参考資料２

匿名化についての考え方（案）

2022.5

東京大学政策評価研究教育センター

１．本資料の趣旨

どの程度の匿名化が必要かは、２（２）で述べるとおり、加工対象となる個人情報データベース等の性質も勘案して個別具体的に判断する必要があります。本資料は、自治体様がデータを匿名化処理するに当たり、どのような匿名化処理を行えば、十分な匿名性が確保できるかについて、法令や匿名化処理についての研究等を基に御提案するものです。

この「考え方（案)」について、御了解いただけましたら、東京大学政策評価研究教育センター（CREPE）で、この「考え方（案)」を実行するプログラムを配布したいと考えています。

２．法令上の位置付け

（１）本プロジェクトの適用関係

本プロジェクトは、デジタル社会形成整備法（令和３年法律第 37 号。自治体関係は令和５年４月１日から施行）による個人情報保護法の改正前に行うものであり、国の個人情報保護法の適用を受けず、各自治体の個人情報保護条例によって規律されることになります。

また、デジタル社会形成整備法による改正後の個人情報保護法においても、「学術研究の目的のために保有個人情報を提供するとき」は、「利用目的以外の目的のために保有個人情報を……提供することができる。」（改正後の第 69 条第２項第４号）こととされており、個人情報保護法上の匿名加工情報制度がそのまま適用されるわけではありません。

しかし、３で述べるとおり、本プロジェクトでは、自治体様の多様なニーズに応えるため、匿名加工情報制度において求められる水準の匿名加工を行うという選択肢も御用意しています。

（２）匿名加工情報制度の概要

個人情報保護法上の匿名加工情報制度については、以下のような定めがあります。

○個人情報の保護に関する法律（平成十五年法律第五十七号）
（定義）
第二条　（略）
２～８　（略）

9　この法律において「匿名加工情報」とは、次の各号に掲げる個人情報の区分に応じて当該各号に定める措置を講じて特定の個人を識別することができないように個人情報を加工して得られる個人に関する情報であって、当該個人情報を復元することができないようにしたものをいう。

一　第一項第一号に該当する個人情報　当該個人情報に含まれる記述等の一部を削除すること（当該一部の記述等を復元することのできる規則性を有しない方法により他の記述等に置き換えることを含む。）。

二　第一項第二号に該当する個人情報　当該個人情報に含まれる個人識別符号の全部を削除すること（当該個人識別符号を復元することのできる規則性を有しない方法により他の記述等に置き換えることを含む。）。

10　（略）

（匿名加工情報の作成等）

第三十六条　個人情報取扱事業者は、匿名加工情報（匿名加工情報データベース等を構成するものに限る。以下同じ。）を作成するときは、特定の個人を識別すること及びその作成に用いる個人情報を復元することができないようにするために必要なものとして個人情報保護委員会規則で定める基準に従い、当該個人情報を加工しなければならない。

２～６　（略）

○個人情報の保護に関する法律施行規則（平成二十八年個人情報保護委員会規則第三号）

（匿名加工情報の作成の方法に関する基準）

第十九条　法第三十六条第一項の個人情報保護委員会規則で定める基準は、次のとおりとする。

一　個人情報に含まれる特定の個人を識別することができる記述等の全部又は一部を削除すること（当該全部又は一部の記述等を復元することのできる規則性を有しない方法により他の記述等に置き換えることを含む。）。

二　個人情報に含まれる個人識別符号の全部を削除すること（当該個人識別符号を復元することのできる規則性を有しない方法により他の記述等に置き換えることを含む。）。

三　個人情報と当該個人情報に措置を講じて得られる情報とを連結する符号（現に個人情報取扱事業者において取り扱う情報を相互に連結する符号に限る。）を削除すること（当該符号を復元することのできる規則性を有しない方法により当該個人情報と当該個人情報に措置を講じて得られる情報を連結することができない符号に置き換えることを含む。）。

四　特異な記述等を削除すること（当該特異な記述等を復元することのできる規則性を有しない方法により他の記述等に置き換えることを含む。）。

五　前各号に掲げる措置のほか、個人情報に含まれる記述等と当該個人情報を含む個人情報データベース等を構成する他の個人情報に含まれる記述等との差異その他の当該個人情報データベース等の性質を勘案し、その結果を踏まえて適切な措置を講ずること。

また、法律や規則以外にも、「個人情報の保護に関する法律についてのガイドライン（匿名加工情報編）」（平成 28 年 11 月（平成 29 年 3 月一部改正）個人情報保護委員会）、「個人情報保護委員会事務局レポート～パーソナルデータの利活用促進と消費者の信頼性確保の両立に向けて～」（2017 年 2 月個人情報保護委員会事務局）、国立情報学研究所匿名加工情報に関する技術検討ワーキンググループ「匿名加工情報の適正な加工の方法に関する報告書」（2017 年 2 月 21 日版）など、参考になるものはありますが、いずれも機械的に判断で

きるものではなく、特に5号については、「加工対象となる個人情報データベース等の性質によって加工の対象及び加工の程度は変わり得るため、どの情報をどの程度加工する必要があるかは、加工対象となる個人情報データベース等の性質も勘案して個別具体的に判断する必要がある」（同ガイドラインp. 13）とされています。

なお、個人情報保護法の「匿名加工情報」の要件である「特定の個人を識別することができない」こと、及び「個人情報を復元することができない」ことは、一般人及び一般的な事業者の能力、手法等を基準として判断され、あらゆる手法によって特定・復元できないよう技術的側面から全ての可能性を排除することまで求めるものではないとされています。

○「個人情報の保護に関する法律についてのガイドライン（匿名加工情報編）」p. 4

法において「特定の個人を識別することができる」とは、情報単体又は複数の情報を組み合わせて保存されているものから社会通念上そのように判断できるものをいい、一般人の判断力又は理解力をもって生存する具体的な人物と情報の間に同一性を認めるに至ることができるかどうかによるものである。匿名加工情報に求められる「特定の個人を識別することができない」という要件は、あらゆる手法によって特定することができないよう技術的側面から全ての可能性を排除することまでを求めるものではなく、少なくとも、一般人及び一般的な事業者の能力、手法等を基準として当該情報を個人情報取扱事業者又は匿名加工情報取扱事業者が通常の方法により特定できないような状態にすることを求めるものである。

……「当該個人情報を復元することができないようにしたもの」という要件は、あらゆる手法によって復元することができないよう技術的側面から全ての可能性を排除することまでを求めるものではなく、少なくとも、一般人及び一般的な事業者の能力、手法等を基準として当該情報を個人情報取扱事業者又は匿名加工情報取扱事業者が通常の方法により復元できないような状態にすることを求めるものである。

3．具体的匿名化処理

2（1）で述べたとおり、本プロジェクトでは、個人情報保護法上の匿名加工情報制度がそのまま適用されるわけではありませんが、自治体様の多様なニーズに応えるため、匿名化に関して、2つの選択肢を御用意しています。

Ⅰ　簡易な匿名化
Ⅱ　高度な匿名化（匿名加工情報制度において求められる水準の匿名加工）

Ⅰでは、匿名加工情報制度において求められる水準の匿名加工とは必ずしも言えないものの、氏名、個人番号（いわゆるマイナンバー）、住所等の削除や、宛名番号、世帯番号のハッシュ化（後述）等を行います。Ⅱはより匿名性が高いものとなりますが、その分、データの有用性は低下します。ⅠはⅡと比べて匿名性が低いものとなりますが、CREPEが御提供する分析結果は質・精度の高いものとなります。自治体様におかれては、こうしたトレードオフに御留意のうえ、いずれの匿名化手法を取るか御判断ください。

「Ⅰ　簡易な匿名化」を選択する場合、黒丸数字❶〜❸の流れ、「Ⅱ　高度な匿名化」を

選択する場合、白丸数字⓪～⑦の流れで匿名化処理を行うことを想定しています。

（事前処理）

❶⓪　氏名、個人番号（いわゆるマイナンバー）、住所等の事前削除

（匿名化処理）

①　特異な世帯の世帯番号の秘匿

❶②　宛名番号、世帯番号のハッシュ化

❷③　宛名番号をキーにして、複数年のデータの結合

④　所得、賦課額等のトップコーディング

❸⑤　生年月日の月単位への丸め

⑥　生年月、性別等について、「3-匿名性」を判定し、「3-匿名性」を満たさないものは、「3-匿名性」を満たすまで秘匿

⑦　各世帯番号（のハッシュ値）について、ランダムサンプリング

【解説】

① 特異な世帯の世帯番号の秘匿

特異な世帯は、それだけで個人識別性が高くなります。

そこで、特異な世帯[1]の世帯番号を秘匿します。

❶② 宛名番号、世帯番号のハッシュ化

宛名番号、世帯番号[2]については、ハッシュ化[3]を行います。

これらは通常、一般に知りうるものではなく、知りえたとしてもそこからすぐに個人情報にアクセスできるものでもありませんが、自治体内で様々なシステムを通じて横断的に使われていることもあり、万が一、流出したときの影響が大きいため、ハッシュ化を行います。年度をまたがる個人の紐付けは、このハッシュ化した宛名番号により行います。

ハッシュ化の方法としては、鍵付きハッシュ関数を用いるものとし（鍵となる秘密の文字列は自治体限りとしていただきます）、事務局レポート p. 21 において利用が推奨されている CRYPTREC（Cryptography Research and Evaluation Committees）により公開されている電子政府推奨暗号リスト[4]において挙げられているハッシュ関数のアルゴリズムを用います。

❷③ 宛名番号をキーにして、複数年のデータの結合

宛名番号[5]をキーにして、複数年のデータを個人単位で結合します。

「II　高度な匿名化」の場合は、生年月、性別、郵便番号（郵便番号がデータセットに入

---

[1] 「特異な世帯」の基準は公表していません。プロジェクト参加確定後、お示しします。

[2] 自治体内において世帯を一意に識別するために付番した番号のことです。

[3] ハッシュ化とは、元のデータから一定の計算手順に従ってハッシュ値と呼ばれる規則性のない値を求め、その値によって元のデータを置き換えることにより、データを不可逆的に別の形に変える方法をいいます。

例としては以下のようなイメージです。

宛名番号	ハッシュ化した宛名番号
12345678	aks;ldfjpawnefdoiewadlksfdajf
23456789	k3298refcsna3489elsjodslerire
34567890	kjlz-9erj8ufdjiawe8re;ofjwloe

ここで、「aks;ldfjpawnefdoiewadlksfdajf」から、元の「12345678」を復元することは不可能です。そのため、ハッシュ化を行えば当方が宛名番号を知りえないまま、年度をまたぐ個人の紐付けが可能となります。

[4] CRYPTREC 暗号リスト（電子政府推奨暗号リスト）（http://www.cryptrec.go.jp/list.html）。

[5] 自治体内において個人、法人を一意に識別するために付番した番号のことです。自治体によって、「個人コード」など他の名称で呼ばれることもあります。「個人番号」、「住記個人番号」と呼ばれることもありますが、番号法に基づく「個人番号」（いわゆるマイナンバー）とは異なります。

っている場合。⑥において同じ。）が年度によって異なる場合は、最も古いものを優先します。自治体内で転居した場合は、年度によって郵便番号が異なることがありますが、ある郵便番号の場所から別の郵便番号の場所に転居したという情報は、個人識別性が高いため、最も古い郵便番号を保持する（転居情報を隠す）ことによって、匿名性を確保します。

④　所得、賦課額等のトップコーディング

所得、賦課額、控除額等の金額については、同じ性別・年代ごとに 0.5%を対象として（ただし、この数が 10 人に満たない場合は 10 人を対象とします。以下同じ）、トップコーディングを行います。トップコーディングとは、加工対象に含まれる情報のうち、特に数値の大きいグループについてまとめる処理のことです。

今回は、同じ性別・年代（西暦の生年の上 3 桁）ごとに各項目（例：所得であれば所得、賦課額であれば賦課額）の上位 0.5%を、同じ値（上位 0.5%の平均値）に変換します。これにより、例えば、「男性・1970 年代生まれ」が 1 万人いる自治体であれば、上位 50 人が同じ所得となり、人数の少ない高所得者層であっても高所得であるからという理由から個人を特定するのが困難になります。

この結果、データセット中に男女それぞれで 1930 年代～2020 年代生まれが少なくとも 10 人ずついるとすると、10 人×2（男女）×10（年代）＝200 人が少なくともトップコーディングされることとなり、人口 4 万以上の場合はこれより多くの人(少なくとも人口の 0.5%)がトップコーディングの対象となります。

❸⑤　生年月日の月単位への丸め

生年月日については、「日」の情報を落として、月単位に丸めます。このときの丸め方としては、生年月日の前日の年と月に丸めます。

例えば、2001 年 1 月 1 日生まれの人と 2001 年 1 月 2 日生まれの人がいれば、前者は 2021 年 1 月 1 日時点で 20 歳であり、後者は 19 歳です。行政の取扱い上、1 日時点の年齢で判断されることが多いため（学齢等）、前者については前日（2020 年 12 月 31 日）の年と月である「2020 年 12 月」に、後者については前日（2021 年 1 月 1 日）の年と月である「2021 年 1 月」に丸めます。

もっとも、生まれ月まで入っていると、個人を特定するリスクが高くなりますが、この問題については⑥の処理で対応します。

⑥　生年月、性別、郵便番号について、「3-匿名性」を判定し、「3-匿名性」を満たさないものは、「3-匿名性」を満たすまで秘匿

データセットにある変数のうち、生年月、性別、郵便番号については、他者が容易に情報を得ることができることが多く、これらから個人を特定するリスクを避ける必要がありま

す。「3-匿名性」[6]が確保されれば、生年月、性別、郵便番号が同じ個人が少なくとも 3 人（自分を除けば 2 人）いることが保証されるので、生年月、性別、郵便番号の情報を得ても、当該個人である確率は 1/3 以下となります（さらに、後述の⑦の処理を行うことにより 1/6 以下となります）。

しかし、例えば、特に高齢である方、過疎地域に住む方などは、「3-匿名性」を満たさないことがあります。その場合、「3-匿名性」を満たすよう、生年月、性別、郵便番号のいずれかを秘匿[7]します。

秘匿の順序は、以下のとおりとします。

1. 生年月を四半期（1～3 月、4～6 月、7～9 月、10～12 月）にする
2. 郵便番号の 7 桁目を秘匿
3. 郵便番号の 6 桁目を秘匿
4. 郵便番号の 5 桁目を秘匿
5. 郵便番号の 4 桁目を秘匿
6. 郵便番号を秘匿
7. 生年月を半期（1～6 月、7～12 月）にする
8. 生年月の月を秘匿
9. 生年を 5 年単位で丸め
10. 生年の下一桁を秘匿
11. 生年を秘匿

これにより、すべての個人が「3-匿名性」を満たすこととなります。

⑦　各世帯番号（のハッシュ値）について、ランダムサンプリング

⑥で得られたデータセットのうち各世帯番号（のハッシュ値）について、50%の確率でサンプルに入れるかどうかをランダムで決めます（ランダムサンプリング）[8]。

これは、個人の特定ができない状況でも、個人の情報が得られてしまうリスクに対処するためです。

例えば、「1932 年 1 月生まれ・女性・郵便番号 9876543」が 3 人いる場合、これらの情報

---

[6] 「k-匿名性を満たす」とは、対象となるデータセット内に、同じ準識別子（今回の場合は、生年月、性別、郵便番号）の組合せを持つデータが k 件以上存在することをいいます。事務局レポート p. 32 では、「匿名加工情報は、……一般公開されるものではないから、上記で準識別子とされている情報の項目について、匿名加工情報データベース等との関係で k ≧ 2 となるように加工することは必ずしも求められない。ただし、匿名加工情報が第三者に提供される態様や利用形態を考慮した上で、必要に応じてこのような考え方を取り入れることが望ましい。」としています。

[7] 空欄に置き換えるか、別の値に置き換えますが、いずれにせよ、「3-匿名性」が満たされるまでこの処理を行います。

[8] この際、ある個人が年度間で異なる世帯番号を持つことが考えられるため、データセット中のいずれの年度でも、1 世帯内でサンプルに入っている構成員と入っていない構成員がいないように処理します。

からは 3 人のうち誰かを特定できず、「3-匿名性」を満たします。しかし、データセットの中で、「1932 年 1 月生まれ・女性・郵便番号 9876543」の 3 人の所得が全員「0 円」だったとすると、3 人の誰かは特定できなくても、所得が「0 円」であることは分かってしまいます。

そこで、50%のランダムサンプリングを行うことにより、個人がデータセットに入っている確率が、個人がデータセットに入っていない確率を上回らないようにします。そうすることにより、上記のような推論が成り立つ可能性が、上記のような推論が成り立たない確率を上回らないことになります。

また、個人単位ではなく、世帯単位でサンプリングを行うのは、1 世帯内でサンプルに入っている構成員と入っていない構成員がいないようにするためです。

なお、50%は、あくまで確率であるため、サンプル数が元のデータセットのレコード数のちょうど 50%に一致しない可能性があります。例えば、人口 10 万ちょうどの自治体について⑦の処理を行った場合、サンプル数は平均して 5 万ちょうどとなりますが、場合によって 49892 人になったり 50006 人になったりします。

４．CREPE におけるデータ取扱規則等

CREPE では、個人識別行為の禁止等を定めた「自治体税務データ活用プロジェクトにおけるデータ取扱規則」、「自治体税務データ活用プロジェクトにおける安全管理措置等に関する規程」を定めており、個人情報の管理に万全を尽くしています。