  
587 de-identified data and the data subject. Re-identification is not the only mode of failure of de-  
588 identification techniques, as information about individuals can be inferred from their data, even  
589 without restoring an association between a data subject and the de-identified data.

590 *redaction* is a kind of de-identifying technique that relies on suppression or removal of  
591 information. In general, redaction alone is not sufficient to provide formal privacy guarantees,  
592 such as differential privacy. Redaction may also reduce the usefulness of the remaining data.

593 *anonymization* is another term that is used for de-identification. The term is defined as “process  
594 that removes the association between the identifying dataset and the data subject.”<sup>44</sup> Some  
595 authors use the terms “de-identification” and “anonymization” interchangeably. Others use “de-  
596 identification” to describe a process and “anonymization” to denote a specific kind of de-  
597 identification that cannot be reversed. In health care, the term anonymization is sometimes used  
598 to describe the destruction of a table that maps pseudonyms to real identifiers.<sup>45</sup> However, the  
599 term anonymization conveys the perception that the de-identified data *cannot* be re-identified.  
600 Absent formal methods for privacy protection, it is not possible to place mathematical bounds on  
601 the amount of privacy loss that might result from the release of de-identified data. This is  
602 because techniques such as k-anonymity and traditional Statistical Disclosure Limitation based  
603 their estimates of re-identification risk on availability or lack of information that could be used to  
604 link to the de-identified dataset. Therefore, the word *anonymization* should be avoided, as it

---

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

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

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

605 makes a promise that cannot be mathematically supported.

606 Because of the inconsistencies in the use and definitions of the word “anonymization,” this  
607 document avoids the term except in this section and in the titles of some references. Instead, it  
608 uses the term “de-identification,” with the understanding that sometimes de-identified  
609 information can sometimes be re-identified, and sometimes it cannot. So, where other  
610 references<sup>46</sup> might use the term *anonymized file* to describe a dataset that has been de-identified,  
611 this publication uses the terms *de-identified file* and *de-identified dataset*, in recognition that the  
612 term *de-identified* is descriptive while the term *anonymized* is aspirational.

613 *pseudonymization* is a “particular type of anonymization that both removes the association with a  
614 data subject and adds an association between a particular set of characteristics relating to the data  
615 subject and one or more pseudonyms.”<sup>47</sup> The term *coded* is frequently used in the healthcare  
616 setting to describe data that has been pseudonymized. Pseudonymization is commonly used so  
617 that multiple observations of an individual over time can be matched, and so that an individual  
618 can be re-identified if there is a policy reason to do so. Although re-identification is typically  
619 done by consulting the key, which may be highly protected, the existence of the pseudonym  
620 identifiers frequently increases the risk of re-identification through other means.

621 Many government documents use the phrases *personally identifiable information* (PII) and  
622 *personal information*.<sup>48,49</sup> PII is typically used to indicate information that contains identifiers  
623 specific to individuals, although there are a variety of definitions for PII in various laws,  
624 regulations, and agency guidance documents. Because of these differing definitions, it is possible  
625 to have information that *singles out* individuals but which does not meet a particular definition of  
626 PII. An added complication is that some documents use the phrase PII to denote any information  
627 that is attributable to individuals, or information that is uniquely attributable to a specific  
628 individual, while others use the term strictly for data that are in fact identifying.

629 This document avoids the term “personally identifiable information.” Instead, the phrase  
630 *personal information* is used to denote information relating to individuals, and *identifying*  
631 *information* is “information that can be used to distinguish or trace an individual's identity, such  
632 as their name, social security number, biometric records, etc. alone, or when combined with  
633 other personal or identifying information which is linked or linkable to a specific individual,  
634 such as date and place of birth, mother’s maiden name, etc.”<sup>50</sup> Therefore, identifying information  
635 is personal information, but personal information is not necessarily identifying information. *Non-*  
636 *public personal information* is used to describe information that is in a dataset that is not publicly

---

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

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

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

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

<sup>50</sup> OMB M-07-16

637 available. Non-public personal information is not necessarily identifying.

638 This document envisions a *de-identification process* in which an *original dataset* containing  
639 personal information is algorithmically processed to produce a *de-identified* result. The result  
640 may be a *de-identified dataset*, *aggregate statistics* such as summary tables, or a *synthetic*  
641 *dataset*, in which the data are created by a model. This kind of de-identification is envisioned as  
642 a batch process. Alternatively, the de-identification process may be a system that accepts queries  
643 and returns responses that do not leak identifying information. De-identified results may be  
644 corrected or updated and re-released on a periodic basis. The accumulated leakage of information  
645 from multiple releases may be significant, even if the leakage from a single release is small.  
646 Issues arising from multiple releases are discussed in §3.4, “Data Release Models.”

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

656 *Disclosure limitation* is a general term for the practice of allowing summary information or  
657 queries on data within a dataset to be released without revealing information about specific  
658 individuals whose personal information is contained within the dataset. De-identification is thus  
659 a kind of disclosure limitation technique. Every disclosure limitation process introduces  
660 inaccuracy into the results.<sup>52</sup> A primary goal of disclosure limitation is to protect the privacy of  
661 individuals while avoiding the introduction of *non-ignorable biases*<sup>53</sup> (for example, bias that  
662 might lead a social scientist to come to the wrong conclusion) into the de-identified dataset. One  
663 way to measure the amount of bias that has been introduced is to compare statistics or models  
664 generated by analyzing the original dataset with those that are generated by analyzing the de-  
665 identified datasets.

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

---

<sup>51</sup> Statistical Policy Working Paper 22 (Second version, 2005), Report on Statistical Disclosure Limitation Methodology, Federal Committee on Statistical Methodology, December 2005. <https://fcsm.sites.usa.gov/reports/policy-wp/>

<sup>52</sup> For example, see See Olivia Angiuli, Joe Blitzstein, and Jim Waldo, How to De-Identify Your Data, *Communications of the ACM*, December 2015, 58:12, pp. 48-55. DOI: 10.1145/2814340

<sup>53</sup> 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/>



668 K-anonymity<sup>54</sup> is a framework for quantifying the amount of manipulation required of the quasi-  
 669 identifiers to achieve a given desired level of privacy. The technique is based on the concept of  
 670 an *equivalence class*, the set of records that have the same values on the quasi-identifiers. (A  
 671 quasi-identifier is a variable that can be used to identify an individual through association with  
 672 other information.) A dataset is said to be *k-anonymous* if, for every specific combination of  
 673 quasi-identifiers, there are no fewer than *k* matching records. For example, if a dataset that has  
 674 the quasi-identifiers (birth year) and (state) has *k=4* anonymity, then there must be at least four  
 675 records for every combination of (birth year, state). Subsequent work has refined *k-anonymity* by  
 676 adding requirements for diversity of the sensitive attributes within each equivalence class  
 677 (known as *l-diversity*<sup>55</sup>) and requiring that the resulting data are statistically close to the original  
 678 data (known as *t-closeness*).<sup>56</sup>

679 Differential privacy<sup>57</sup> is a model based on a mathematical definition of privacy that considers the  
 680 risk to an individual from the release of a query on a dataset containing their personal  
 681 information. Differential privacy is also a set of mathematical techniques that can achieve the  
 682 differential privacy definition of privacy. Differential privacy prevents both identity and attribute  
 683 disclosure by adding non-deterministic noise (usually small random values) to the results of  
 684 mathematical operations before the results are reported.<sup>58</sup> Unlike k-anonymity and other de-  
 685 identification frameworks, differential privacy is based on information theory and makes no  
 686 distinction between what is private data and what is not. Differential privacy does not require  
 687 that values be classified as direct identifiers, quasi-identifiers, and non-identifying values.  
 688 Instead, differential privacy assumes that *all values* in a record might be identifying and  
 689 therefore all must potentially be manipulated.

690 Differential privacy's mathematical definition holds that the result of an analysis of a dataset  
 691 should be roughly the same before and after the addition or removal of the data from any  
 692 individual. This works because the amount of noise added masks the contribution of any  
 693 individual. The degree of sameness is defined by the parameter  $\epsilon$  (epsilon). The smaller the  
 694 parameter  $\epsilon$ , the more noise is added, and the more difficult it is to distinguish the contribution of  
 695 a single individual. The result is increased privacy for all individuals, both those in the sample

---

<sup>54</sup> 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>

<sup>55</sup> 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.

<sup>56</sup> Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian (2007). "t-Closeness: Privacy beyond k-anonymity and l-diversity". *ICDE (Purdue University)*.

<sup>57</sup> Cynthia Dwork. 2006. Differential privacy. In *Proceedings of the 33rd international conference on Automata, Languages and Programming - Volume Part II (ICALP'06)*, Michele Bugliesi, Bart Preneel, Vladimiro Sassone, and Ingo Wegener (Eds.), Vol. Part II. Springer-Verlag, Berlin, Heidelberg, 1-12. DOI Foundations of Differential Privacy, in *Foundations and Trends in Theoretical Computer Science* Vol. 9, Nos. 3-4 (2014) 211-407, <https://www.cis.upenn.edu/~aaroth/Papers/privacybook.pdf>; [http://dx.doi.org/10.1007/11787006\\_1](http://dx.doi.org/10.1007/11787006_1)

<sup>58</sup> Cynthia Dwork, *Differential Privacy*, in *ICALP*, Springer, 2006

696 and those in the population from which the sample is drawn who are not present in the dataset.  
697 The research literature describes differential privacy being used to solve a variety of tasks  
698 including statistical analysis, machine learning, and data sanitization.<sup>59</sup> Differential privacy can  
699 be implemented in an online query system or in a batch mode in which an entire dataset is de-  
700 identified at one time. In common usage, the phrase “differential privacy” is used to describe  
701 both the formal mathematical framework for evaluating privacy loss, and for algorithms that  
702 provably provide those privacy guarantees.

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

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

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

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

---

<sup>59</sup> Cynthia Dwork and Aaron Roth, *The Algorithmic Foundations of Differential Privacy* (Foundations and Trends in Theoretical Computer Science). Now Publishers, August 11, 2014. <http://www.cis.upenn.edu/~aaroht/privacybook.html>

## 724 **3 Governance and Management of Data De-Identification**

725 The decisions and practices regarding the de-identification and release of government data can  
726 be integral to the mission and proper functioning of a government agency. As such, these  
727 activities should be managed by an agency's leadership in a way that assures performance and  
728 results that are consistent with the agency's mission and legal authority. As discussed above, the  
729 need for attention arises because of the conflicting goals of data transparency and privacy  
730 protection. Although many agencies once assumed that it is relatively straightforward to remove  
731 privacy-sensitive data from a dataset so that the remainder could be released without restriction,  
732 experience has shown that this is not the case.<sup>60</sup>

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

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

### 744 **3.1 Identifying Goals and Intended Uses of De-Identification**

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

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

751 For example:

- 752 ● **Federal Statistical Agencies** that collect, process, and publish data for use by  
753 researchers, business planners, and other well-established purposes. These agencies are  
754 likely to have in place established standards and methodologies for de-identification. As  
755 these agencies evaluate new approaches for de-identification, they should document their  
756 rationale for adopting new approaches, how successful their approaches seem to have

---

<sup>60</sup> NISTIR 8053 §2.4, §3.6

757           been over time, and inconsistencies with previous data releases.

758           ● **Federal Awarding Agencies** are allowed under OMB Circular A-110 to require that  
 759           institutions of higher education, hospitals, and other non-profit organizations receiving  
 760           federal grants provide the US Government with “the right to (1) obtain, reproduce,  
 761           publish or otherwise use the data first produced under an award; and (2) authorize others  
 762           to receive, reproduce, publish, or otherwise use such data for Federal Purposes.”<sup>61</sup>  
 763           Realizing this policy, awarding agencies can require that awardees establish data  
 764           management plans (DMPs) for making research data publicly available. Such data are  
 765           used for a variety of purposes, including transparency and reproducibility. In general,  
 766           research data that contains personal information should be de-identified by the awardee  
 767           prior to public release. Awarding agencies may establish de-identification standards to  
 768           ensure the protection of personal information.

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

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

### 777   **3.2 Evaluating Risks Arising from De-Identified Data Releases**

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

782           NIST provides detailed information on how to conduct risk assessments in NIST Special  
 783           Publication 800-30, *Guide for Conducting Risk Assessments*.<sup>62</sup>

784           Risk assessments should be based on scientific, objective factors and consider the best interests  
 785           of the individuals in the dataset, the responsibilities of the agency holding the data, and the  
 786           anticipated benefits to society. The goal of a risk evaluation is not to eliminate risk, but to  
 787           identify which risks can be reduced while still meeting the objectives of the data release, and  
 788           then deciding whether the residual risk is justified by the goals of the data release. An agency  
 789           decision making process may choose to accept or reject the risk resulting from a release of de-

---

<sup>61</sup> 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](https://www.whitehouse.gov/omb/circulars_a110)

<sup>62</sup> 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>

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

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

### 800 3.2.1 Probability of Re-Identification

801 As discussed in Section 2.3, “Terminology,” the potential impacts on individuals from the  
802 release and use of de-identified data include:<sup>63</sup>

- 803 ● **Identity disclosures** — Associating a specific individual with the corresponding  
804 record(s) in the data set with high probability. Identity disclosure can result from  
805 insufficient de-identification, re-identification by linking, or pseudonym reversal.
- 806 ● **Attribute disclosure** — determining that an attribute described in the dataset is held by a  
807 specific individual with high probability, even if the record(s) associated with that  
808 individual is (are) not identified. Attribute disclosure can occur without identity  
809 disclosure if the de-identified dataset contains data from a significant number of  
810 relatively homogeneous individuals.<sup>64</sup> In these cases, traditional de-identification does  
811 not protect against attribute disclosure, although differential privacy can.
- 812 ● **Inferential disclosure** — being able to make an inference about an individual (typically  
813 a member of a group) with high probability, even if the individual was not in the dataset  
814 prior to de-identification. “Inferential disclosure is of less concern in most cases as  
815 inferences are designed to predict aggregate behavior, not individual attributes, and thus  
816 are often poor predictors of individual data values.”<sup>65</sup> Inferential disclosure does not  
817 disclose identity, and traditional de-identification do not protect against inferential  
818 disclosure; differential privacy can only protect against it if the potential for disclosure  
819 results from the individual’s presence in the dataset. Therefore, when considering  
820 inferential disclosure, it is important to distinguish between inferences about individuals  
821 that rely on the fact that the individual’s data was used, and those that result from the  
822 individual’s membership in a group that has been subject to data collection and analysis.

---

<sup>63</sup> 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.

<sup>64</sup> NISTIR 8053 §2.4, p 13.

<sup>65</sup> Vagelis Hristidis, *Information Discovery on Electronic Health Records*, CRC Press, Dec. 2009, p. 198. 331 pages.

823 Although these disclosures are commonly thought to be discrete events involving the release of  
 824 specific data, such as an individual’s name matched to a record, disclosures can result from the  
 825 release of data that merely changes an adversary’s probabilistic belief. For example, a disclosure  
 826 might change an adversary’s estimate that a specific individual is present in a dataset from a 50%  
 827 probability to 90%. The adversary still doesn’t *know* if the individual is in the dataset or not (and  
 828 the individual might not, in fact, be in the dataset), but a probabilistic disclosure has still  
 829 occurred, because the adversary’s estimate of the individual has been changed by the data  
 830 release.

831 *Re-identification probability*<sup>66</sup> is the probability that an attacker will be able to use information  
 832 contained in a de-identified dataset to make identity-related inferences about individuals.  
 833 Different kinds of re-identification probabilities can be calculated, including:

- 834 • *Known Inclusion Re-Identification Probability (KIRP)*. The probability of finding the  
 835 record that matches a specific individual known to be in the population corresponding to  
 836 a specific record. KIRP can be expressed as the probability for a specific individual, or  
 837 the probability averaged over the entire dataset (AKIRP).<sup>67</sup>
- 838 • *Unknown Inclusion Re-Identification Probability (UIRP)*. The probability of finding the  
 839 record that matches a specific individual, without first knowing if the individual is or is  
 840 not in the dataset. UIRP can be expressed as a probability for an individual record in the  
 841 dataset averaged over the entire population (AUIRP).<sup>68</sup>
- 842 • *Recording matching probability (RMP)*. The probability of finding the record that  
 843 matches a specific individual chosen from the population. RMP can be expressed as the  
 844 probability for a specific record (RMP), the probability averaged over the entire dataset  
 845 (ARMP), or the maximum probability over the entire dataset.
- 846 • *Inclusion probability (IP)*, the probability that a specific individual’s presence in the  
 847 dataset can be inferred.

848 Whether or not it is necessary to calculate these probabilities depends upon the specifics of each  
 849 intended data release. For example, many cities publicly disclose whether or not the taxes have  
 850 been paid on a given property. Given that this information is already public, it may not be

---

<sup>66</sup> Note that previous publications described identification probability as “re-identification risk” and used scenarios such as a journalist seeking to discredit a national statistics agency and a prosecutor seeking to find information about a suspect as the basis for probability calculations. That terminology is not presented in this document because of possible unwanted connotations of those terms, and in the interest of bringing the terminology of de-identification into agreement with the terminology used in contemporary risk analyses processes. See Elliot M, Dale A. *Scenarios of attack: the data intruder’s perspective on statistical disclosure risk*, Netherlands Official Statistics 1999;14(Spring):6-10.

<sup>67</sup> Some texts refer to KIRP as “prosecutor risk.” The scenario is that a prosecutor is looking for records belonging to a specific, named individual.

<sup>68</sup> 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.

851 necessary to consider inclusion probably when a dataset of property taxpayers for a specific  
852 dataset is released. Likewise, there may be some attributes in a dataset that are already public  
853 and thus may not need to be protected with disclosure limitation techniques. However, the  
854 existence of such attributes may themselves pose a re-identification risk for other information in  
855 this dataset, or in other de-identified datasets. Also, the mere fact that information is public may  
856 not negate the responsibility of an agency to provide protection for that information, as the  
857 aggregation and distribution of information may cause privacy risk that was not otherwise  
858 present.

859 It may be difficult to calculate specific re-identification probabilities, as the ability to re-identify  
860 depends on the original dataset, the de-identification technique, the technical skill of the attacker,  
861 the attacker's available resources, and the availability of additional data (publicly available or  
862 privately held) that can be linked with the de-identified data. It is likely that the probability of  
863 re-identification increases over time as techniques improve and more contextual information  
864 becomes available to attackers.

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

- 871 • Although a specific de-identified dataset may not be seen as sensitive, re-identifying that  
872 dataset may be an important step in re-identifying another dataset that is sensitive.  
873 Alternatively, the adversary may merely wish to embarrass the government agency. Thus,  
874 adversaries may have a strong incentive to re-identify datasets that are seemingly  
875 innocuous.
- 876 • Although the public may not be skilled in re-identification in general, many resources on  
877 the Internet make it easy to acquire specialized datasets, tools, and experts for specific re-  
878 identification challenges. Also, family members, friends, colleagues, and others may  
879 possess substantial personal knowledge about individuals in the data that can be used for  
880 re-identification.

881 Instead, de-identification practitioners should assume that de-identified government datasets will  
882 be subjected to sustained, world-wide re-identification attempts, and they should gauge their de-  
883 identification requirements accordingly. However, it is unrealistic to assume that all of the  
884 world's resources will be used to attempt to re-identify every publicly released file. Therefore, it  
885 is also necessary to gauge de-identification requirements using a risk assessment. More  
886 information on conducting risk assessments can be found in NIST Special Publication 800-30,

887 *Guide for Conducting Risk Assessments*<sup>69</sup>.

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

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

### 899 **3.2.2 Adverse Impacts Resulting from Re-Identification**

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

903 Potential adverse impacts on individuals include:

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

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

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



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

920 ● Financial impact resulting from the harm to the individuals (e.g. lawsuits).

921 ● Civil or criminal sanctions against employees or contractors resulting from a data release  
922 contrary to US law.

923 Potential adverse impacts on society as a whole include:

924 ● It may undermine the reputation of researchers in general and the willingness of the  
925 public to support/tolerate research and to provide accurate information to government  
926 agencies and to researchers.

927 ● It may engender a lack of trust in government. Individuals may stop consenting to the use  
928 of their data, or even stop providing their data or provide false data.

929 ● Damage to the practice of using de-identification information. De-identification is an  
930 important tool for promoting research and accountability. Poorly executed de-  
931 identification efforts may negatively impact the public's view of this technique and limit  
932 its use as a result.

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

937 As part of a risk analysis process, organizations should enumerate specific measures that they  
938 will take to minimize the risk of successful re-identification. Organizations may wish to consider  
939 both the actual risk and the perceived risk on the part of those in the dataset as well as the  
940 broader community.

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

### 946 **3.2.3 Impacts other than re-identification**

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

950 ● The risk of inferential disclosures.

951 ● The risk that the de-identification process might introduce bias or inaccuracies into the

952 dataset that result in incorrect decisions.<sup>70</sup>

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

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

### 960 3.2.4 Remediation

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

## 965 3.3 Data Life Cycle

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

---

<sup>70</sup> 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, 23<sup>rd</sup> *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. Barth-Jones, DC. *The Debate Over 'Re-Identification' Of Health Information: What Do We Risk?* Health Affairs Blog, August 10, 2012. <http://healthaffairs.org/blog/2012/08/10/the-debate-over-re-identification-of-health-information-what-do-we-risk/>

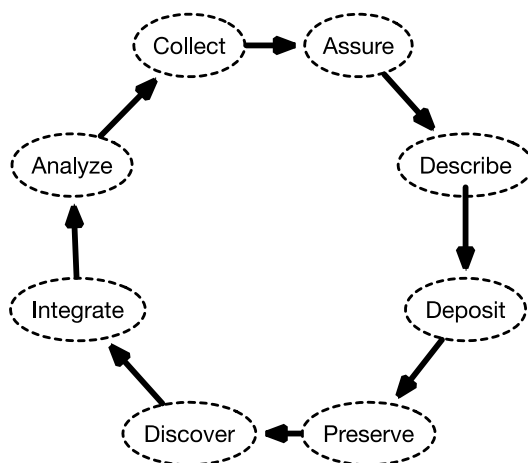
<sup>71</sup> NIST Special Publication 1500-1, *NIST Big Data Interoperability Framework: Volume 1, Definitions*. NIST Big Data Public Working Group, Definitions and Taxonomies Subgroup. September 2015. <http://dx.doi.org/10.6028/NIST.SP.1500-1>

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

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

975 Currently there is no standardized model for the data life cycle.

976 Michener et al. describe the data life cycle as a true cycle of Collect → Assure → Describe →  
 977 Deposit → Preserve → Discover → Integrate → Analyze → Collect:<sup>74</sup>



978

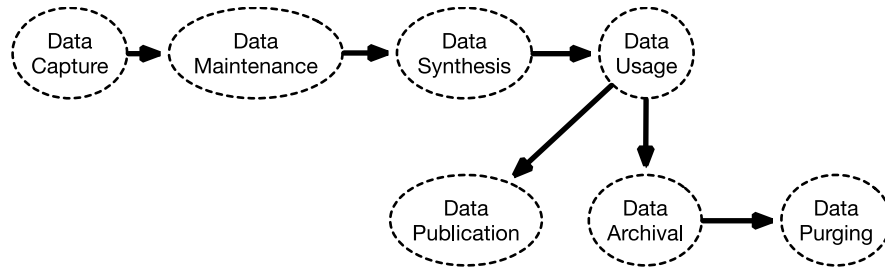
979 **Figure 1 Michener et al.'s view of the data lifecycle is a true cycle, with analysis guiding future collection.**

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

986 Chisholm and others describe the data life cycle as a linear process that involves Data Capture →  
 987 Data Maintenance → Data Synthesis → Data Usage → {Data Publication & Data Archival} →  
 988 Data Purging:<sup>75</sup>

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

<sup>75</sup> 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>

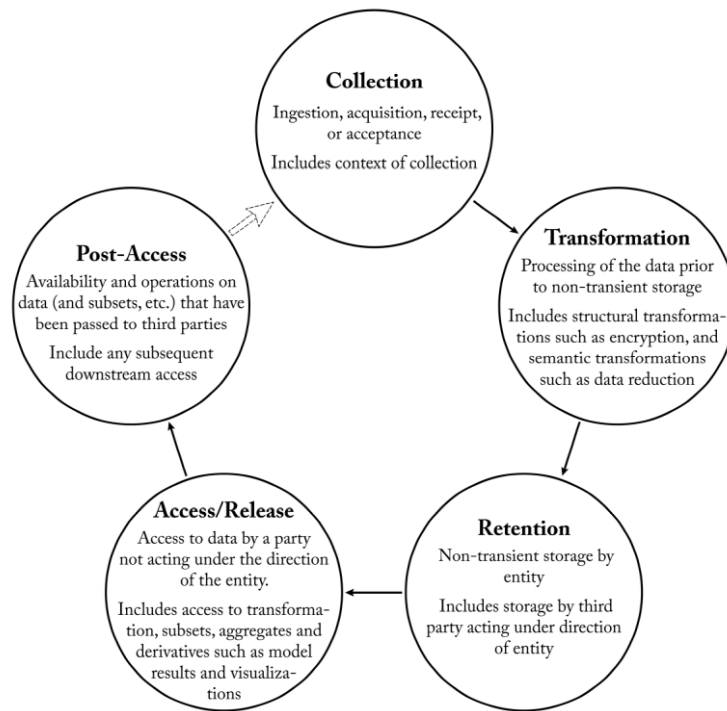


989

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

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

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

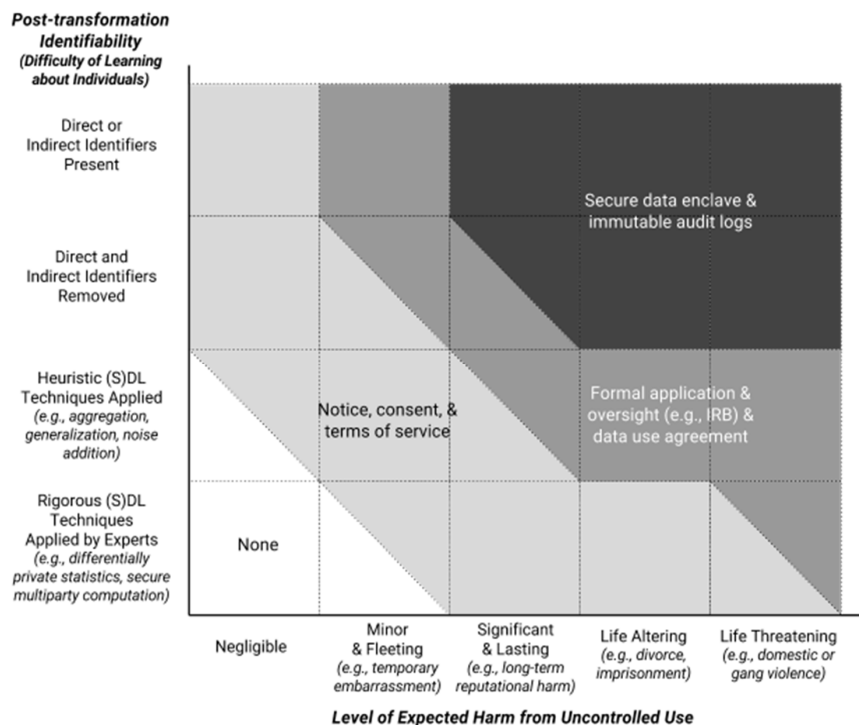


1000

1001

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

1002



1003

1004 **Figure 4 Conceptual diagram of the relationship between post-transformation identifiability, level of expected**  
 1005 **harm, and suitability of selected privacy controls for a data release. From Altman et al.**

1006 Agencies performing de-identification should document that:

- 1007 • Techniques used to perform the de-identification are theoretically sound and generally  
 1008 accepted.
- 1009 • Software used to perform the de-identification is reliable for the intended task.
- 1010 • Individuals who performed the de-identification were suitably qualified.
- 1011 • Tests were used to evaluate the effectiveness of the de-identification.
- 1012 • Ongoing monitoring is in place to assure the continued effectiveness of the de-  
 1013 identification strategy.

1014 No matter where de-identification is applied in the data life cycle, agencies should document the  
 1015 answers of these questions for each de-identified dataset:

- 1016 • Are direct identifiers collected with the dataset?
- 1017 • Even if direct identifiers are not collected, is it nevertheless still possible to identify the  
 1018 data subjects through the presence of quasi-identifiers?

- 1019       • Where in the data life cycle is de-identification performed? Is it performed in only one  
1020       place, or is it performed in multiple places?
- 1021       • Is the original dataset retained after de-identification?
- 1022       • Is there a key or map retained, so that specific data elements can be re-identified later?
- 1023       • How are decisions made regarding de-identification and re-identification?
- 1024       • Are there specific datasets that can be used to re-identify the de-identified data? If so,  
1025       what controls are in place to prevent intentional or unintentional re-identification?
- 1026       • Is it a problem if a dataset is re-identified?
- 1027       • Is there a mechanism that will inform the de-identifying agency if there is an attempt to  
1028       re-identify the de-identified dataset? Is there a mechanism that will inform the agency if  
1029       the attempt is successful?

### 1030    **3.4 Data Sharing Models**

1031    Agencies should decide the data release model that will be used to make the data available  
1032    outside the agency after the data have been de-identified.<sup>76</sup> Possible models include:

- 1033       • **The Release and Forget Model:**<sup>77</sup> The de-identified data may be released to the public,  
1034       typically by being published on the Internet. It can be difficult or impossible for an  
1035       organization to recall the data once released in this fashion and may limit information for  
1036       future releases.
- 1037       • **The Data Use Agreement (DUA) Model:** The de-identified data may be made available  
1038       under a legally binding data use agreement that details what can and cannot be done with  
1039       the data. Typically, data use agreements may prohibit attempted re-identification, linking  
1040       to other data, and redistribution of the data without a similarly binding DUA. A DUA  
1041       will typically be negotiated between the data holder and qualified researchers (the  
1042       “qualified investigator model”<sup>78</sup>) or members of the general public (e.g. citizen scientists  
1043       or the media), although they may be simply posted on the Internet with a click-through

---

<sup>76</sup> NISTIR 8053 §2.5, p. 14

<sup>77</sup> Ohm, Paul, Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization. *UCLA Law Review*, Vol. 57, p. 1701, 2010

<sup>78</sup> 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

1044 license agreement that must be agreed to before the data can be downloaded (the “click-  
1045 through model”<sup>79</sup>).

1046 ● **The Synthetic Data with Verification Model:** Statistical Disclosure Limitation  
1047 techniques are applied to the original dataset and used to create a synthetic dataset that  
1048 contains many of the aspects of the original dataset, but which does not contain  
1049 disclosing information. The synthetic dataset is released, either publically or to vetted  
1050 researchers. The synthetic dataset can then be used as a proxy for the original dataset, and  
1051 if constructed well, the results of statistical analyses should be similar. If used in  
1052 conjunction with an enclave model as below, researchers may use the synthetic dataset to  
1053 develop queries and/or analytic software; these queries and/or software can then be taken  
1054 to the enclave or provided to the agency and be applied on the original data.

1055 ● **The Enclave Model:**<sup>80,81,82</sup> The de-identified data may be kept in a segregated enclave  
1056 that restricts the export of the original data, and instead accepts queries from qualified  
1057 researchers, runs the queries on the de-identified data, and responds with results.  
1058 Enclaves can be physical or virtual, and can themselves operate under a variety of  
1059 different models. For example, vetted researchers may travel to the enclave to perform  
1060 their research, as is done with the Federal Statistical Research Data Centers operated by  
1061 US Census Bureau. Enclaves may be used to implement the verification step of the  
1062 Simulated Data with Verification Model. Queries made in the enclave model may be  
1063 vetted either automatically or manually (e.g., by the DRB). Vetting can try to screen for  
1064 queries that might violate privacy or are inconsistent with the stated purpose of the  
1065 research.

1066 Sharing models should consider the possibility of multiple or periodic releases. Just as repeated  
1067 queries to the same dataset may leak personal data from the dataset, repeated de-identified  
1068 releases (whether from the same dataset or from different datasets containing some of the same  
1069 individuals) by an agency may result in compromising the privacy of individuals unless each  
1070 subsequent release is viewed in light of the previous release. Even if a contemplated release of a  
1071 de-identified dataset does not directly reveal identifying information, Federal agencies should  
1072 ensure that the release, combined with previous releases, will also not reveal identifying  
1073 information.<sup>83</sup>

---

<sup>79</sup> Ibid.

<sup>80</sup> Ibid.

<sup>81</sup> 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

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

<sup>83</sup> See Joel Havermann, plaintiff - Appellant, v. Carolyn W. Colvin, Acting Commissioner of the Social Security Administration,

1074 Instead of sharing an entire dataset, the data owner may choose to release a sample. If only a  
1075 subsample is released, the probability of re-identification decreases, because an attacker will not  
1076 know if a specific individual from the data universe is present in the de-identified dataset.<sup>84</sup>  
1077 However, releasing only a subset may decrease the statistical power of tests on the data, may  
1078 cause users to draw incorrect inferences on the data if proper statistical sampling methods are not  
1079 used, may obscure the ability to draw correct inferences, and may not align with agency goals  
1080 regarding transparency and accountability.

### 1081 3.5 The Five Safes

1082 The Five Safes is a popular framework created for “designing, describing and evaluating” data  
1083 access systems, and especially access systems designed for the sharing of information from a  
1084 national statistics with a research community.<sup>85</sup> The framework proposes five “risk (or access)  
1085 dimensions:”

- 1086 ● **Safe projects** — Is this use of the data appropriate?
- 1087 ● **Safe people** — Can the researchers be trusted to use it in an appropriate manner?
- 1088 ● **Safe data** — Is there a disclosure risk in the data itself?
- 1089 ● **Safe settings** — Does the access facility limit unauthorized use?
- 1090 ● **Safe outputs** — Are the statistical results non-disclosive?

1091 Each of these dimensions is intended to be *independent*. That is, the legal, moral and ethical  
1092 review of the research proposed by the “safe projects” dimension should be evaluated  
1093 independently of the people proposing to conduct the research, and the location where the  
1094 research will be conducted.

1095 One of the positive aspects of the Five Safes framework is that it forces data controllers to  
1096 consider many different aspects of data release when considering or evaluating data access  
1097 proposals. Frequently, the authors write, it is common for data owners to “focus on one, and only  
1098 one, particular issue (such as the legal framework surrounding access to their data, or IT  
1099 solutions).” With a framework such as the Five Safes, people who may be specialists in one area  
1100 are forced to consider (or to explicitly not consider) a variety of different aspects of privacy

---

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

<sup>84</sup> 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>

<sup>85</sup> 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>



1101 protection.

1102 The Five Safes framework can be used as a tool for designing access systems, for evaluating  
1103 existing systems, for communication and for training. Agencies should consider using a  
1104 framework such as The Five Safes for organizing risk analysis of data release efforts.

### 1105 **3.6 Disclosure Review Boards**<sup>86</sup>

1106 Disclosure Review Boards (DRBs), also known as Data Release Boards, are administrative  
1107 bodies created within an organization that are charged with assuring that a data release meets the  
1108 policy and procedural requirements of that organization. DRBs should be governed by a written  
1109 *mission statement* and *charter* that are, ideally, approved by the same mechanisms that the  
1110 organization uses to approve other organization-wide policies.

1111 The DRB should have a mission statement that guides its activities. For example, the US  
1112 Department of Education’s DRB has the mission statement:

1113 “The Mission of the Department of Education Disclosure Review Board (ED-DRB) is to  
1114 review proposed data releases by the Department’s principal offices (POs) through a  
1115 collaborate technical assistance, aiding the Department to release as much useful data as  
1116 possible, while protecting the privacy of individuals and the confidentiality of their data, as  
1117 required by law.”<sup>87</sup>

1118 The DRB charter specifies the mechanics of how the mission is implemented. A formal, written  
1119 charter promotes transparency in the decision-making process, and assures consistency in the  
1120 applications of its policies. It is envisioned that most DRBs will be established to weigh the  
1121 interests of data release against those of individual privacy protection. However, a DRB may also  
1122 be chartered to consider *group harms*<sup>88</sup> that can result from the release of a dataset beyond harm  
1123 to individual privacy. Such considerations should be framed within existing organizational  
1124 policy, regulation, and law. Some agencies may balance these concerns by employing data use  
1125 models other than de-identification—for example, by establishing data enclaves where a limited  
1126 number of vetted researchers can gain access to sensitive datasets in a way that provides data  
1127 value while attempting to minimize the possibility for harm. In those agencies, a DRB would be  
1128 empowered to approve the use of such mechanisms.

1129 The DRB charter should specify the DRB’s composition. To be effective, the DRB should  
1130 include representatives from multiple groups, and should include experts in both technology and  
1131 policy of privacy. Specifically, DRBs may wish to have as members:

---

<sup>86</sup> 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.

<sup>87</sup> 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>

<sup>88</sup> NISTIR 8053 §2.4, p. 13

1132 • Individuals representing the interests of potential users; such individuals need not come  
1133 from outside of the organization.

1134 • Representation from among the public, and specifically from groups represented in the  
1135 data sets if they have a limited scope.

1136 • Representation from the organization's leadership team. Such representation helps  
1137 establish the DRB's credibility with the rest of the organization.

1138 • A representative of the organization's senior privacy official.

1139 • Subject matter experts.

1140 • Outside experts.

1141 The charter should establish rules for ensuring quorum, and specify if members can designate  
1142 alternates on a standing or meeting-by-meeting basis. The DRB should specify the mechanism  
1143 by which members are nominated and approved, their tenure, conditions for removal, and  
1144 removal procedures.<sup>89</sup>

1145 The charter should set policy expectations for recording keeping and reporting, including  
1146 whether records and reports are considered public or restricted. The charter should indicate if it is  
1147 possible to exclude sensitive decisions from these requirements and the mechanism for doing so.

1148 To meet its requirement of evaluating data releases, the DRB should require that written  
1149 applications be submitted to the DRB that specify the nature of the dataset, the de-identification  
1150 methodology, and the result. An application may require that the proposer present the re-  
1151 identification risk, the risk to individuals if the dataset is re-identified, and a proposed plan for  
1152 detecting and mitigating successful re-identification. In addition, the DRB should require that,  
1153 when individuals are informed that their information will be de-identified, that they also be  
1154 informed that privacy risks may remain despite de-identification.

1155 DRBs may wish to institute a two-step process, in which the applicant first proposes and receives  
1156 approval for a specific de-identification process that will be applied to a specific dataset, then  
1157 submits and receives approval for the release of the dataset that has been de-identified according  
1158 to the proposal. However, because it is theoretically impossible to predict the results of applying  
1159 an arbitrary process to an arbitrary dataset,<sup>90,91</sup> the DRB should be empowered to reject release

---

<sup>89</sup> 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

<sup>90</sup> Church, A. 1936. 'A Note on the Entscheidungsproblem'. Journal of Symbolic Logic, 1, 40-41.

<sup>91</sup> 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

1160 of a dataset even if it has been de-identified in accordance with an approved procedure, because  
 1161 performing the de-identification may demonstrate that the procedure was insufficient to protect  
 1162 privacy. The DRB may delegate the responsibility of reviewing the de-identified dataset, but it  
 1163 should not be delegated to the individual that performed the de-identification.

1164 The DRB charter should specify if the Board needs to approve each data release by the  
 1165 organization or if it may grant blanket approval for all data of a specific type that is de-identified  
 1166 according to a specific methodology. The charter should specify duration of the approval. Given  
 1167 advances in the science and technology of de-identification, it is inadvisable that a Board be  
 1168 empowered to grant release authority for an indefinite amount of time.

1169 In most cases a single privacy protection methodology will be insufficient to protect the varied  
 1170 datasets that an agency may wish to release. That is, different techniques might best optimize the  
 1171 tradeoff between re-identification risk and data usability, depending on the specifics of each kind  
 1172 of dataset. Nevertheless, the DRB may wish to develop guidance, recommendations and training  
 1173 materials regarding specific de-identification techniques that are to be used. Agencies that  
 1174 standardize on a small number of de-identification techniques will gain familiarity with these  
 1175 techniques and are likely to have results that have a higher level of consistency and success than  
 1176 those that have no such guidance or standardization.

1177 Although it is envisioned that DRBs will work in a cooperative, collaborative and congenial  
 1178 manner with those inside an agency seeking to release de-identified data, there will at times be a  
 1179 disagreement of opinion. For this reason, the DRB's charter should state if the DRB has the final  
 1180 say over disclosure matters or if the DRB's decisions can be overruled, by whom, and by what  
 1181 procedure. For example, an agency might give the DRB final say over disclosure matters, but  
 1182 allow the agency's leadership to replace members of the DRB as necessary. Alternatively, the  
 1183 DRB's rulings might merely be advisory, with all data releases being individually approved by  
 1184 agency leadership or its delegates.<sup>92</sup>

1185 Finally, agencies should decide whether or not the DRB charter will include any kind of  
 1186 performance timetables or be bound by a service level agreement (SLA) that defines a level of  
 1187 service to which the DRB commits.

1188 Key elements of a DRB:

- 1189 • Written mission statement and charter.
- 1190 • Members represent different groups within the organization, including leadership.
- 1191 • Board receives written applications to release de-identified data.

---

<sup>92</sup> 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,

- 1192 • Board reviews *both* proposed methodology *and* the results of applying the methodology.
  - 1193 • Applications should identify risk associated with data release, including re-identification  
1194 probability, potentially adverse events that would result if individuals are re-identified,  
1195 and a mitigation strategy if re-identification takes place.
  - 1196 • Approvals may be valid for multiple releases, but should not be valid indefinitely.
  - 1197 • Mechanisms for dispute resolution.
  - 1198 • Timetable or service level agreement (SLA).
  - 1199 • Legal and technical understanding of privacy.
- 1200 Example outputs of a DRB include specifying access methods for different kinds of data  
1201 releases, establishing acceptable levels of re-identification risk (e.g. values of  $k$  or  $\epsilon$ ), and  
1202 maintaining detailed records of previous data releases—ideally including the dataset that  
1203 was released and the privacy-preserving methodology that was employed.
- 1204 There is some similarity between DRBs as envisioned here and the Institutional Review Board  
1205 (IRBs) system created by the Common Rule<sup>93</sup> for regulating human subjects research in the  
1206 United States. However, there are important differences:
- 1207 • While the purpose of IRBs is to protect human subjects, DRBs are charged with  
1208 protecting data subjects, institutions, and potentially society as a whole.
  - 1209 • Whereas IRBs are required to have “at least one member whose primary concerns are in  
1210 nonscientific areas” and “at least one member who is not otherwise affiliated with the  
1211 institution and who is not part of the immediate family of a person who is affiliated with  
1212 the institution,” there is no need for such members in a DRB.
  - 1213 • Whereas IRBs give approval for research and then typically receive reports only during  
1214 an annual review or when a research project terminates, DRBs may be involved at  
1215 multiple points during the process.
  - 1216 • Whereas approval of an IRB is required before research with human subjects can  
1217 commence, DRBs are typically involved after research has taken place, prior to data or  
1218 other research findings being released.
  - 1219 • Whereas service on an IRB requires knowledge of the Common Rule and an

---

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

1220 understanding of ethics, service on a DRB requires knowledge of statistics, computation  
1221 and public policy.

## 1222 **3.7 De-Identification Standards**

1223 Agencies can rely on de-identification standards to provide a standardized terminology,  
1224 procedures, and performance criteria for de-identification efforts. Agencies can adopt existing  
1225 de-identification standards or create their own. De-identification standards can be prescriptive or  
1226 performance-based.

### 1227 **3.7.1 Benefits of Standards**

1228 De-identification standards assist agencies in the process of de-identifying data prior to public  
1229 release. Without standards, data owners may be unwilling to share data, as they may be unable to  
1230 assess if a procedure for de-identifying data is sufficient to minimize privacy risk.

1231 Standards can increase the availability of individuals with appropriate training by providing a  
1232 specific body of knowledge and practice that training should address. Absent standards, agencies  
1233 may forego opportunities to share data. De-identification standards can help practitioners to  
1234 develop a community, certification and accreditation processes.

1235 Standards decrease uncertainty and provide data owners and custodians with best practices to  
1236 follow. Courts can consider standards as acceptable practices that should generally be followed.  
1237 In the event of litigation, an agency can point to the standard and say that it followed good data  
1238 practice.

### 1239 **3.7.2 Prescriptive De-Identification Standards**

1240 A prescriptive de-identification standard specifies an algorithmic procedure that, if followed,  
1241 results in data that are de-identified.

1242 The “Safe Harbor” method of the HIPAA Privacy Rule<sup>94</sup> is an example of a prescriptive de-  
1243 identification standard. The intent of the Safe Harbor method is to “provide covered entities with  
1244 a simple method to determine if the information is adequately de-identified.”<sup>95</sup> It does this by  
1245 specifying 18 kinds of identifiers that, once removed, results in the de-identification of Protected  
1246 Health Information (PHI) and the subsequent relaxing of privacy regulations. Although the  
1247 Privacy Rule does state that a covered entity employing the Safe Harbor method must have no  
1248 “actual knowledge” that the PHI, once de-identified, could still be used to re-identify individuals,  
1249 covered entities are not obligated to employ experts or mount re-identification attacks against  
1250 datasets to verify that the use of the Safe Harbor method has in fact resulted in data that cannot

---

<sup>94</sup> Health Insurance Portability and Accountability Act of 1996 (HIPAA) Privacy Rule Safe Harbor method §164.514(b)(2).

<sup>95</sup> *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](http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/guidance.html#_edn32)

1251 be re-identified.

1252 Prescriptive standards have the advantages of being relatively easy for users to follow, but  
1253 developing, testing, and validating such standards can be burdensome. Because prescriptive de-  
1254 identification standards are not changed on a case-by-case basis, there is a tendency for them to  
1255 be more conservative than is necessary, resulting in the unnecessary decrease in data quality for  
1256 corresponding levels of risk.

1257 Agencies creating prescriptive de-identification standards should assure that data de-identified  
1258 according to the rules have a sufficiently small risk of being re-identified that is consistent with  
1259 the intended data use; such assurances frequently cannot be made unless formal privacy  
1260 techniques such as *differential privacy* are employed. However, agencies may determine that  
1261 public policy goals furthered by having an easy-to-use prescriptive standard outweighs the risk  
1262 of a standard that does not have provable privacy guarantees.

1263 Prescriptive de-identification standards carry the risk that the procedure specified in the standard  
1264 may not sufficiently de-identify to avoid the risk of re-identification, especially as methodology  
1265 advances and more data sources become available.

### 1266 **3.7.3 Performance-Based De-Identification Standards**

1267 A performance based de-identification standard specifies properties that the de-identification  
1268 procedure must have.

1269 The “Expert Determination” method of the HIPAA Privacy Rule is an example of a performance  
1270 based de-identification standard. Under the rule, a technique for de-identifying data is sufficient  
1271 if an appropriate expert “determines that the risk is very small that the information could be used,  
1272 alone or in combination with other reasonably available information, by an anticipated recipient  
1273 to identify an individual who is a subject of the information.”<sup>96</sup> The rule does not require that  
1274 experts describe the methodology used, nor does it put the expert’s work under the jurisdiction of  
1275 HHS.

1276 Performance-based standards have the advantage of allowing users many different ways to solve  
1277 a problem. As such, they leave room for innovation. Such standards also have the advantage that  
1278 they can embody the desired outcome.

1279 Performance-based standards should be sufficiently detailed that they can be performed in a  
1280 manner that is reliable and repeatable. For example, standards that call for the use of experts  
1281 should specify how an expert’s expertise is to be determined. Standards that call for the reduction  
1282 of risk to an acceptable level should provide a procedure for determining that level.

---

<sup>96</sup> The Health Insurance Portability and Accountability Act of 1996 (HIPAA) Privacy Rule Expert Determination Method §164.514(b)(1).

### 1283 **3.8 Education, Training and Research**

1284 De-identifying data in a manner that preserves privacy can be a complex mathematical,  
1285 statistical, administrative, and data-driven process. Frequently the opportunities for identity  
1286 disclosure will vary from dataset to dataset. Privacy protecting mechanisms developed for one  
1287 dataset may not be appropriate for others. For these reasons, agencies engaging in de-  
1288 identification should ensure that their workers have adequate education and training in the  
1289 subject domain. Agencies may wish to establish education or certification requirements for those  
1290 who work directly with the datasets or to adopt industry standards such as the HITrust De-  
1291 Identification Framework.<sup>97</sup> Because de-identification techniques are modality dependent,  
1292 agencies using de-identification may need to institute research efforts to develop and test  
1293 appropriate data release methodologies.

1294

---

<sup>97</sup> Health Information Trust Alliance, De-Identification Framework, 2016 <https://hitrustalliance.net/de-identification/>.

## 1295 **4 Technical Steps for Data De-Identification**

1296 The goal of de-identification is to transform data in a way that protects privacy while preserving  
 1297 the validity of inferences drawn on that data within the context of a target use-case. This section  
 1298 discusses technical options for performing de-identification and verifying the result of a de-  
 1299 identification procedure.

1300 Agencies should adopt a detailed, written process for de-identifying data prior to commencing  
 1301 work on a de-identification project. The details of the process will depend on the particular de-  
 1302 identification approach that is pursued. In developing technical steps for data de-identification,  
 1303 agencies may wish to consider existing de-identification standards, such as the HIPAA Privacy  
 1304 Rule or the IHE De-Identification Handbook<sup>98</sup> or the HITRUST De-Identification Framework.<sup>99</sup>

### 1305 **4.1 Determine the Privacy, Data Usability, and Access Objectives**

1306 Agencies intent on de-identifying data for release should understand the nature of the data that  
 1307 they intended to de-identification and determine the policies and standards that will be used to  
 1308 determine acceptable levels of data quality, de-identification, and risk of re-identification. For  
 1309 example:

- 1310 ● Where did the data come from?
- 1311 ● What promises were made when the data were collected?
- 1312 ● What are the legal and regulatory requirements regarding privacy and release of the data?
- 1313 ● What is the purpose of the data release?
- 1314 ● What is the intended use of the data?
- 1315 ● What data sharing model (§3.4) will be used?
- 1316 ● Which standards for privacy protection or de-identification will be used?
- 1317 ● What is the level of risk that the project is willing to accept?
- 1318 ● How should compliance with that level of risk be determined?
- 1319 ● What are the goals for limiting re-identification? That only a few people be re-identified?
- 1320 That only a few people can be re-identified in theory, but no one will actually be re-

<sup>98</sup> IHE IT Infrastructure Handbook, De-Identification, Integrating the Healthcare Enterprise, June 6, 2014.  
[http://www.ihe.net/User\\_Handbooks/](http://www.ihe.net/User_Handbooks/)

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



1321 identified in practice? That there will be a small percentage chance that everybody will be  
1322 re-identified?

- 1323 ● What harm might result from re-identification, and what techniques that will be used to  
1324 mitigate those harms?

1325 Some goals and objectives are synergistic, while others are in opposition.

## 1326 4.2 Conducting a Data Survey

1327 Different kinds of data require different kinds of de-identification techniques. As a result, an  
1328 important early step in the de-identification of government data is to identify the data modalities  
1329 that are present in the dataset and formulate a plan for de-identification that takes into account  
1330 goals for data release, data quality, privacy protection, and the best available science.

1331 For example:

- 1332 ● **Tabular numeric and categorical data** is the subject of the majority of de-identification  
1333 research and practice. These datasets are most frequently de-identified by using  
1334 techniques based on the designation and removal of direct identifiers and the  
1335 manipulation of quasi-identifiers. The chief criticism of de-identification based on direct  
1336 and quasi-identifiers is that administrative determinations of quasi-identifiers may miss  
1337 variables that can be uniquely identifying when combined and linked with external  
1338 data—including data that are not available at the time the de-identification is performed,  
1339 but become available in the future. De-identification can be evaluated using frameworks  
1340 such as Statistical Disclosure Limitation (SDL) or k-anonymity. However, *risk*  
1341 *determinations based on this kind of de-identification will be incorrect if direct and*  
1342 *quasi-identifiers are not properly classified.* For example, if there exist quasi-identifiers  
1343 that are not identified as such and not subjected to SDL, then it may be easy to re-identify  
1344 records in the de-identified dataset.

1345  
1346 Tabular data may also be used to create a synthetic dataset that preserves some inference  
1347 validity but does not have a 1-to-1 correspondence to the original dataset.

- 1348 ● **Dates and times** require special attention when de-identifying, because all dates within a  
1349 dataset are inherently linked to the natural progression of time. Some dates and times are  
1350 highly identifying, while others are not. Dates which refer to matters of public record  
1351 (e.g., date of birth, death or home purchase) should be routinely taken as having high re-  
1352 identification potential. Dates may also form the basis of linkages between dataset  
1353 records or even within a record—for example, a record may contain the date of  
1354 admission, the date of discharge, and the number of days in residence. Thus, care should  
1355 be taken when de-identifying dates to locate and properly handle potential linkages and  
1356 relationships: applying different techniques to different fields may result in information  
1357 being left in a dataset that can be used for re-identification. Specific issues regarding date  
1358 de-identification are discussed below in §4.2.2.

- 1359       • **Geographic and map data** also require special attention when de-identifying, as some  
1360       locations can be highly identifying, other locations are not identifying at all, and some  
1361       locations are only identifying at specific times. As with dates and times, the challenge of  
1362       de-identifying geographic locations comes from the fact that locations inherently link to  
1363       an external reality. Identifying locations can be de-identified through the use of  
1364       perturbation or generalization. The effectiveness such de-identification techniques for  
1365       protecting privacy in the presence of external information has not been well  
1366       characterized.<sup>100,101</sup> Specific issues regarding geographical de-identification are discussed  
1367       below in §4.2.3.
- 1368       • **Unstructured text** may contain direct identifiers, such as a person’s name, or may  
1369       contain additional information that can serve as a quasi-identifier. Finding such  
1370       identifiers and distinguishing them from non-identifiers invariably requires domain-  
1371       specific knowledge.<sup>102</sup> Note that unstructured text may be present in tabular datasets and  
1372       require special attention.<sup>103</sup>
- 1373       • **Photos and video** may contain identifying information such as printed names (e.g. name  
1374       tags), as well as metadata in the file format. There also exists a range of biometric  
1375       techniques for matching photos of individuals against a dataset of photos and  
1376       identifiers.<sup>104</sup>
- 1377       • **Medical imagery** poses additional problems over photographs and video due to the  
1378       presence of many kinds of identifiers. For example, identifying information may be  
1379       present in the image itself (e.g. a photo may show an identifying scar or tattoo), an  
1380       identifier may be “burned in” to the image area, or an identifier may be present in the file  
1381       metadata. The body part in the image itself may also be recognized using a biometric  
1382       algorithm and dataset.<sup>105</sup>
- 1383       • **Genetic sequences** and other kinds of sequence information can be identified by  
1384       matching to existing databanks that match sequences and identities. There is also  
1385       evidence that genetic sequences from individuals who are not in datasets can be matched

---

<sup>100</sup> NISTIR 8053, §4.5 p. 37

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

<sup>102</sup> NISTIR 8053, §4.1 p. 30

<sup>103</sup> For an example of how unstructured text fields can damage the policy objectives and privacy assurances of a larger structured dataset, see Andrew Peterson, *Why the names of six people who complained of sexual assault were published online by Dallas police*, The Washington Post, April 29, 2016. <https://www.washingtonpost.com/news/the-switch/wp/2016/04/29/why-the-names-of-six-people-who-complained-of-sexual-assault-were-published-online-by-dallas-police/>

<sup>104</sup> NISTIR 8053, §4.2 p. 32

<sup>105</sup> NISTIR 8053, §4.3 p. 35

1386 through genealogical triangulation, a process that uses genetic information and other  
 1387 information as quasi-identifiers to single-out a specific identity.<sup>106</sup> At the present time  
 1388 there is no known method to reliably de-identify genetic sequences. Specific issues  
 1389 regarding the de-identification of genetic information is discussed below in §4.2.4.

1390 In many cases data may be complex and contain multiple modalities. Such mixtures may  
 1391 complicate risk determinations.

1392 A dataset that is thought to contain purely tabular data may be found, upon closer examination,  
 1393 to include unstructured text or even photograph data.

### 1394 **4.3 De-identification by removing identifiers and transforming quasi-** 1395 **identifiers**

1396 De-identification based on the removal of identifiers and transformation of quasi-identifiers is  
 1397 one of the most common approaches for de-identification currently in use. This approach has the  
 1398 advantage of being conceptually straightforward and there being a long institutional history in  
 1399 using this approach within both federal statistical agencies and the healthcare industry. This  
 1400 approach has the disadvantage of being not based on formal methods for assuring privacy  
 1401 protection. The lack of formal methods does not mean that this approach cannot protect privacy,  
 1402 but it does mean that privacy protection is not assured.

1403 Below is a sample process for de-identifying data by removing identifiers and transforming  
 1404 quasi-identifiers:<sup>107</sup>

1405 Step 1. Determine the re-identification risk threshold. The organization determines  
 1406 acceptable risk for working with the dataset and possibly mitigating controls, based on  
 1407 strong precedents and standards (e.g., Working Paper 22: Report on Statistical Disclosure  
 1408 Control).

1409 Step 2. Determine the information in the dataset that could be used to identify the data  
 1410 subjects. Identifying information can include:

- 1411 a. **Direct identifiers**, such as names, phone numbers, and other information that  
 1412 unambiguously identifies an individual.
- 1413 b. **Quasi-identifiers** that could be used in a linkage attack. Typically, quasi-  
 1414 identifiers identify multiple individuals and can be used to triangulate on a  
 1415 specific individual.

---

<sup>106</sup> NISTIR 8053, §4.4 p. 36

<sup>107</sup> This process is based on a process developed by Professors Khaled El Emam and Bradley Malin. See K. El Emam and B. Malin, "Appendix B: Concepts and Methods for De-Identifying Clinical Trial Data," in *Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk*, Institute of Medicine of the National Academies, The National Academies Press, Washington, DC. 2015

- 1416           c. **High-dimensionality data**<sup>108</sup> that can be used to single out data records and thus  
 1417           constitute a unique pattern that could be identifying, if these values exist in a  
 1418           secondary source to link against.<sup>109</sup>
- 1419       Step 3.     Determine the direct identifiers in the dataset. An expert determines the elements  
 1420           in the dataset that serve only to identify the data subjects.
- 1421       Step 4.     Mask (transform) direct identifiers. The direct identifiers are either removed or  
 1422           replaced with pseudonyms. Options for performing this operation are discussed below in  
 1423           §4.3.1.
- 1424       Step 5.     Perform threat modeling. The organization determines the additional information  
 1425           they might be able to use for re-identification, including both quasi-identifiers and non-  
 1426           identifying values that an adversary might use for re-identification.
- 1427       Step 6.     Determine the minimal acceptable data quality. In this step, the organization  
 1428           determines what uses can or will be made with the de-identified data.
- 1429       Step 7.     Determine the transformation process that will be used to manipulate the quasi-  
 1430           identifiers. Pay special attention to the data fields containing dates and geographical  
 1431           information, removing or recoding as necessary.
- 1432       Step 8.     Import (sample) data from the source dataset. Because the effort to acquire data  
 1433           from the source (identified) dataset may be substantial, some researchers recommend a  
 1434           test data import run to assist in planning.<sup>110</sup>
- 1435       Step 9.     Review the results of the trial de-identification. Correct any coding or algorithmic  
 1436           errors that are detected.
- 1437       Step 10.    Transform the quasi-identifiers for the entire dataset.
- 1438       Step 11.    Evaluate the actual re-identification risk. The actual identification risk is  
 1439           calculated. As part of this evaluation, every aspect of the released dataset should be  
 1440           considered in light of the question, “can *this* information be used to identify someone?”
- 1441       Step 12.    Compare the actual re-identification risk with the threshold specified by the  
 1442           policy makers.

---

<sup>108</sup> 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.

<sup>109</sup> 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

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

1443 Step 13. If the data do not pass the actual risk threshold, adjust the procedure and repeat  
 1444 Steps 11 and 12. For example, additional transformations may be required. Alternatively,  
 1445 it may be necessary to remove outliers.

#### 1446 **4.3.1 Removing or Transformation of Direct Identifiers**

1447 There are many possible processes for removing direct identifiers from a dataset, including:

- 1448 ● Removal and replacement with the value used by the database to indicate a missing  
 1449 value, such as Null or NA.
- 1450 ● Masking with a repeating character, such as XXXXXX or 999999.
- 1451 ● Encryption. After encryption, the cryptographic key should be discarded to prevent  
 1452 decryption or the possibility of a brute force attack. However, the key must not be  
 1453 discarded if there is a desire to employ the same transformation at a later point in time,  
 1454 but rather stored in a secure location separate from the de-identified dataset. Encryption  
 1455 used for this purpose carries special risks which need to be addressed with specific  
 1456 controls; see below for further information. Encryption is a pseudonymization technique.
- 1457 ● Hashing with a keyed hash, such as a Hash-based Message Authentication Code  
 1458 (HMAC)<sup>111</sup>. The hash key should have sufficient randomness to defeat a brute force  
 1459 attack aimed at recovering the hash key. For example, SHA-256 HMAC with a 256-bit  
 1460 randomly generated key. As with encryption, the key should be discarded unless there is  
 1461 a desire for repeatability. Hashing used for this purpose carries special risks which need  
 1462 to be addressed with specific controls; see below for further information.
- 1463 ● Replacement with keywords, such as transforming “George Washington” to “PATIENT.”
- 1464 ● Replacement by realistic surrogate values, such as transforming “George Washington” to  
 1465 “Abraham Polk.”<sup>112</sup> If the replacement by realistic surrogate values is consistent and  
 1466 surrogates are not reused, then replacement is a pseudonymization technique.

1467 The technique used to remove direct identifiers should be clearly documented for users of the  
 1468 dataset, especially if the technique of replacement by realistic surrogate names is used.

1469 If the agency plans to make data available for longitudinal research and contemplates multiple  
 1470 data releases, then the transformation process should be repeatable, and the resulting transformed

---

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

<sup>112</sup> 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.

1471 identities are *pseudonyms*. The mapping between the direct identifier and the pseudonym is  
 1472 performed using a lookup table or a repeatable transformation. In either case, the release of the  
 1473 lookup table or the information used for the repeatable transformation will result in the  
 1474 compromise identities. Thus, the lookup table or the information for the transformation must be  
 1475 highly protected. When using a lookup table, the pseudonym must be randomly assigned. A  
 1476 significant risk of using a repeatable transformation is that an attacker will be able to determine  
 1477 the transformation, and in so doing gain the capability to re-identify all of the records in the  
 1478 dataset.

1479 **SPECIAL SECURITY NOTE REGARDING**  
 1480 **THE ENCRYPTION OR HASHING OF DIRECT IDENTIFIERS**

1481 The transformation of direct identifies through encryption or hashing carries special risks, as  
 1482 errors in procedure or the release of the encryption key can result in the compromise of identity.  
 1483 When information is protected with encryption, the security of the encrypted data depends  
 1484 entirely on the security of the encryption key. If a key is improperly chosen, it may be possible  
 1485 for an attacker to find it using a brute force search. Because there is no visual difference between  
 1486 data that are encrypted with a strong encryption key and data that are encrypted with a weak key,  
 1487 it is necessary for an organization to relies on encryption to assure through administrative  
 1488 controls that keys are used that are both unpredictable and suitably protected. The use of  
 1489 encryption or hashing to protect direct identifiers is not recommended unless justified by specific  
 1490 extenuating circumstances.

1491 **4.3.2 De-Identifying Numeric Quasi-Identifiers**

1492 Once a determination is made regarding quasi-identifiers, they should be transformed. A variety  
 1493 of techniques are available to transform quasi-identifiers:

- 1494 ● **Top and bottom coding.** Outlier values that are above or below certain values are coded  
 1495 appropriately. For example, the HIPAA Privacy Rules calls for ages over 89 to be  
 1496 “aggregated into a single category of age 90 or older.”<sup>113</sup>
- 1497 ● **Micro aggregation**, in which individual microdata are combined into small groups that  
 1498 preserve some data analysis capability while providing for some disclosure protection.<sup>114</sup>
- 1499 ● **Generalize categories with small values.** When preparing contingency tables, several  
 1500 categories with small values may be combined. For example, rather than reporting that  
 1501 there is 1 person with blue eyes, 2 people with green eyes, and 1 person with hazel eyes,  
 1502 it may be reported that there are 4 people with blue, green or hazel eyes.

---

<sup>113</sup> HIPAA § 164.514 (b).

<sup>114</sup> J. M. Mateo-Sanz, J. Domingo-Ferrer, a comparative study of microaggregation methods, *Qüestió*, vol. 22, 3, p. 511-526, 1998.

- 1503       • **Data suppression.** Cells in contingency tables with counts lower than a predefined  
1504       threshold can be suppressed to prevent the identification of attribute combinations with  
1505       small numbers.<sup>115</sup>
- 1506       • **Blanking and imputing.** Specific values that are highly identifying can be removed and  
1507       replaced with imputed values.
- 1508       • **Attribute or record swapping,** in which attributes or records are swapped between  
1509       records representing individuals. For example, data representing families in two similar  
1510       towns within a county might be swapped with each other. “Swapping has the additional  
1511       quality of removing any 100-percent assurance that a given record belongs to a given  
1512       household,”<sup>116</sup> while preserving the accuracy of regional statistics such as sums and  
1513       averages. For example, in this case the average number of children per family in the  
1514       county would be unaffected by data swapping.
- 1515       • **Noise infusion.** Also called “partially synthetic data,” small random values may be added  
1516       to attributes. For example, instead of reporting that a person is 84 years old, the person  
1517       may be reported as being 79 years old. Noise infusion increases variance and leads to  
1518       attenuation bias in estimated regression coefficients and correlations among attributes.<sup>117</sup>
- 1519       These techniques (and others) are described in detail in several publications, including:
- 1520       • *Statistical Policy Working Paper #2* (Second version, 2005) by the Federal Committee on  
1521       Statistical Methodology.<sup>118</sup> This 137-page paper also includes worked examples of  
1522       disclosure limitation, specific recommended practices for Federal agencies, profiles of  
1523       federal statistical agencies conducting disclosure limitation, and an extensive  
1524       bibliography.
- 1525       • *The Anonymisation Decision-Making Framework*, by Mark Elliot, Elaine MacKey,  
1526       Kieron O’Hara and Caroline Tudor, UKAN, University of Manchester, Manchester, UK.  
1527       2016. This 156-page book provides tutorials and worked examples for de-identifying data  
1528       and calculating risk.

---

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

<sup>116</sup> *Census Confidentiality and Privacy: 1790-2002*, US Census Bureau, 2003, p. 31

<sup>117</sup> George T. Duncan, Mark Elliot, Juan-José Salazar-Gonzalez, *Statistical Confidentiality: Principles and Practice*, Springer, 2011, p. 113, cited in John M. Abowd and Ian M. Schmutte, *Economic Analysis and Statistical Disclosure Limitation*, Brookings Papers on Economic Activity, March 19, 2015. <https://www.brookings.edu/bpea-articles/economic-analysis-and-statistical-disclosure-limitation/>

<sup>118</sup> *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.

- 1529       • *IHE IT Infrastructure Handbook, De-Identification, Integrating the Healthcare*  
1530       Enterprise, June 6, 2014. [http://www.ihe.net/User\\_Handbooks/](http://www.ihe.net/User_Handbooks/)
- 1531       Swapping and noise infusion both introduce noise into the dataset, such that records literally  
1532       contain incorrect data. These techniques can introduce sufficient noise to provide formal privacy  
1533       guarantees.
- 1534       All of these techniques impact data quality, but whether they impact data *utility* depends upon  
1535       the downstream uses of the data. For example, top-coding household incomes will not impact a  
1536       measurement of the 90-10 quantile ratio, but it will impact a measurement of the top 1% of  
1537       household incomes.<sup>119</sup>
- 1538       As currently practiced, statistical agencies typically do not document in detail the specific  
1539       statistical disclosure technique that they use to transform quasi-identifiers when performing  
1540       statistical disclosure limitation. Likewise, statistical agencies do not document the parameters  
1541       used in the transformations, nor the amount of data that have been transformed, as documenting  
1542       these techniques can allow an adversary to reverse-engineer the specific values, eliminating the  
1543       privacy protection.<sup>120</sup> This lack of transparency can result in erroneous conclusions on the part  
1544       of data users.
- 1545       **4.3.3 De-identifying dates**
- 1546       Dates can exist many ways in a dataset. Dates may be in particular kinds of typed columns, such  
1547       as a date of birth or the date of an encounter. Dates may be present as a number, such as the  
1548       number of days since an epoch such as January 1, 1900. Dates may be present in the free text  
1549       narratives. Dates may be present in photographs—for example, a photograph that shows a  
1550       calendar or a picture of a computer screen that shows date information.
- 1551       Several strategies have been developed for de-identifying dates:
- 1552       • Under the HIPAA Privacy Rule, dates must be generalized to no greater specificity than  
1553       the year (e.g. July 4, 1776 becomes 1776).
- 1554       • Dates within a single person’s record can be systematically adjusted by a random amount.  
1555       For example, dates of a hospital admission and discharge might be systematically moved  
1556       the same number of days — a date of admission and discharge of July 4, 1776 and July 9,  
1557       1776 become Sept. 10, 1777 and Sept. 15, 1777<sup>121</sup>). However, this does not eliminate the

---

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

<sup>120</sup> 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/>

<sup>121</sup> Office of Civil Rights, “Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule”, US Department of Health and Human Services, 2010. <http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/guidance.html>



- 1558 risk an attacker will make inferences based on the interval between dates.
- 1559 ● In addition to a systematic shift, the intervals between dates can be perturbed to protect  
1560 against re-identification attacks involving identifiable intervals while still maintaining the  
1561 ordering of events.
- 1562 ● Some dates cannot be arbitrarily changed without compromising data quality. For  
1563 example, it may be necessary to preserve day-of-week, whether a day is a work day or a  
1564 holiday, or a relationship to a holiday or event.
- 1565 ● Likewise, some ages can be randomly adjusted without impacting data quality, while  
1566 others cannot. For example, in many cases the age of an individual can be randomly  
1567 adjusted  $\pm 2$  years if the person is over the age of 25, but not if their age is between 1 and  
1568 3.

#### 1569 **4.3.4 De-identifying geographical locations**

1570 Geographical data can exist in many ways in a dataset. Geographical locations may be indicated  
1571 by map coordinates (e.g. 39.1351966, -77.2164013), street address (e.g. 100 Bureau Drive), or  
1572 postal code (20899). Geographical locations can also be embedded in textual narratives.

1573 Some geographical locations are not identifying (e.g. a crowded train station), while others may  
1574 be highly identifying (e.g. a house in which a single person lives). Other positions may be  
1575 identifying at some times but not at others. Single locations may be not identifying, but may  
1576 become identifying if they represent locations linked to a single individual that are recorded over  
1577 time.

1578 The amount of noise required to de-identify geographical locations significantly depends on  
1579 external factors. Identity may be shielded in an urban environment by adding  $\pm 100\text{m}$ , whereas a  
1580 rural environment may require  $\pm 5\text{Km}$  to introduce sufficient ambiguity.

1581 A prescriptive rule, even one that accounts for varying population densities, may still not be  
1582 applicable, if it fails to consider the other quasi-identifiers in the data set. Noise should also be  
1583 added with caution to avoid the creation of inconsistencies in underlying data—for example,  
1584 moving the location of a residence along a coast into a body of water or across geo-political  
1585 boundaries.

1586 De-identification of geographical data is especially challenging when location of individuals is  
1587 recorded over time, because behavioral time-location patterns can act as fingerprints for re-  
1588 identification purposes even with a small number of recorded locations per individual.<sup>122</sup>

---

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

#### 1589 4.3.5 De-identifying genomic information

1590 Deoxyribonucleic acid (DNA) is the molecule inside human cells that carries genetic instructions  
1591 used for the proper functioning of living organisms. DNA present in the cell nucleus is inherited  
1592 from both parents; DNA present in the mitochondria is only inherited from an organism's  
1593 mother.

1594 DNA is a repeating polymer that is made from four chemical bases: adenine (A), guanine (G),  
1595 cytosine (C) and thymine (T). Human DNA consists of roughly 3 billion bases, of which 99% is  
1596 the same in all people.<sup>123</sup> Modern technology allows the complete specific sequence of an  
1597 individual's DNA to be chemically determined, although this is rarely done in practice. With  
1598 current technology, it is far more common to use a DNA microarray to probe for the presence or  
1599 absence of specific DNA sequences at predetermined points in the genome. This approach is  
1600 typically used to determine the presence or absence of specific single nucleotide polymorphisms  
1601 (SNPs).<sup>124</sup> DNA sequences and SNPs are the same for identical twins, individuals resulting from  
1602 divided embryos, and clones. With these exceptions, it is believed that no two humans have the  
1603 same complete DNA sequence.

1604 Individual SNPs may be shared by many individuals, but it a sufficiently large number of SNPs  
1605 that show sufficient variability is generally believed to produce a combination that is unique to  
1606 an individual. Thus, there are some sections of the DNA sequence and some combinations of  
1607 SNPs that have high variability within the human population and others that have significant  
1608 conservation between individuals within a specific population or group.

1609 When there is high variability, DNA sequences and SNPs can be used to match an individual  
1610 with a historical sample that has been analyzed and entered into a dataset. The inheritability of  
1611 genetic information has also allowed researchers to determine the surnames and even the  
1612 complete identities of some individuals.<sup>125</sup>

1613 As the number of individuals that have their DNA and SNPs measured increases, scientists are  
1614 realizing that the characteristics of DNA and SNPs in individuals may be more complicated than  
1615 the preceding paragraphs imply. DNA changes as individuals age because of senescence,  
1616 transcription errors, and mutation. DNA methylation, which can impact the functioning of DNA,  
1617 also changes over time.<sup>126</sup> Some individuals are made up with DNA from multiple individuals,  
1618 typically the result of fusion of twins in early pregnancy; these people are known as *chimera* or  
1619 *mosaic*. In 2015 a man in the US failed a paternity test because the genes in his saliva were

---

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

<sup>124</sup> 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

<sup>125</sup> Gymrek *et al.*, Identifying Personal Genomes by Surname Inference, *Science* 18 Jan 2013, 339:6117.

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

1620 different from those in his sperm.<sup>127</sup> A human chimera was identified in 1953 as a result of  
 1621 having a blood that a mixture of two blood types, A and O.<sup>128</sup> The incidence of human chimeras  
 1622 is unknown.

1623 Because of the high variability inherent in DNA, complete DNA sequences may be identifiable.  
 1624 Likewise, biological samples for which DNA can be extracted may be identifiable. Subsections  
 1625 of an individual's DNA sequence and collections of highly variable SNPs may be identifiable  
 1626 unless there it is known that there are many individuals that share the region of DNA or those  
 1627 SNPs. Furthermore, genetic information may not only identify an individual, but may also be  
 1628 able to identify an individual's ancestors, siblings, and descendants.

#### 1629 **4.3.6 Challenges Posed by Aggregation Techniques**

1630 Aggregation does not necessarily provide privacy protection, especially when data is presented  
 1631 as part of multiple data releases. Consider the hypothetical example of a school uses aggregation  
 1632 to report the number of students performing below, at, and above grade level:

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

1633

1634 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

1635

1636 By comparing the two tables, one can readily infer that the student who joined the school is  
 1637 performing above grade level. Because aggregation does not inherently protect privacy,

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

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

1638 aggregation alone is not sufficient to provide formal privacy guarantees. However, the  
1639 differential privacy literature provides many methods for performing aggregation that are both  
1640 formally private and highly accurate on large datasets. These methods work through the  
1641 additional of carefully calibrated “random noise.”  
1642

#### 1643 **4.3.7 Challenges posed by High-Dimensionality Data**

1644 Even after removing all of the unique identifiers and manipulating the quasi-identifiers, some  
1645 data can still be identifying if it is of sufficient high-dimensionality, and if there exists a way to  
1646 link the supposedly non-identifying values with an identity.<sup>129</sup>

#### 1647 **4.3.8 Challenges Posed by Linked Data**

1648 Data can be linked in many ways. Pseudonyms allow data records from the same individual to be  
1649 linked together over time. Family identifiers allow data from parents to be linked with their  
1650 children. Device identifiers allow data to be linked to physical devices, and potentially link  
1651 together all data coming from the same device. Data can also be linked to geographical locations.

1652 Data linkage increases the risk of re-identification by providing more attributes that can be used  
1653 to distinguish the true identity of a data record from others in the population. For example,  
1654 survey responses that are linked together by household are more readily re-identified than survey  
1655 responses that are not linked. For example, heart rate measurements may not be considered  
1656 identifying, but given a long sequence of tests, each individual in a dataset would have a unique  
1657 constellation of heart rate measurements, and thus the data set could be susceptible to being  
1658 linked with another data set that contains these same values. (Note that this is different than  
1659 characterizing an individual’s heartbeat pattern so that it could be used as a biometric. In this  
1660 case, it is a specific sequence of heartbeats that is recognized.) Geographical location data can,  
1661 when linked over time create individual behavioral time-location patterns can act as fingerprints  
1662 for re-identification purposes even with a small number of recorded locations per individual.<sup>130</sup>

1663 Dependencies between records may result in record linkages even when there is no explicit  
1664 linkage identifier. For example, it may be that an organization has new employees take a  
1665 proficiency test within 7 days of being hired. This information would allow links to be drawn  
1666 between an employee dataset that accurately reported an employee’s start date and a training  
1667 dataset that accurately reported the date that the test was administered, even if the sponsoring  
1668 organization did not intend for the two datasets to be linkable.

---

<sup>129</sup> 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

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

### 1669 **4.3.9 Challenges Posed by Composition**

1670 In computer science, the term *composition* refers to combining multiple functions together to  
 1671 create more complicated ones. One of the defining characteristics of complex systems is that  
 1672 complicated functions created by composition can have unpredictable results, even when they  
 1673 are composed from very simple components. The challenge of composition is to develop  
 1674 approaches for understanding when composition will have unpredictable results and to address  
 1675 those results proactively.

1676 When de-identifying, it is important to understand if the techniques that are used will retain their  
 1677 privacy guarantees when they are subject to composition. For example, if the same dataset is  
 1678 made available through two different de-identification regimes, attention must be paid to whether  
 1679 the privacy guarantees will remain if the two downstream datasets are re-combined.

1680 Composition concerns can arise when the same dataset is provided to multiple downstream users,  
 1681 when the dataset is published on a periodic basis, or when changes in computer technology result  
 1682 in new aspects of a dataset being made available. Privacy risk can result from unanticipated  
 1683 composition, which is one of the reasons that released datasets should be subjected to periodic  
 1684 review and reconsideration.

### 1685 **4.3.10 Post-Release Monitoring**

1686 Following the release of a de-identified dataset, the releasing agency should monitor to assure  
 1687 that the assumptions made during the de-identification remain valid. This is because the  
 1688 identifiability of a dataset may increase over time.

1689 For example, the de-identified dataset may contain information that can be linked to an internal  
 1690 dataset that is later the subject of a data breach. In such a situation, the data breach could also  
 1691 result in the re-identification of the de-identified dataset.

## 1692 **4.4 Synthetic Data**

1693 An alternative to de-identifying using the technique presented in the previous section is to use  
 1694 the original dataset to create a synthetic dataset.

1695 Synthetic data can be created by two approaches:<sup>131</sup>

- 1696 ● Sampling an existing dataset and either adding noise to specific cells likely to have a high  
 1697 risk of disclosure, or replacing these cells with imputed values. (A “partially synthetic  
 1698 dataset.”)
- 1699 ● Using the existing dataset to create a model and then using that model to create a

---

<sup>131</sup> 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>

1700 synthetic dataset. (A “fully synthetic dataset.”)

1701 In both cases, formal privacy techniques can be used to quantify the privacy protection offered  
 1702 by the synthetic dataset.

1703 It is also possible to create *test data* that is syntactically valid but which does not convey  
 1704 accurate information when analyzed. Such data can be used for software development. When  
 1705 creating test data, it is useful for the names, addresses and other information in the data to be  
 1706 conspicuously non-natural, so that the test data is not inadvertently confused with actual data.

1707 Other terms have been used to describe synthetic data; Table 1 presents terms that have been  
 1708 collected from the academic literature.

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

Table 1 Terms for Synthetic Data

1709

1710 **4.4.1 Partially Synthetic Data**

1711 A partially synthetic dataset is one in which some of the data is inconsistent with the original  
 1712 dataset. For example, data belonging to two families in adjoining towns may be swapped to

1713 protect the identity of the families. Alternatively, the data for an outlier variable may be removed  
 1714 and replaced with a range value that is incorrect (for example, replacing the value “60” with the  
 1715 range “30-35”). It is considered best practice that the data publisher indicate that some values  
 1716 have been modified or otherwise imputed, but not to reveal the specific values that have been  
 1717 modified.

#### 1718 **4.4.2 Fully Synthetic Data**

1719 A fully synthetic dataset is a dataset for which there is no one-to-one mapping between data in  
 1720 the original dataset and in the de-identified dataset. One approach to create a fully synthetic  
 1721 dataset is to use the original dataset to create a high-fidelity model, and then to use the model to  
 1722 produce individual data elements consistent with the model using a simulation. Special efforts  
 1723 must be taken to maintain marginal and join probabilities when creating fully synthetic data.

1724 Fully synthetic datasets cannot provide more information to the downstream user than was  
 1725 contained in the original model. Nevertheless, some users may prefer to work with the fully  
 1726 synthetic dataset instead of the model:

- 1727 ● Synthetic data provides users with the ability to develop queries and other techniques that  
 1728 can be applied to the real data, without exposing real data to users during the  
 1729 development process. The queries and techniques can then be provided to the data owner,  
 1730 which can run the queries or techniques on the real data and provide the results to the  
 1731 users.
- 1732 ● Analysts may discover things from the synthetic data that they don't see in the model,  
 1733 even though the model contains the information. However, such discoveries should be  
 1734 evaluated against the real data to assure that the things that were discovered were actually  
 1735 in the original data, and not an artifact of the synthetic data generation.
- 1736 ● Some users may place more trust in a synthetic dataset than in a model.
- 1737 ● When researchers form their hypotheses working with synthetic data and then verify their  
 1738 findings on actual data, they are protected from pretest estimation and false-discovery  
 1739 bias.<sup>132</sup>

1740 Both high-fidelity models and synthetic data generated from models may leak personal  
 1741 information that is potentially re-identifiable; the amount of leakage can be controlled using  
 1742 formal privacy models (such as differential privacy) that typically involve the introduction of  
 1743 noise.

1744 There are several advantages to agencies that chose to release de-identified data as a fully

---

<sup>132</sup> 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/>

1745 synthetic dataset:

- 1746 ● It can be very difficult or even impossible to map records to actual people.
- 1747 ● The privacy guarantees can potentially be mathematically established and proven (cf. the  
1748 section below on “Creating a synthetic dataset with differential privacy”).
- 1749 ● The privacy guarantees can remain in force even if there are future data releases.

1750 Fully synthetic data also has these disadvantages and limitations:

- 1751 ● It is not possible to create pseudonyms that map back to actual people, because the  
1752 records are fully synthetic.
- 1753 ● The data release may be less useful for accountability or transparency. For example,  
1754 investigators equipped with a synthetic data release would be unable to find the actual  
1755 “people” who make up the release, because they would not actually exist.
- 1756 ● It is impossible to find meaningful correlations or abnormalities in the synthetic data that  
1757 are not represented in the model. For example, if a model is built by considering all  
1758 possible functions of 1 and 2 variables, then any correlations found of 3 variables will be  
1759 a spurious artifact of the way that the synthetic data were created, and not based on the  
1760 underlying real data.
- 1761 ● Users of the data may not realize that the data are synthetic. Simply providing  
1762 documentation that the data are fully synthetic may not be sufficient public notification,  
1763 since the dataset may be separated from the documentation. Instead, it is best to indicate  
1764 in the data itself that the values are synthetic. For example, names like “SYNTHETIC  
1765 PERSON” may be placed in the data. Such names could follow the distribution of real  
1766 names but obviously be not real.

#### 1767 **4.4.3 Synthetic Data with Validation**

1768 Agencies that share or publish synthetic data can optionally make available a validation service  
1769 that takes queries or algorithms developed with synthetic data and applies them to actual data.  
1770 The results of these queries or algorithms can then then be compared with the results of running  
1771 the same queries on the synthetic data and the researchers warned if the results are different.  
1772 Alternatively, the results can be provided to the researchers after the application of statistical  
1773 disclosure limitation.

#### 1774 **4.4.4 Synthetic Data and Open Data Policy**

1775 Releases of synthetic data can be confusing to the lay public. Specifically, synthetic data may  
1776 contain synthetic individuals who appear quite similar to actual individuals in the population.  
1777 Furthermore, fully synthetic datasets do not have a zero disclosure risk, because they still convey  
1778 some non-public personal information about individuals. The disclosure risk may be greater



1779 when synthetic data are created with traditional data imputing techniques, rather than those based  
 1780 on formal privacy models such as differential privacy, as the formal models have provisions for  
 1781 tracking the accumulated privacy loss budget resulting from multiple data operations.

1782 One of the advantages of synthetic data is that the privacy loss budget can be spent in creating  
 1783 the synthetic dataset, rather than in responding to interactive queries. The danger in using the  
 1784 privacy loss budget to respond to interactive queries is that each query decreases the budget. As  
 1785 the number of queries continues, the data controller needs to respond by increasing the amount of  
 1786 noise, by accepting a higher level of privacy risk, or by ceasing to answer questions. This can  
 1787 result in equity issues, if the first users to query the dataset are able to obtain better answers than  
 1788 later users.<sup>133</sup>

#### 1789 **4.4.5 Creating a synthetic dataset with differential privacy**

1790 A growing number of mathematical algorithms have been developed for creating synthetic  
 1791 datasets that meet the mathematical definition of privacy provided by differential privacy.<sup>134</sup>  
 1792 Most of these algorithms will transform a dataset containing private data into a new dataset that  
 1793 contains synthetic data that nevertheless provides reasonably accurate results in response to a  
 1794 variety of queries. However there is no algorithm or implementation currently in existence that  
 1795 can be used by a person who is unskilled in the area of differential privacy.

1796 The classic definition of differential privacy is that if results of function calculated on a dataset  
 1797 are indistinguishable within a certain privacy metric  $\epsilon$  (epsilon) no matter whether any  
 1798 possible individual is included in the dataset or removed from the dataset,<sup>135</sup> then that  
 1799 function is said to provide  $\epsilon$ -differential privacy.

1800 In the mathematical formulation of differential privacy, the two datasets (with and without the  
 1801 individual) are denoted by  $D_1$  and  $D_2$ , and the function that is said to be differential private is  $\kappa$ .  
 1802 The formal definition of differential privacy is then:

1803 **Definition 2.**<sup>136</sup> A randomized function  $\kappa$  gives  $\epsilon$ -differential privacy if for all datasets  $D_1$

---

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

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

<sup>135</sup> 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.

<sup>136</sup> 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](http://dx.doi.org/10.1007/11787006_1). Definition 1 is not important for this publication.

1804 and  $D_2$  differing on at most one element, and all  $S \subseteq \text{Range}(\kappa)$ ,

1805 
$$Pr[\kappa(D_1) \in S] \leq e^\epsilon \times Pr[\kappa(D_2) \in S]$$

1806 This definition that may be easier to understand if rephrased as a dataset  $D$  with an arbitrary  
 1807 person  $p$ , and dataset  $D - p$ , the dataset without a person, and the multiplication operator  
 1808 replaced by a division operator, e.g.:

1809 
$$\frac{Pr[\kappa(D - p) \in S]}{Pr[\kappa(D) \in S]} \leq e^\epsilon$$

1810 That is, the ratio between the probable outcomes of function  $\kappa$  operating on the datasets with and  
 1811 without person  $p$  should be less than  $e^\epsilon$ . If the two probabilities are equal, then  $e^\epsilon = 1$ , and  
 1812  $\epsilon = 0$ . If the difference between the two probabilities is potentially infinite—that is, there is  
 1813 no privacy—then  $e^\epsilon = \infty$  and  $\epsilon = \infty$ .

1814 For values of epsilon that are small,  $e^\epsilon$  is approximately equal to  $1 + \epsilon$ . Intuitively, this means  
 1815 that small values of  $\epsilon$  result in high privacy outcomes, while large values of  $\epsilon$  result in low  
 1816 privacy outcomes.

1817 What this means in practice for the creation of a synthetic dataset with differential privacy and a  
 1818 sufficiently large  $\epsilon$  is that functions computed on the so-called “privatized” dataset will have a  
 1819 similar probability distribution no matter whether any person in the original data that was used to  
 1820 create the model is included or excluded. In practice, this similarity is provided by adding noise  
 1821 to the model. For datasets drawn from a population with a large number of individuals, the model  
 1822 (and the resulting synthetic data) will have a small amount of noise added. For models and  
 1823 results created from a small population (or for contingency tables with small cell counts), this  
 1824 will require the introduction of a significant amount of noise. The amount of noise added is  
 1825 determined by the differential privacy parameter  $\epsilon$ , the number of individuals in the dataset, and  
 1826 the specific differential privacy mechanism that is employed.

1827 Smaller values of  $\epsilon$  provide for more privacy but decreased data quality. As stated above, the  
 1828 value of 0 implies that the function  $\kappa$  provides the same distribution of answers no matter if  
 1829 anyone is removed or a person’s attributes changed, while the value of  $\infty$  allows the original  
 1830 dataset to be released without being subject to disclosure limitation.

1831 Many academic papers on differential privacy have assumed a value for  $\epsilon$  of 1.0 or  $e$  but have not  
 1832 explained the rationale of the choice. Some researchers working in the field of differential  
 1833 privacy have just started the process of mapping existing privacy regulations to the choice of  $\epsilon$ .  
 1834 For example, using a hypothetical example of a school that wished to release a dataset containing  
 1835 the school year and absence days for a number of students, the value of  $\epsilon$  using one set of  
 1836 assumptions might be calculated to 0.3379 (producing a low degree of data quality), but this  
 1837 number can safely be raised to 2.776 (and correspondingly higher data quality) without

1838 significantly undermining the privacy protections.<sup>137</sup>

1839 Another challenge in implementing differential privacy is the demands that the algorithms make  
 1840 on the correctness of implementation. For example, a Microsoft researcher discovered that four  
 1841 publicly available general purpose implementations of differential privacy contained a flaw that  
 1842 potentially leaked non-public personal information because of the binary representation of IEEE  
 1843 floating point numbers used by the implementations.<sup>138</sup>

1844 Since there are relatively few scholarly publications regarding the deployment of differential  
 1845 privacy in real-world situation, combined with the lack of guidance and experience in choosing  
 1846 appropriate values of  $\epsilon$ , agencies that are interested in using differential privacy algorithms to  
 1847 allow querying of sensitive datasets or for the creation of synthetic data should take great care to  
 1848 assure that the techniques are appropriately implemented and that the privacy protections are  
 1849 appropriate to the desired application.

## 1850 **4.5 De-Identifying with an interactive query interface**

1851 Another model for granting the public access to de-identified agency information is to construct  
 1852 an interactive query interface that allows members of the public or qualified investigators to run  
 1853 queries over the agency's dataset. This option has been developed by several agencies and there  
 1854 are many different ways that it can be implemented.

- 1855       • If the queries are run on actual data, the results can be altered through the injection of  
 1856       noise to protect privacy, potentially satisfying a formal privacy model such as differential  
 1857       privacy. Alternatively, the individual queries can be reviewed by agency staff to verify  
 1858       that privacy thresholds are maintained.
- 1859       • Alternatively, the queries can be run on synthetic data. In this case, the agency can also  
 1860       run queries on the actual data and warn the external researchers if the queries run on  
 1861       synthetic data deviate significantly from the queries run on the actual data (taking care to  
 1862       ensure that the warning itself does not compromise privacy of some individual).
- 1863       • Query interfaces can be made freely available on the public internet, or they can be made  
 1864       available in a restricted manner to qualified researchers operating in secure locations.

1865 Care must be taken in implementing interactive query interfaces, as it is possible to reconstruct

---

<sup>137</sup> Jaewoo Lee and Chris Clifton. 2011. How much is enough? choosing  $\epsilon$  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.

<sup>138</sup> 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>

1866 private microdata from a query interface that does not incorporate sufficient noise infusion.<sup>139</sup>  
 1867 For this reason, query interfaces should also log both queries and query results in order to deter  
 1868 and detect malicious use.

## 1869 **4.6 Validating a de-identified dataset**

1870 Agencies should validate datasets after they are de-identified to assure that the resulting dataset  
 1871 meets the agency's goals in terms of both data usefulness and privacy protection.

### 1872 **4.6.1 Validating data usefulness**

1873 De-identification decreases data quality and the usefulness of the resulting dataset. It is therefore  
 1874 important to assure that the de-identified dataset is still useful for the intended application—  
 1875 otherwise there is no reason to go through the expense and added risk of de-identification.

1876 Several approaches exist for validating data usefulness. For example, insiders can perform  
 1877 statistical calculations on both the original dataset and on the de-identified dataset and compare  
 1878 the results to see if the de-identification resulted in changes that are unacceptable. Agencies can  
 1879 engage trusted outsiders to examine the de-identified dataset and determine if the data can be  
 1880 used for the intended purpose.

### 1881 **4.6.2 Validating privacy protection**

1882 Several approaches exist for validating the privacy protection provided by de-identification,  
 1883 including:

- 1884 ● Examining the resulting data files to make sure that no identifying information is  
 1885 included in file data or metadata.
- 1886 ● Examining the resulting data files to make sure that the resulting data meet stated goals  
 1887 for ambiguity under a k-anonymity model, if such a standard is desired.
- 1888 ● Critically evaluating all default assumptions used by software that performs data  
 1889 modification or modeling.
- 1890 ● Conducting a *motivated intruder test* to see if reasonably competent outside individuals  
 1891 can perform re-identification using publicly available datasets. Motivations for a  
 1892 motivated intruder can include prurient interest; the goal of causing embarrassment or  
 1893 harm; revealing private facts about public figures; or engaging in a reputation attack.  
 1894 Details of the motivated intruder test can be found in *Anonymisation: code of practice,*  
 1895 *managing data protection risk*, published by the United Kingdom's Information

---

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

- 1896 Commissioner's Office.<sup>140</sup>
- 1897 ● Providing the team conducting the motivated intruder test with using confidential agency  
1898 data, to simulated what might happen in the result of a breach or a hostile insider.
- 1899 These approaches do not provide provable guarantees on the protection offered by de-  
1900 identification, but they may be useful as part of an overall agency risk assessment.<sup>141</sup>
- 1901 Applications that require provable privacy guarantees should rely on formal privacy methods  
1902 such as differential privacy when planning their data releases.
- 1903 Validating the privacy protection of de-identified data is greatly simplified by using validated de-  
1904 identification software, as discussed in Section 6, "Evaluation."

---

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

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

## 1905 **5 Software Requirements, Evaluation and Validation**

1906 Agencies performing de-identification should clearly define the requirements for de-  
1907 identification algorithms and the software that implements those algorithms. They should be sure  
1908 that the algorithms that they intend to use are validated, that the software that implements the  
1909 algorithms as expected, and the data that results from the operation of the software are correct.<sup>142</sup>

### 1910 **5.1 Evaluating Privacy Preserving Techniques**

1911 There have been decades of research in the field of statistical disclosure limitation and de-  
1912 identification. As the understanding of statistical disclosure limitation and de-identification have  
1913 evolved over time, agencies should not base their technical evaluation of a technique on the mere  
1914 fact that the technique has been published in the peer reviewed literature or that the agency has a  
1915 long history of using the technique and has not experienced any problems. Instead, it is necessary  
1916 to evaluate proposed techniques considering the totality of the scientific experience and with  
1917 regards to current threats.

1918 Traditional statistical disclosure limitation and de-identification techniques base their risk  
1919 assessments, in part, on an expectation of what kinds of data are available to an attacker to  
1920 conduct a linkage attack. Where possible, these assumptions should be documented and  
1921 published along with a technique description of the privacy-preserving techniques that are used  
1922 to transform datasets prior to release, so that they can be reviewed by external experts and the  
1923 scientific community.

1924 Because our understanding of privacy technology and the capabilities of privacy attacks are both  
1925 rapidly evolving, techniques that have been previously established should be periodically  
1926 reviewed. New vulnerabilities may be discovered in techniques that have been previously  
1927 accepted. Alternatively, it may be that new techniques are developed that allow agencies to re-  
1928 evaluate the tradeoffs that they have made with respect to privacy risk and data usability.

### 1929 **5.2 De-Identification Tools**

1930 A de-identification tool is a program that is involved in the creation of de-identified datasets.

#### 1931 **5.2.1 De-Identification Tool Features**

1932 De-identification tools might perform many functions, including:

- 1933 ● Detection of identifying information
- 1934 ● Calculation of re-identification risk
- 1935 ● Performing de-identification

---

<sup>142</sup> Please note that NIST is preparing a separate report on evaluating de-identification software and results.

- 1936       • Mapping identifiers to pseudonyms
- 1937       • Providing for the selective revelation of pseudonyms

1938 De-identification tools may handle a variety of data modalities. For example, tools might be  
 1939 designed for tabular data or for multimedia. Particular tools might attempt to de-identify all data  
 1940 types, or might be developed for specific modalities. A potential risk of using de-identification  
 1941 tools is that a tool might be equipped to handle some but not all of the different modalities in a  
 1942 dataset. For example, a tool might de-identifying the categorical information in a table according  
 1943 to a de-identification standard, but might not detect or attempt to address the presence of  
 1944 identifying information in a text field. For this reason, de-identification tools should be validated.  
 1945 For further information, see Section 6, “Software Requirements, Evaluation.”

1946 Appendix 8.7, “Specific De-Identification Tools,” provides a listing of some de-identification  
 1947 tools that were known at the time of this publication.

## 1948 **5.2.2 Data Provenance and File Formats**

1949 Output files created by de-identification tools and data masking tools can record provenance  
 1950 information, such as metadata regarding input datasets, the de-identification methods used, and  
 1951 the resulting decrease in data quality. Output files can also be explicitly marked to indicate that  
 1952 they have been de-identified. For example, de-identification profiles that are part of the Digital  
 1953 Imaging and Communications in Medicine (DICOM) specification indicate which elements are  
 1954 direct vs quasi identifiers, and which de-identification algorithms have been employed.<sup>143</sup>

## 1955 **5.2.3 Data Masking Tools**

1956 Data masking tools are programs that can perform removal or replacement of designated fields in  
 1957 a dataset while maintaining relationships between tables. These tools can be used to remove  
 1958 direct identifiers but generally cannot identify or modify quasi-identifiers in a manner consistent  
 1959 with a privacy policy or risk analysis.

1960 Data masking tools were developed to allow software developers and testers access to datasets  
 1961 containing realistic data while providing minimal privacy protection. Absent additional controls  
 1962 or data manipulations, data masking tools should not be used for de-identification of datasets that  
 1963 are intended for public release, and data masking tools should not be used as the sole mechanism  
 1964 to assure confidentiality in non-public data sharing.

## 1965 **5.3 Evaluating De-Identification Software**

1966 Once techniques are evaluated and approved, agencies should assure that the techniques are  
 1967 faithfully executed by their chosen software. Privacy software evaluation should consider the

---

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

1968 tradeoff between data usability and privacy protection.

1969 Privacy software evaluation should also seek to detect and minimize the chances of tool error  
1970 and user error.

1971 For example, agencies should verify:

1972 • That the software properly implements the chosen algorithms.

1973 • The software does not leak identifying information including in unexpected ways such as  
1974 through the inaccuracies of floating-point arithmetic or the differences in execution time  
1975 (if observable to an adversary).

1976 • The software has sufficient usability that it can be operated efficiently and without error.

1977 Agencies may also wish to evaluate the performance of the de-identification software, such as:

1978 • Efficiency. How long does it take to run on a dataset of a typical size?

1979 • Scalability. How much does it slow down when moving from a dataset of  $N$  to  $100N$ ?

1980 • Usability. Can users understand the user interface? Can users detect and correct their  
1981 errors? Is the documentation sufficient?

1982 • Repeatability. If the tool is run twice on the same dataset, are the results similar? If two  
1983 different people run the tool, do they get similar results?

1984 Ideally, software should be able to track the accumulated privacy leakage from multiple data  
1985 releases.

## 1986 **5.4 Evaluating Data Quality**

1987 Finally, agencies should evaluate the quality of the de-identified data to verify that it is sufficient  
1988 for the intended use. Approaches for evaluating the data quality include:

1989 • Verifying that single variable statistics and two-variable correlations remain relatively  
1990 unchanged.

1991 • Verifying that statistical distributions do not incur undue bias as a result of the de-  
1992 identification procedure.

1993 Agencies can create or adopt standards regarding the quality and accuracy of de-identified data.  
1994 If data accuracy cannot be well maintained along with data privacy goals, then the release of data  
1995 that is inaccurate for statistical analyses could potentially result in incorrect scientific  
1996 conclusions and incorrect policy decisions.



**1997 6 Conclusion**

1998 Government agencies can use de-identification technology to make datasets available to  
1999 researchers and the general public without compromising the privacy of people contained within  
2000 the data.

2001 Currently there are three primary models available for de-identification: agencies can make data  
2002 available with traditional de-identification techniques relying on suppression of identifying  
2003 information (direct identifiers) and manipulation of information that partially identifying (quasi-  
2004 identifiers); agencies can create synthetic datasets; and agencies can make data available through  
2005 a query interface. These models can be mixed within a single dataset, providing different kinds  
2006 of access for different users or intended uses.

2007 Privacy protection is strengthened when agencies employ formal models for privacy protection  
2008 such as differential privacy, because the mathematical models that these systems use are  
2009 designed to assure privacy protection irrespective of future data releases or developments in re-  
2010 identification technology. However, the mathematics underlying these systems is very new, and  
2011 there is little experience within the government in using these systems. Thus, these systems may  
2012 result in significant and at times unnecessary reduction in data quality when compared with  
2013 traditional de-identification approaches that do not offer formal privacy guarantees.

2014 Agencies that seek to use de-identification to transform privacy sensitive datasets into dataset  
2015 that can be publicly released should take care to establish appropriate governance structures to  
2016 support de-identification, data release, and post-release monitoring. Such structures will typically  
2017 include a Disclosure Review Board as well as appropriate education, training, and research  
2018 efforts.

2019 Finally, different countries have different standards and policies regarding the definition and use  
2020 of de-identified data. Information that is regarded as de-identified in one jurisdiction may be  
2021 regarded as being identifiable in another.

2022

## 2023 7 References

### 2024 7.1 Standards

- 2025 ● ASTM E1869-04(2014) Standard Guide for Confidentiality, Privacy, Access, and Data  
2026 Security Principles for Health Information Including Electronic Health Records
- 2027 ● DICOM PS3.15 2016d – Security and System Management Profiles Chapter E Attribute  
2028 Confidentiality Profiles, DICOM Standards Committee, NEMA 2016.  
2029 [http://dicom.nema.org/medical/dicom/current/output/html/part15.html#chapter\\_E](http://dicom.nema.org/medical/dicom/current/output/html/part15.html#chapter_E)
- 2030 ● HITRUST De-Identification Working Group (2015, March). De-Identification  
2031 Framework: A Consistent, Managed Methodology for the De-Identification of Personal  
2032 Data and the Sharing of Compliance and Risk Information. Frisco, TX:  
2033 HITRUST. Retrieved from <https://hitrustalliance.net/de-identification-license-agreement/>
- 2034 ● ISO/IEC 27000:2014 Information technology -- Security techniques -- Information  
2035 security management systems -- Overview and vocabulary
- 2036 ● ISO/IEC 24760-1:2011 Information technology -- Security techniques -- A framework  
2037 for identity management -- Part 1: Terminology and concepts
- 2038 ● ISO/TS 25237:2008(E) Health Informatics — Pseudonymization. ISO, Geneva,  
2039 Switzerland. 2008.
- 2040 ● ISO/IEC 20889 WORKING DRAFT 2016-05-30, Information technology – Security  
2041 techniques – Privacy enhancing data de-identification techniques. 2016.
- 2042 ● IHE IT Infrastructure Handbook, De-Identification, Integrating the Healthcare Enterprise,  
2043 June 6, 2014. [http://www.ihe.net/User\\_Handbooks/](http://www.ihe.net/User_Handbooks/)

### 2044 7.2 US Government Publications

- 2045 ● *Census Confidentiality and Privacy: 1790-2002, US Census Bureau, 2003.*  
2046 <https://www.census.gov/prod/2003pubs/conmono2.pdf>
- 2047 ● *Disclosure Avoidance Techniques at the US Census Bureau: Current Practices and*  
2048 *Research*, Research Report Series (Disclosure Avoidance #2014-02), Amy Lauger, Billy  
2049 Wisniewski, and Laura McKenna, Center for Disclosure Avoidance Research, US  
2050 Census. Bureau, September 26, 2014. [https://www.census.gov/srd/CDAR/cdar2014-02\\_Discl\\_Avoid\\_Techniques.pdf](https://www.census.gov/srd/CDAR/cdar2014-02_Discl_Avoid_Techniques.pdf)  
2051
- 2052 ● *Privacy and Confidentiality Research and the US Census Bureau, Recommendations*  
2053 *Based on a Review of the Literature*, Thomas S. Mayer, Statistical Research Division, US  
2054 Bureau of the Census. February 7, 2002.

- 2055 <https://www.census.gov/srd/papers/pdf/rsm2002-01.pdf>
- 2056 ● *Frequently Asked Questions—Disclosure Avoidance, Privacy Technical Assistance*  
 2057 *Center*, US Department of Education. October 2012 (revised July 2015)  
 2058 [http://ptac.ed.gov/sites/default/files/FAQ\\_Disclosure\\_Avoidance.pdf](http://ptac.ed.gov/sites/default/files/FAQ_Disclosure_Avoidance.pdf)
- 2059 ● *Guidance Regarding Methods for De-identification of Protected Health Information in*  
 2060 *Accordance with the Health Insurance Portability and Accountability Act (HIPAA)*  
 2061 *Privacy Rule*, U.S. Department of Health & Human Services, Office for Civil Rights,  
 2062 November 26, 2012.  
 2063 [http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveridentities/De-](http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveridentities/De-identification/hhs_deid_guidance.pdf)  
 2064 [identification/hhs\\_deid\\_guidance.pdf](http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveridentities/De-identification/hhs_deid_guidance.pdf)
- 2065 ● *OHRP-Guidance on Research Involving Private Information or Biological Specimens*  
 2066 *(2008)*, Department of Health & Human Services, Office of Human Research Protections  
 2067 (OHRP), August 16, 2008. <http://www.hhs.gov/ohrp/policy/cdebiol.html>
- 2068 ● *Data De-identification: An Overview of Basic Terms*, Privacy Technical Assistance  
 2069 *Center*, U.S. Department of Education. May 2013.  
 2070 [http://ptac.ed.gov/sites/default/files/data\\_deidentification\\_terms.pdf](http://ptac.ed.gov/sites/default/files/data_deidentification_terms.pdf)
- 2071 ● *Statistical Policy Working Paper 22 (Second version, 2005)*, Report on Statistical  
 2072 *Disclosure Limitation Methodology*, Federal Committee on Statistical Methodology,  
 2073 December 2005.
- 2074 ● *The Data Disclosure Decision, Department of Education* (ED) Disclosure Review Board  
 2075 (DRB), A Product of the Federal CIO Council Innovation Committee. Version 1.0, 2015.  
 2076 <http://go.usa.gov/xr68F>
- 2077 ● *National Center for Health Statistics Policy on Micro-data Dissemination*, Centers for  
 2078 *Disease Control*, July 2002.  
 2079 [https://www.cdc.gov/nchs/data/nchs\\_microdata\\_release\\_policy\\_4-02a.pdf](https://www.cdc.gov/nchs/data/nchs_microdata_release_policy_4-02a.pdf)
- 2080 ● *National Center for Health Statistics Data Release and Access Policy for Micro-data and*  
 2081 *Compressed Vital Statistics File*, Centers for Disease Control, April 26, 2011.  
 2082 [http://www.cdc.gov/nchs/nvss/dvs\\_data\\_release.htm](http://www.cdc.gov/nchs/nvss/dvs_data_release.htm)
- 2083 ● *Linking Data for Health Services Research: A Framework and Instructional Guide.*,  
 2084 *Dusetzina SB, Tyree S, Meyer AM, Meyer A, Green L, Carpenter WR. (Prepared by the*  
 2085 *University of North Carolina at Chapel Hill under Contract No. 290-2010-000141.)*  
 2086 *AHRQ Publication No. 14-EHC033-EF. Rockville, MD: Agency for Healthcare Research*  
 2087 *and Quality; September 2014.*

### 2088 7.3 Publications by Other Governments

- 2089       • *Privacy business resource 4: De-identification of data and information*, Office of the  
2090       Australian Information Commissioner, Australian Government, April 2014.  
2091       [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-<br/>2092       resources/privacy_business_resource_4.pdf)
- 2093       • *Opinion 05/2014 on Anonymisation Techniques*, Article 29 Data Protection Working  
2094       Party, 0829/14/EN WP216, Adopted on 10 April 2014
- 2095       • *Anonymisation: Managing data protection risk, Code of Practice 2012*, Information  
2096       Commissioner's Office. [https://ico.org.uk/media/for-  
organisations/documents/1061/anonymisation-code.pdf](https://ico.org.uk/media/for-<br/>2097       organisations/documents/1061/anonymisation-code.pdf). 108 pages
- 2098       • *The Anonymisation Decision-Making Framework*, Mark Elliot, Elaine Mackey, Kieron  
2099       O'Hara and Caroline Tudor, UKAN, University of Manchester, July 2016.  
2100       <http://ukanon.net/ukan-resources/ukan-decision-making-framework/>

### 2101 7.4 Reports and Books:

- 2102       • *Private Lives and Public Policies: Confidentiality and Accessibility of Government  
2103       Statistics (1993)*, George T. Duncan, Thomas B. Jabine, and Virginia A. de Wolf,  
2104       Editors; Panel on Confidentiality and Data Access; [\*Commission on Behavioral and  
2105       Social Sciences and Education; Division of Behavioral and Social Sciences and  
2106       Education\*](#); National Research Council, 1993. <http://dx.doi.org/10.17226/2122>
- 2107       • *Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk, Committee on  
2108       Strategies for Responsible Sharing of Clinical Trial Data, Board on Health Sciences  
2109       Policy*, Institute of Medicine of the National Academies, The National Academies Press,  
2110       Washington, DC. 2015.
- 2111       • P. Doyle and J. Lane, *Confidentiality, Disclosure and Data Access: Theory and Practical  
2112       Applications for Statistical Agencies*, North-Holland Publishing, Dec 31, 2001
- 2113       • George T. Duncan, Mark Elliot, Juan-José Salazar-Gonzalez, *Statistical Confidentiality:  
2114       Principles and Practice*, Springer, 2011
- 2115       • Cynthia Dwork and Aaron Roth, *The Algorithmic Foundations of Differential Privacy*  
2116       (Foundations and Trends in Theoretical Computer Science). Now Publishers, August 11,  
2117       2014. <http://www.cis.upenn.edu/~aaroht/privacybook.html>
- 2118       • Khaled El Emam, *Guide to the De-Identificaion of Personal Health Information*, CRC  
2119       Press, 2013
- 2120       • Khaled El Emam and Luk Arbuckle, *Anonymizing Health Data*, O'Reilly, Cambridge,  
2121       MA. 2013

- 2122 • K El Emam and B Malin, “Appendix B: Concepts and Methods for De-identifying  
2123 Clinical Trial Data,” in *Sharing Clinical Trial Data: Maximizing Benefits, Minimizing*  
2124 *Risk*, Institute of Medicine of the National Academies, The National Academies Press,  
2125 Washington, DC. 2015
- 2126 • Anco Hundepool, Josep Domingo-Ferrer, Luisa Franconi, Sarah Giessing, Eric Schulte  
2127 Nordholt, Keith Spicer, Peter-Paul de Wolf, *Statistical Disclosure Control*, Wiley,  
2128 September 2012.

## 2129 7.5 How-To Articles

- 2130 • Olivia Angiuli, Joe Blitstein, and Jim Waldo, How to De-Identify Your Data,  
2131 Communications of the ACM, December 2015.
- 2132 • Jörg Drechsler, Stefan Bender, Susanne Rässler, Comparing fully and partially synthetic  
2133 datasets for statistical disclosure control in the German IAB Establishment Panel. 2007,  
2134 United Nations, Economic Commission for Europe. Working paper, 11, New York, 8 p.  
2135 <http://fdz.iab.de/342/section.aspx/Publikation/k080530j05>
- 2136 • Ebaa Fayyumi and B. John Oommen, A survey on statistical disclosure control and  
2137 micro-aggregation techniques for secure statistical databases. 2010, *Software Practice*  
2138 *and Experience*. 40, 12 (November 2010), 1161-1188. DOI=10.1002/spe.v40:12  
2139 <http://dx.doi.org/10.1002/spe.v40:12>
- 2140 • Jingchen Hu, Jerome P. Reiter, and Quanli Wang, Disclosure Risk Evaluation for Fully  
2141 Synthetic Categorical Data, *Privacy in Statistical Databases*, pp. 185-199, 2014.  
2142 [http://link.springer.com/chapter/10.1007%2F978-3-319-11257-2\\_15](http://link.springer.com/chapter/10.1007%2F978-3-319-11257-2_15)
- 2143 • Matthias Templ, Bernhard Meindl, Alexander Kowarik and Shuang Chen, Introduction to  
2144 Statistical Disclosure Control (SDC), IHSN Working Paper No. 007, International  
2145 Household Survey Network, August 2014.  
2146 [http://www.ihsn.org/home/sites/default/files/resources/ihsn-working-paper-007-  
Oct27.pdf](http://www.ihsn.org/home/sites/default/files/resources/ihsn-working-paper-007-<br/>2147 Oct27.pdf)
- 2148 • Natalie Shlomo, Statistical Disclosure Control Methods for Census Frequency Tables,  
2149 *International Statistical Review* (2007), 75, 2, 199-217.  
2150 <https://www.jstor.org/stable/41508461>

2151

2152

2153 **7.6 Glossary**

2154 Selected terms used in the publication are defined below. Where noted, the definition is sourced  
2155 to another publication.

2156 **attribute:** “inherent characteristic.” (ISO 9241-302:2008)

2157 **attribute disclosure:** re-identification event in which an entity learns confidential information  
2158 about a data principal, without necessarily identifying the data principal (ISO/IEC 20889  
2159 WORKING DRAFT 2 2016-05-27)

2160 **anonymity:** “condition in identification whereby an entity can be recognized as distinct, without  
2161 sufficient identity information to establish a link to a known identity” (ISO/IEC 24760-1:2011)

2162 **anticipated re-identification rate:** when an organization contemplates performing re-  
2163 identification, the re-identification rate that the resulting de-identified data are likely to have.

2164 **attacker:** person seeking to exploit potential vulnerabilities of a system

2165 **attribute:** “characteristic or property of an entity that can be used to describe its state,  
2166 appearance, or other aspect” (ISO/IEC 24760-1:2011)<sup>144</sup>

2167 **brute force attack:** in cryptography, an attack that involves trying all possible combinations to  
2168 find a match

2169 **coded:** “1. identifying information (such as name or social security number) that would enable  
2170 the investigator to readily ascertain the identity of the individual to whom the private information  
2171 or specimens pertain has been replaced with a number, letter, symbol, or combination thereof  
2172 (i.e., the code); and 2. a key to decipher the code exists, enabling linkage of the identifying  
2173 information to the private information or specimens.”<sup>145</sup>

2174 **control:** “measure that is modifying risk. Note: controls include any process, policy, device,  
2175 practice, or other actions which modify risk.” (ISO/IEC 27000:2014)

2176 **covered entity:** under HIPAA, a health plan, a health care clearinghouse, or a health care  
2177 provider that electronically transmits protected health information (HIPAA Privacy Rule)

2178 **data subjects:** “persons to whom data refer” (ISO/TS 25237:2008)

---

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

<sup>145</sup> 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>

- 2179 **data use agreement:** executed agreement between a data provider and a data recipient that  
2180 specifies the terms under which the data can be used.
- 2181 **data universe:** All possible data within a specified domain.
- 2182 **dataset:** collection of data
- 2183 **dataset with identifiers:** a dataset that contains information that directly identifies individuals.
- 2184 **dataset without identifiers:** a dataset that does not contain direct identifiers
- 2185 **de-identification:** “general term for any process of removing the association between a set of  
2186 identifying data and the data subject” (ISO/TS 25237-2008)
- 2187 **de-identification model:** approach to the application of data de-identification techniques that  
2188 enables the calculation of re-identification risk (ISO/IEC 20889 WORKING DRAFT 2 2016-05-  
2189 27)
- 2190 **de-identification process:** “general term for any process of removing the association between a  
2191 set of identifying data and the data principal” [ISO/TS 25237:2008]
- 2192 **de-identified information:** “records that have had enough PII removed or obscured such that the  
2193 remaining information does not identify an individual and there is no reasonable basis to believe  
2194 that the information can be used to identify an individual” (SP800-122)
- 2195 **direct identifying data:** “data that directly identifies a single individual” (ISO/TS 25237:2008)
- 2196 **disclosure:** “divulging of, or provision of access to, data” (ISO/TS 25237:2008)
- 2197 **disclosure limitation:** “statistical methods [] used to hinder anyone from identifying an  
2198 individual respondent or establishment by analyzing published [] data, especially by  
2199 manipulating mathematical and arithmetical relationships among the data.”<sup>146</sup>
- 2200 **effectiveness:** “extent to which planned activities are realized and planned results achieved”  
2201 (ISO/IEC 27000:2014)
- 2202 **entity:** “item inside or outside an information and communication technology system, such as a  
2203 person, an organization, a device, a subsystem, or a group of such items that has recognizably  
2204 distinct existence” (ISO/IEC 24760-1:2011)
- 2205 **expert determination:** within the context of de-identification, expert determination refers to the  
2206 Expert Determination method for de-identifying protected health information in accordance with

---

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

- 2207 the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule.
- 2208 **Federal Committee on Statistical Methodology (FCSM):** “an interagency committee  
2209 dedicated to improving the quality of Federal statistics. The FCSM was created by the Office of  
2210 Management and Budget (OMB) to inform and advise OMB and the Interagency Council on  
2211 Statistical Policy (ICSP) on methodological and statistical issues that affect the quality of Federal  
2212 data.” (fscm.sites.usa.gov)
- 2213 **genomic information:** information based on an individual’s genome, such as a sequence of  
2214 DNA or the results of genetic testing
- 2215 **harm:** “any adverse effects that would be experienced by an individual (i.e., that may be  
2216 socially, physically, or financially damaging) or an organization if the confidentiality of PII were  
2217 breached” (SP800-122)
- 2218 **Health Insurance Portability and Accountability Act of 1996 (HIPAA):** the primary law in  
2219 the United States that governs the privacy of healthcare information
- 2220 **HIPAA:** see *Health Insurance Portability and Accountability Act of 1996*
- 2221 **HIPAA Privacy Rule:** “establishes national standards to protect individuals’ medical records  
2222 and other personal health information and applies to health plans, health care clearinghouses, and  
2223 those health care providers that conduct certain health care transactions electronically” (HIPAA  
2224 Privacy Rule, 45 CFR 160, 162, 164)
- 2225 **identification:** “process of using claimed or observed attributes of an entity to single out the  
2226 entity among other entities in a set of identities” (ISO/TS 25237:2008)
- 2227 **identifying information:** information that can be used to distinguish or trace an individual’s  
2228 identity, such as their name, social security number, biometric records, etc. alone, or when  
2229 combined with other personal or identifying information which is linked or linkable to a specific  
2230 individual, such as date and place of birth, mother’s maiden name, etc. (OMB M-07-16)
- 2231 **identifier:** “information used to claim an identity, before a potential corroboration by a  
2232 corresponding authenticator” (ISO/TS 25237:2008)
- 2233 **imputation:** “a procedure for entering a value for a specific data item where the response is  
2234 missing or unusable.” (OECD Glossary of Statistical Terms)
- 2235 **inference:** “refers to the ability to deduce the identity of a person associated with a set of data  
2236 through “clues” contained in that information. This analysis permits determination of the  
2237 individual’s identity based on a combination of facts associated with that person even though



- 2238 specific identifiers have been removed, like name and social security number” (ASTM E1869<sup>147</sup>)
- 2239 **k-anonymity:** a technique “to release person-specific data such that the ability to link to other  
2240 information using the quasi-identifier is limited.”<sup>148</sup> k-anonymity achieves this through  
2241 suppression of identifiers and output perturbation.
- 2242 **l-diversity:** a refinement to the k-anonymity approach which assures that groups of records  
2243 specified by the same identifiers have sufficient diversity to prevent inferential disclosure<sup>149</sup>
- 2244 **masking:** the process of systematically removing a field or replacing it with a value in a way that  
2245 does not preserve the analytic utility of the value, such as replacing a phone number with  
2246 asterisks or a randomly generated pseudonym<sup>150</sup>
- 2247 **motivated intruder test:** “The ‘motivated intruder’ is taken to be a person who starts without  
2248 any prior knowledge but who wishes to identify the individual from whose personal data the  
2249 anonymised data has been derived. This test is meant to assess whether the motivated intruder  
2250 would be successful.”<sup>151</sup>
- 2251 **noise:** “a convenient term for a series of random disturbances borrowed through communication  
2252 engineering, from the theory of sound. In communication theory noise results in the possibility of  
2253 a signal sent, x, being different from the signal received, y, and the latter has a probability  
2254 distribution conditional upon x. If the disturbances consist of impulses at random intervals it is  
2255 sometimes known as “shot noise”.” (OECD Glossary of Statistical Terms)
- 2256 **non-deterministic noise:** a random value that cannot be predicted
- 2257 **non-public personal information:** information about a person that is not publicly known; called  
2258 “private information” in some other publications.
- 2259 **personal identifier:** “information with the purpose of uniquely identifying a person within a  
2260 given context” (ISO/TS 25237:2008)
- 2261 **personal data:** “any information relating to an identified or identifiable natural person (*data*

---

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

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

<sup>149</sup> 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.

<sup>150</sup> El Emam, Khaled and Luk Arbuckle, Anonymizing Health Data, O’Reilly, Cambridge, MA. 2013

<sup>151</sup> Anonymisation: code of practice, managing data protection risk. Information Commissioner’s Office. 2012.  
<https://ico.org.uk/media/1061/anonymisation-code.pdf>

2262 *subject*)” (ISO/TS 25237:2008)

2263 **personally identifiable information (PII):** “Any information about an individual maintained by  
2264 an agency, including (1) any information that can be used to distinguish or trace an individual’s  
2265 identity, such as name, social security number, date and place of birth, mother’s maiden name, or  
2266 biometric records; and (2) any other information that is linked or linkable to an individual, such  
2267 as medical, educational, financial, and employment information.”<sup>152</sup> (SP800-122)

2268 **perturbation-based methods:** “Perturbation-based methods falsify the data before publication  
2269 by introducing an element of error purposely for confidentiality reasons. This error can be  
2270 inserted in the cell values after the table is created, which means the error is introduced to the  
2271 output of the data and will therefore be referred to as output perturbation, or the error can be  
2272 inserted in the original data on the microdata level, which is the input of the tables one wants to  
2273 create; the method with then be referred to as data perturbation—input perturbation being the  
2274 better but uncommonly used expression. Possible methods are: - rounding; - random  
2275 perturbation; - disclosure control methods for microstatistics applied to macrostatistics.” (OECD  
2276 Glossary of Statistical Terms)

2277 **privacy:** “freedom from intrusion into the private life or affairs of an individual when that  
2278 intrusion results from undue or illegal gathering and use of data about that individual” (ISO/IEC  
2279 2382-8:1998, definition 08-01-23)

2280 **protected health information (PHI):** “individually identifiable health information: (1) Except  
2281 as provided in paragraph (2) of this definition, that is: (i) Transmitted by electronic media;  
2282 (ii) Maintained in electronic media; or (iii) Transmitted or maintained in any other form or  
2283 medium. (2) *Protected health information* excludes individually identifiable health information  
2284 in: (i) Education records covered by the Family Educational Rights and Privacy Act, as  
2285 amended, [20 U.S.C. 1232g](#); (ii) Records described at [20 U.S.C. 1232g\(a\)\(4\)\(B\)\(iv\)](#); and  
2286 (iii) Employment records held by a covered entity in its role as employer.” (HIPAA Privacy  
2287 Rule, 45 CFR 160.103)

2288 **pseudonymization:** a particular type of de-identification that both removes the association with  
2289 a data subject and adds an association between a particular set of characteristics relating to the  
2290 data subject and one or more pseudonyms.<sup>153</sup> Typically, pseudonymization is implemented by  
2291 replacing direct identifiers with a pseudonym, such as a randomly generated value.

2292 **pseudonym:** “personal identifier that is different from the normally used personal identifier.”  
2293 (ISO/TS 25237:2008)

---

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

<sup>153</sup> 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.”

- 2294 **quasi-identifier:** a variable that can be used to identify an individual through association with  
2295 another variable
- 2296 **recipient:** “natural or legal person, public authority, agency or any other body to whom data are  
2297 disclosed” (ISO/TS 25237:2008)
- 2298 **re-identification:** general term for any process that restores the association between a set of de-  
2299 identified data and a data subject
- 2300 **re-identification risk:** a measure of the extent to which an entity is threatened by the re-  
2301 identification of records within a dataset, typically a function of: (i) the adverse impacts that  
2302 would arise if the re-identification would occur; and (ii) the likelihood of occurrence.
- 2303 **re-identification rate:** the percentage of records in a dataset that can be re-identified.
- 2304 **risk:** “A measure of the extent to which an entity is threatened by a potential circumstance or  
2305 event, and typically a function of: (i) the adverse impacts that would arise if the circumstance or  
2306 event occurs; and (ii) the likelihood of occurrence.” (CNSSI No. 4009)
- 2307 **risk assessment:** “The process of identifying, estimating, and prioritizing risks to organizational  
2308 operations (including mission, functions, image, reputation), organizational assets, individuals,  
2309 other organizations, and the Nation, resulting from the operation of an information system. Part  
2310 of risk management, incorporates threat and vulnerability analyses, and considers mitigations  
2311 provided by security controls planned or in place. Synonymous with risk analysis.” (NIST SP  
2312 800-39)
- 2313 **safe harbor:** within the context of de-identification, safe harbor refers to the Safe Harbor  
2314 method for de-identifying protected health information in accordance with the Health Insurance  
2315 Portability and Accountability Act (HIPAA) Privacy Rule.
- 2316 **synthetic data generation:** a process in which seed data are used to create artificial data that has  
2317 some of the statistical characteristics as the seed data

## 2318 **7.7 Specific De-Identification Tools**

2319 This appendix provides a list of de-identification tools.

2320 **NOTE**

2321 *Specific products and organizations identified in this report were used in order to perform the*  
2322 *evaluations described. In no case does such identification imply recommendation or*  
2323 *endorsement by the National Institute of Standards and Technology, nor does it imply that*  
2324 *identified are necessarily the best available for the purpose.*

### 2325 **7.7.1 Tabular Data**

2326 Most de-identification tools designed for tabular data implement the k-Anonymity model. Many  
 2327 directly implement the HIPAA Privacy Rule's Safe Harbor standard. Tools that are currently  
 2328 available include:

2329 **AnonTool** is a German-language program that supports the k-anonymity framework.  
 2330 [http://www.tmf-ev.de/Themen/Projekte/V08601\\_AnonTool.aspx](http://www.tmf-ev.de/Themen/Projekte/V08601_AnonTool.aspx)

2331 **ARX** is an open source data de-identification tool written in Java that implements a variety of  
 2332 academic de-identification models, including k-anonymity, Population uniqueness,<sup>154</sup> k-Map,  
 2333 Strict-average risk,  $\ell$ -Diversity,<sup>155</sup> t-Closeness,<sup>156</sup>  $\delta$ -Disclosure privacy,<sup>157</sup> and  $\delta$ -presence.  
 2334 <http://arx.deidentifier.org/>

2335 **Cornell Anonymization Toolkit** is an interactive tool that was developed by the Computer  
 2336 Science Department at Cornell University<sup>158</sup> for performing de-identification. It can perform data  
 2337 generalization, risk analysis, utility evaluation, sensitive record manipulation, and visualization  
 2338 functions. <https://sourceforge.net/projects/anony-toolkit/>

2339 **Open Anonymizer** implements the k-anonymity framework.  
 2340 <https://sourceforge.net/projects/openanonymizer/>

2341 **Privacy Analytics Eclipse** is a comprehensive de-identification platform that can de-identify  
 2342 multiple linked tabular datasets to HIPAA or other de-identification standards. The program runs  
 2343 on Apache SPARK to allow de-identification of massive datasets, such as those arising in  
 2344 medical research. <http://www.privacy-analytics.com/software/privacy-analytics-core/>

2345  **$\mu$ -ARGUS** was developed by Statistics Netherlands for microdata release. The program was  
 2346 originally written in Visual Basic and was rewritten into C/C++ for an Open Source release. The  
 2347 program runs on Windows and Linux. <http://neon.vb.cbs.nl/casc/mu.htm>

2348 **sdcMicro** is a package for the popular open source R statistical platform that implements a  
 2349 variety of statistical disclosure controls. A full tutorial is available, as are prebuilt binaries for

---

<sup>154</sup> 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

<sup>155</sup> 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>

<sup>156</sup> 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

<sup>157</sup> 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>

<sup>158</sup> X. Xiao, G. Wang, and J. Gehrke. Interactive anonymization of sensitive data. In *SIGMOD Conference*, pages 1051–1054, 2009.

2350 Windows and OS X. <https://cran.r-project.org/web/packages/sdcMicro/>

2351 **SECRET**A, a tool for evaluating and comparing anonymizations. According to the website,  
2352 “SECRETA supports Incognito, Cluster, Top-down, and Full subtree bottom-up algorithms for  
2353 datasets with relational attributes, and COAT, PCTA, Apriori, LRA and VPA algorithms for  
2354 datasets with transaction attributes. Additionally, it supports the RMERGER, TMERGER, and  
2355 RTMERGER bounding methods, which enable the anonymization of RT-datasets by combining  
2356 two algorithms, each designed for a different attribute type (e.g., Incognito for relational  
2357 attributes and COAT for transaction attributes).” <http://users.uop.gr/~poulis/SECRET>A/

2358 **UTD Anonymization Toolbox** is an open source tool developed by the University of Texas  
2359 Dallas Data Security and Privacy Lab using funding provided by the National Institutes of  
2360 Health, the National Science Foundation, and the Air Force Office of Scientific Research.

### 2361 7.7.2 Free Text

2362 **BoB, a best-of-breed automated text de-identification system for VHA clinical**  
2363 **documents**,<sup>159</sup> developed by the Meystre Lab at the University of Utah School of Medicine.  
2364 <http://meystrelab.org/automated-ehr-text-de-identification/>

2365 **MITRE Identification Scrubber Toolkit (MIST)** is an open source tool for de-identifying free  
2366 format text. <http://mist-deid.sourceforge.net>

2367 **Privacy Analytics Lexicon** performs automated de-identification of unstructured data (text).  
2368 <http://www.privacy-analytics.com/software/privacy-analytics-lexicon/>

### 2369 7.7.3 Multimedia

2370 **DicomCleaner** is an open source tool that removes identifying information from medical  
2371 imagery in the DICOM format. DicomCleaner. The program can remove both metadata from the  
2372 DICOM file and black out identifying information that has been “burned in” to the image area.  
2373 DicomCleaner can perform redaction directly of compressed JPEG blocks so that the medical  
2374 image does not need to be decompressed and re-compressed, a procedure that can introduce  
2375 artifacts. <http://www.dclunie.com/pixelmed/software/webstart/DicomCleanerUsage.html>

---

<sup>159</sup> [BoB, a best-of-breed automated text de-identification system for VHA clinical documents](#). Ferrández O, South BR, Shen S, Friedlin FJ, Samore MH, Meystre SM. J Am Med Inform Assoc. 2013 Jan 1;20(1):77-83. doi: 10.1136/amiajnl-2012-001020. Epub 2012 Sep 4.