2B4-OS-15a-4

匿名化の実社会での利用に向けての技術課題

Technical Problem of Anonymity for Real Society Application

中川裕志[*1]
Hiroshi Nakagawa

[*1] 東京大学
The University of Tokyo

This document describes several technical problems which we are possibly to face when personal data anonymity would actually be applied to real society.

1. はじめに

ビッグデータとりわけパーソナルデータの活用が 2013 年 6 月に政府指針として打ち出された. 前後して JR 東日本が収集した顧客の Suica データを, 日立を経由しての他業者への提供することに対する反対意見が噴出するという騒動が起こった. この状況の下で政府のパーソナルデータに関する検討会が行われ, 技術検討ワーキング WG 報告[佐藤 2013](以下では「報告書」と略記する.)が同年 12 月 10 日に公表された.

この報告書は技術的レベルの高い内容だが, そこで示された方向性に対して IT 業界から, パーソナルデータを含むビッグデータの扱うビジネスを萎縮させるとして反対論があがっている.

一方, パーソナルデータに関連する法制度に関しては, 日本は不十分であるとして, EU からはゲノム情報などの有用な情報の輸入を禁止されているという状況を考慮すると, それを改善する法整備は喫緊と課題といえる.

本論文では, この報告を念頭において, 匿名化を現実社会で使うにあたっての技術課題, 制度設計について述べる.

2. 匿名化における基本概念

個人データの発生源である人をデータ源の個人, パーソナルデータを何らかの手段で多数の個人から集め, それを使った事業を行う人や組織をデータ事業者とする. データ事業者がさらに別の人や組織に自らが収集したデータを渡す場合, 受け手をデータ受領者とする.

個人データの収集において, データ源の個人は, データ事業者が示すデータ利用の許諾に関する文書に同意すれば, 許諾文の範囲でデータ事業者はデータ利用ができる. ただし, 許諾文で想定されるすべての利用法を網羅的に記述することは困難である.

収集される個人データは通常以下の要素からなる.

I. 個人 ID(氏名)
II. 疑似 ID(性別, 住所, 年齢, 国籍, など)
III. その他のデータ
IV. センシティブ情報:III.のその他の情報のうち, 人種, 宗教, 病名, 収入など他人に知られたくない情報をセンシティブ情報という. ただし, 実はその定義が難しい. これについては5節で触れる.

従来, 個人情報保護法では, これらのうち個人を特定する個人情報とは, ほぼ「個人ID」だけを意味しているとされてきた. だが, データ源の個人を特定できうる情報として II. 疑似 ID も個人情報と見なせる. しかし, III. それ以外の情報でも, 滞在場所, 通勤経路, 購買履歴などが集積すると個人を特定する可能性があるため, 疑似 ID と見なせるので, 当然個人情報とみなせる. 以上の考察から, 上記 I.II.III.(IV.)のすべてを個人を識別する情報と解釈できる.

報告書の主張のひとつは識別と特定を以下のように精密に定義したことである.

- 「特定」とは,「ある情報が誰の情報であるかが分かること」
- 「識別」とは,「ある情報が誰か一人の情報であることが分かること」

この定義は匿名化の処理範囲を明確化し, 技術的な検討がしやすくなった点で大きな前進である.

3. 完全な匿名化の不可能性

上記の報告書をまとめるにあたっての規制改革会議からの要請は「データ源の個人が同意しなくてもパーソナルデータを転売も含めて自由に使えるための匿名化の基準作り」であった.

しかし, 報告書では, 比較的簡単でよく知られている k-匿名化を基本に置くとすると, 規制改革会議の要求に沿えるような完全な匿名化は不可能であるとした. 以下で少し詳しく説明する.

k-匿名化:個人 ID を消去ないし仮名化したうえで, 疑似 ID の情報の一部を消去あるいは精度を落とす. 例えば, 住所の記述から番地を削除する. その結果の疑似 ID が同じである人が k 人以上存在するようにデータベースを変更する k-匿名化が報告書では念頭におかれた. つまり, データベースを k-匿名化すれば, データ源の個人は疑似 ID から一意的な識別ができない. 当然, 匿名化されていない外部のデータベースなどの外部情報と突き併せても個人を特定できない.

しかし, 実質的には住所, 年齢, 性別など少数の情報に疑似 IDが固定されているわけではない. 例えば, データ業者 A のデータベースは疑似IDが k-匿名化されているが, 個人を特定はできなくても識別できる購買履歴も含まれていたとしよう. 一方別のデータ業者Bは購買履歴と, 行動履歴(通勤などの乗降駅)からなるデータベースを持っていたとする. すると, データ業者Aのデータベースをデータ業者Bが入手すれば, 購買履歴によって個人を一意に識別でき, その個人の行動履歴を知ることができる. したがって, 突き合わせに使う外部データベースを予見しきれない以上, 識別を防ぐにはデータ業者は疑似ID以外

連絡先:中川裕志, 東京大学, 文京区本郷 7-3-1, 03-5841-2729, nakagawa@dl.itc.u-tokyo.ac.jp

の全情報も合わせて k-匿名化しなければならない. しかし, そうするとデータベースの精度は悪化し, データの価値は激減する. よって k-匿名化は実質的に不可能ということになる.

個人情報の保護の目的のためには, データ受領者からさらに別のデータ受領者への提供においても匿名性を担保しなければならない. 匿名化できない場合はさらなる提供はできないことを法制化が必要である. 報告書では, これを以下に記す米国の FTC3 要件をベースに検討している.

ＦＴＣ3 要件

1. データ事業者はそのデータの非識別化を確保するために合理的な措置を講ずるべきである.
2. データ事業者は，そのデータを非識別化された形態で保有及び利用し，そのデータの再識別化を試みないことを，公に約束すべきである.
3. データ事業者が非識別化されたデータを他の事業者に提供する場合には，それがサービス提供事業者であろうとその他の第三者であろうと，その事業者がデータの再識別化を試みることを契約で禁止する.

※個人を識別可能なデータと，ここで説明した非識別化のための措置を講じたデータの双方を保有及び利用する場合には，これらのデータは別々に保管すべきである.

注意しなければならないのは，この要件においてはデータ受領者がデータ事業者となって，他のデータ受領者へのデータの移管を認めていることである．よって，上記の k-匿名化の説明で述べたように，データ受領者＝データ事業者が使う外部データベースを予見することがますます難しくなってくる．かくして，どのような危険性が存在するかを事前に把握しきれない．この状況においては，データ源の個人から同意をとることは難しくなってくると思われる.

その場合でもなお可能なのは，いわゆる統計データである．ただし，ある集合中の個人が別人として識別されたり，いわんや実世界でのリアルな個人として特定されることができないような統計データの明確な定義を与える必要がある．実際，統計法では外部データベースとの突き合わせも勘案して以下ように匿名データを規定している.

統計法第 2 条 12 項　この法律において「匿名データ」とは，一般の利用に供することを目的として調査票情報を特定の個人又は法人その他の団体の識別（他の情報との照合による識別を含む．）ができないように加工したものをいう.

この条文中の「識別ができないように加工」に関して「匿名データの作成・提供に係るガイドライン」において,

1）識別情報の削除，2）匿名データの再ソート（配列順の並べ替え），3）識別情報のトップ（ボトム）・コーディング，4） 識別情報のグルーピング（リコーディング），5）リサンプリング，6）スワッピング ，7）誤差の導入

などの処理が列挙されているが，匿名化の基準については，調査票情報の特性は統計調査ごとに異なることから，各統計調査について一律に匿名化の基準を設定することは困難である．このため，提供機関は，匿名化する統計調査ごとにその特性を勘案し，一橋大学における匿名標本データの試行的提供の事例及び諸外国の統計機関における同様の提供の事例等を参考に匿名化の基準となる値，例えば，最小値が２件以下とならない等を定める.

としており，ケースバイケースでの処理をデータ事業者に委ねている．よって，匿名化の基準については我々自身が説明責任を果たせるものを提示しなければならない.

4. ケースバイケースの匿名化の展望

報告書では 3 節で述べたように, 一般的なデータに対して完全な匿名化ができないとしたが, 同時に, 個別のデータベースと個別応用によっては匿名化ができる可能性があるので, 検討するとしている. ただし, 報告書では具体策, 具体例を提示していないので, 以下で検討する.

3 章の議論により, 個人情報すべてが疑似IDになりうることと, 突き合わせる外部データベースの予見不可能性が k-匿名化を妨げているので, この条件を回避できる個別ケースでは匿名化の可能性がある. したがって, 匿名化の要件は

a. 疑似ID(住所, 年齢, 性別などの典型的なもの)の有無
b. 外部可知／不可知： III. の「それ以外の情報」が収集されデータベースに格納されていることが外部の第三者に知られうるか／否か

となる.

ここで, 外部可知／不可知について説明する. 外部不可知の場合の例を示す.

例1:病院である検査をしたことは, 通常病院関係者以外には外部不可知.

例2:在宅ヘルスケアで, センサーから計測した心拍数などの健康情報を無線 LAN などで担当病院に送信する. これは, 同居家族ででもなければ外部不可知だし, データの値自体は本人ですら知らないこともありえる.

例3:カードでの購買履歴はカード会社以外には外部不可知.

まとめれば, 医療情報(病名など), 健康状態のセンサーデータ, 財産, 金融資産状況などは, データ収集者である病院, 金融機関など以外には知られないので, 第三者には外部不可知である. 一方, 滞在位置情報, 行動履歴, コンビニでの購買履歴などは物理的な動きを伴うので他人から観察できるので外部可知である. この観点からすると例えば, Suica の行動履歴は特定の人物に目をつけているストーカーなどから可知である. また, 公共の場所や店舗に設置された監視カメラの映像に写っているかどうかも第三者から可知である.

上記 a.と b.の組み合わせは表1. に示す各ケースとなる.

表1.　場合分け

III. それ以外の情報	疑似ID無	疑似ID有
外部不可知	不可知 & 疑ID無	不可知 & 疑ID有
外部可知	可知 & 疑ID無	可知 & 疑ID有

以下で表1の各ケースについて検討する.

- 外部不可知 & 疑ID無： データベースに格納されているか否かも知られず, かつ疑似 ID もないとなると, 仮にデータが公開されても本人特定は究めて困難である. k-匿名化はしていなくても特定はできない. ただし, 本人のデータ自体が万人周知で一意的である場合, 例えば10億円の宝石を購入したなどは外部可知である. この場合は, トップコーディングのような既存の手法で不可知化できる.
- 外部不可知 & 疑ID有： データベースへの格納の有無は知られていないので, 識別, 特定の手がかりは疑似 ID だけである. この場合は, 疑似 ID から識別, 特定されなければよいので, 同じ疑似 ID の人が k 人以上いるように疑似 ID の精度を落とす k-匿名化が有効である.

- 外部可知 & 疑ID無: データベースへの格納が知られており, データ収集事象を外部から観察できると, データが入手できれば, 疑似 ID の有無にかかわらず, データと観察日時などから本人特定が可能である. では, データ自体をk-匿名化すればよいのではないかというとそれも難しい. なぜなら, 長期にわたって収集されたデータが大きくなると, データ自体の個別性が高まり k-匿名化が困難になる. つまり, k-匿名化するにはデータの精度を大幅に落とさなければならないが, そうなるとデータの価値自体が大きく下がってしまう. また, 個人 ID を仮名化し, その仮名化を 1 日単位など頻繁に取り替えることは有力であるが, 同一の個人の行動履歴ではなくなるため, やはりデータの価値は下がってしまう.
- 外部可知 & 疑ID有: この場合は, 格納されているデータと疑似IDを連結したデータに対して k-匿名化を施すので, 前記の外部可知 & 疑ID無の場合よりもさらにデータの価値は下がってしまう.

以上をまとめる.

●データベースの個人データが格納されていることが外部可知の場合は, 外部からデータ収集していることを観察あるいは他のデータベースとの突き合わせによって, 識別ないし特定が可能なので, 疑似IDとデータ自体を併せて k-匿名化しなければいけないため, データの価値は大きく減少する. よって, 完全な匿名化手法はない.

●データベースの個人データが格納されていることが不可知の場合は, 疑似IDがなければ k-匿名化は不要, 疑似 ID があれば疑似 ID を対象にした k-匿名化が有効となる.

●収集した全パーソナルデータからサンプリングしてデータベースを作る方法もある. この場合, 個人毎にはサンプリングされたかどうか分からないので, データベースにある人情報が格納されているかは, 確率的に不可知／可知である. よって, プライバシーの安全性は確率モデルを作って評価する必要があり, 今後の課題である.

医療情報, ゲノム情報, 個人のセンサーから収集される健康情報, カードでのネット通販による購入, 個人の金融資産状況などは, 通常は外部観察できないので, 匿名化した上での活用ができそうである. 一方, 行動履歴などは, 匿名化が困難であり, 活用が難しい. これは, 一見, 重要度の低そうな行動履歴データがかえって活用できないという矛盾した結論にみえる. しかし, 外部不可知なデータというのは, 元来が個人の物理的あるいは法制度的にプライベートな場面で収集されるので, プライバシーという意味では最初から堅守されている. 一方, 行動履歴はプライバシー情報として堅守されていない. よって, 常識に沿った結論になっている.

5. センシティブ情報

III. のその他のデータのうち, 個人にとって他人に知られると不都合なデータをセンシティブ情報という. しかし,「不都合」とは何かを精密に定義していないので, センシティブ情報の定義は明白ではない. 加えて, 何がセンシティブ情報かは個人ごとに異なる. この節ではこの問題を扱う.

●コアなセンシティブ情報:

誰にとっても他人に知られたくない情報をコアなセンシティブ情報とする. ゲノム情報, 病気などの生体情報ないし健康情報, 財産, 債務, 学業成績, 親族などがあげられるが, 何を選ぶかは社会常識によるしかない. 逆に言えば, その定義には社会常識程度の安定性はある.

ところで, EU では滞在場所の情報はセンシティブ情報を超えて氏名と同じレベルの個人IDと見なす Data Protection Directive が昨年の欧州議会で可決されている. 日本では, 滞在場所, 移動履歴がどの個人IDなのかセンシティブ情報なのかの議論すら進んでいない状況である. ゲノム情報は個人IDに準じるとする考えがでてきている.

●状況依存センシティブ情報

上記の滞在場所や移動履歴がセンシティブ情報かどうかは個人ごとに異なる. 例えば, ストーカー行為を受けている人にとっては, 相手に知られたくない情報なので, センシティブ情報であろう. しかし, 他人につきまとわれることのない人であればセンシティブ情報ではない. 議論を簡単にするためには EU のように個人 ID としてしまうのもひとつの策である. ただし, 滞在場所や行動履歴はビジネスに役立つ情報なので, できれば活用したいものである.

購買履歴も個人ないし状況依存である. たとえば, 薬剤の購入は場合によってはセンシティブ情報になりうる.

宗教, 政治信条, 友人関係, 親類関係も状況依存性が高い. 友人関係は, 本人だけではなく, その友人に累が及ぶ可能性があるので, センシティブ情報になりやすい. たとえば, ある売り込み業者が自分の名前をかたって友人に売り込みをすると, 友人関係が悪くなる可能性がある.

このように状況依存のセンシティブ情報は一律な扱いが困難である. センシティブ情報であっても 1) 外部不可知であり, 2) トップコーディングなどで個人の特定ができない状態になっており, 3) さらに疑似IDも存在しない, ないし k-匿名化されているなら, データ事業者が第三者に再識別や特定をしないという条件で提供しても危険性はない. だが, それ以外の場合だと, 第三者提供するにはデータ収集時に本人同意が必要であろう. だが, データ収集時に, そのデータの利用方法をすべて列挙することは不可能である. 一方, データ源の個人にとっても, 収集されたデータが後になってセンシティブ情報になる, あるいはセンシティブ情報だと気づくかもしれない. こういった事態に対応については後に述べる.

6. k-匿名化が誘発する濡れ衣

まず, 以下の表2のデータベースの例について考えてみよう.

表 2 滞在場所のデータベース例

名前	年	性	住所	N 月 M 日 P 時の所
一郎	35	男	文京区本郷	K 消費者金融店舗
次郎	30	男	文京区湯島	T 大学
三子	33	男	文京区弥生	T 大学

最左列は人名だが, これは匿名化されなければならない. 2, 3, 4 列は, 疑似ＩＤで, 1～5 列が総合されると, 就活や婚活中に人にとっては, 最右列の所在地に消費者金融が記載されていることは芳しくない. そこで, 名前を A, B, C と仮名化し, 疑似ＩＤの情報を粗いものに変更して表 3 のように改変する. こうすると疑似 IDは 3 人とも同じになるので, 3-匿名化が実現でき, 消費者金融に行った人を特定できない. ところが, 疑似ＩＤでは 3 人を区別できないので消費者金融に行っていない残りの 2 人も消費者金融に行ったことを疑われる. これを k-匿名化が誘発する濡れ衣と呼ぶ[中川 2013]. 濡れ衣を防ぐには２つの方法がある.

表３　3-匿名化したデータベース

仮名	年齢	性別	住所	N 月 M 日 P 時の所在
A	30 代	男	文京区	K 消費者金融店舗
B	30 代	男	文京区	T 大学
C	30 代	男	文京区	T 大学

第１の方法は，k-匿名化の k を大きくすることである．例えば，表 2，3 のような例で，k=30 であるなら，消費者金融店舗に出入りしたのが１名であると，わざわざ他の 29 名を疑う労力は骨折りであるという心理が働くであろう．[中川 2013]では，このことをコストの観点から分析している．ただし，k を大きくすると，データの精度が下がり価値が低下する．

第２の方法は，k-匿名化をしないことである．疑似ＩＤを変化させて，一致する人を増やす操作をしないので，通常は消費者金融店舗に出入りした人は１人と識別される．よって，濡れ衣を疑われる人はいない．ただし，他のデータベースと突き合わせると本人の特定がしやすくなる．

濡れ衣は無実の罪という側面が強いので，7節で述べる自己情報コントロール権の問題として扱うのが適当であろう．

7. 自己情報コントロール権

以上の考察から，k-匿名化のような識別の曖昧さ導入を狙った匿名化では，実社会での応用場面をカバーしきれないことが分かった．

表１における外部不可知 & 疑ＩＤ有，外部不可知 & 疑ＩＤ無の場合には，ひとたび流出したら取り返しがつかないゲノム，健康情報が当てはまる．データ処理を病院などの収集した組織内に限定し，他のデータ事業者への提供は禁止すべきである．医療情報においては，これは現状と同じレベルの情報管理と考えられる．他の医療機関や研究所との協力に際しては，データ受領者である組織からさらに他の組織に提供することは禁止すべきである．

表１の場合分けのうち，k-匿名化が有効に作用する可能性は，外部不可知 & 疑ＩＤ有　の場合だけであり，パーソナルデータの利活用において k-匿名化だけに頼ることは難しい．行動履歴や購買履歴などビジネス的価値が高いデータは，外部可知 & 疑ＩＤ有であり，k-匿名化には本質的に馴染まない．

この状況での現実的方策を考えてみよう．

- 個人 ID を消去ないし仮名化すること．さらに仮名の変更を頻繁に行うこと．この基礎的方策により，簡単には識別や特定ができなくなるので，必須である．
- 疑似ＩＤはデータベース内に含ませないことをデフォールトとする．疑似ＩＤも必要な場合は，それだけをデータベースから分離して別のデータベースとして，仮名化されている個人ＩＤとの対応テーブルは暗号化などでさらに管理を厳重化する．疑似ＩＤが存在しなければ，個人の特定は難度が高い．
- 自己情報コントロール：上記の方策でも，III.その他の情報が集積すると完全な匿名化は難しい．その場合にはデータ源である個人が自己の情報の利用され方を開示要求して閲覧できること，および消去要求でき実際の消去を確認できることが重要となる．

自己情報の開示と消去の権利は2013年12月に欧州議会で可決されたEUのData Protection DirectiveのProposal for a directive Recital 16の改正案に記載されており，日本の法制度をEUレベルにするなら必要な改革となる．EUとの対比を離れても，この自己情報コントロール権は以下の観点からも重要である．

すなわち，規制改革会議に要請を見ると，個人データを利用するにあたっては，データ源の個人が同意さえすれば自由に使えると思われているようだが，これは必ずしも正しくない．つまり，同意の内容で無制限な個人データ利用を可能だと書くと，多くの人々から同意が得られなくなる恐れがある．[Schornberger2013]の 9 章には，ビッグデータの利用法は収集の前には予め列挙できないので，利用法を指定しての同意取得は実効性がないと述べられている．同様にデータ提供先を同意時点で列挙しきれないであろう．

そこで，データ源の個人に安心して同意してもらうためには，個人 ID の消去や仮名化に加え，上記の自己情報コントロール権（開示と消去）の実施が（データ公開前に）確実に行えることを保証することが有効であろう．私見ではあるが，k-匿名化は，k が 5 以下の小さな数だと，データ源の個人に安心してもらえるかどうか疑問がある．それよりは，自己情報コントロールのほうが直観的に理解を得やすいと思われる．

濡れ衣の問題も，自己情報コントロール権があれば，かけられた疑いを晴らすことが原理的には可能になり，大きく軽減されるであろう．

以上の考察により，プライバシーバイデザイン[Cavoukian 2010]の考え方を採用するなら，自己情報コントロール権の実効的な実装の組み込みをシステム設計時からしておくべきであろう．なぜなら，複雑に改変されたデータベースで個人データの追跡を行うことは，不可能に近いからである．

8. おわりに

個人情報の技術的，法制度的扱いについては，欧米，さらに日本においても急速に変化してきている．ここの述べた提言がひとつの方向性を示していると期待したい．

参考文献

[佐藤 2013]　佐藤一郎，他: 技術検討 WG 報告書，パーソナルデータに関する検討会，2013．http://www.kantei.go.jp /jp/singi/it2/pd/dai5/siryou2-1.pdf

[Schörnberger 2013]　V. Mayer-Schörnberger, K,Cukier: BIG DATA A Revolution That Will Transform How We Live, Work, and Think,邦訳：ビッグデータの正体．Houghton Miffin Harcourt Publishing Company．2013

[中川 2013]　中川裕志，角野為耶: 滞在場所の k-匿名化と濡れ衣. 情報処理学会.第 62 電子化知的財産・社会基盤研究発表会(EIP 研究会) Vol.2013-EIP-62, No.12.2013

[板倉 2014] 板倉陽一郎，他:「完全な匿名化」幻想を超えて．SCIS 2014 3D1-4 IEICE. 2014.

[Fun 2008] B.C.M.Fun, K.Wang, R.Che, P.S.Yu: Privacy-Preserving Data Publishing: A Survey of Recent Development, ACM Computing Surveys, Vol. 42, No. 4, Article 14, 2010.

[Cavoukian 2010]A. Cavoukian： 7 Foundational Principles of Privacy by Design. Information and Privacy, Commissioner/Ontario.http://www.privacybydesign.ca/index.php/about-pbd/7-foundational-principles/