社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
SITE2011-29(2011-12)

情報処理技術向上に伴う 個人に由来するデータの匿名性の変化に関する一考察

森下壮一郎[†,††]　　横井　浩史[††,†]

† 東京大学大学院 情報学環
〒 113-0033 東京都文京区本郷 7-3-1
†† 電気通信大学大学院 情報理工学研究科
〒 182-8585 東京都調布市調布ヶ丘 1-5-1
E-mail: †smori@hi.mce.uec.ac.jp

あらまし　情報処理技術の向上に伴って，従来は個人情報とは見なされなかったデータが個人情報としての性質を帯びるようになる．具体的には，DNA の塩基配列や脳波などの生体由来のデータや，ID タグのトラッキングなどによる行動履歴データである．政府の指針によれば，個人を識別できる情報を取り除く処理，すなわち匿名化が施されたデータは個人情報とは見なされない．一方でデータマイニングによるバイオインフォマティクスやライフログ解析などの研究が進められており，それぞれ成果が上がっている．これらは，あまりに膨大で複雑であるために従来は解釈が困難であったデータから意味を見出すものである．その発展により，ゆくゆくは匿名化が施されたデータでもいずれ個人同定が可能になりかねない．以上の背景を元に本稿では，現状の個人情報保護施策および匿名化の手法について概観し，考え得る問題点とその対策について考察する．
キーワード　個人情報，匿名化，データマイニング，情報倫理

A study for the change of anonymous nature of data from individuals with development of information processing techniques

Soichiro MORISHITA[†,††] and Hiroshi YOKOI[††,†]

† Interfaculty Initiative in Information Studies, The University of Tokyo
Hongo 7-3-1 Bunkyo-ku, Tokyo, 113-0033, Japan
†† Graduate School of Informatics and Engineering, The University of Electro-Communications
Chofugaoka 1-5-1, Chofu-shi, Tokyo, 182-8585, Japan
E-mail: †smori@hi.mce.uec.ac.jp

Abstract　With the improvement of information processing technology, the data which were not considered to be personal data come to have a property as information on individuals. Specifically, they are the data derived from the living bodies such as base sequence of DNA or brain waves, tracking data of mobile terminals, and so on. According to the guideline of Japan's government, the anonymized data — the data removed the information that can distinguish an individual — are not seen as personal information. On the other hand, studies such as bioinformatics or the life log analysis by using data mining are being promoted, and they improve success. Because I am too enormous, and these are complicated, interpretation finds a meaning from difficult data conventionally. They find meanings from too enormous and complicated data to interpret. the development may allow personal identification from anonymized data sometime soon. For this background, in this paper, we survey current personal information protection measures and anonymity techniques. Then we discuss conceivable problems and the measures for them.
Key words　personal data, anonymization, data mining, information ethics

1. はじめに

計算機で処理可能なデータ量の拡大および処理速度の劇的な向上は，研究開発や産業の各分野に情報処理の質的な変化をもたらした．人力では到底処理不可能なほど膨大かつ複雑なデータを広大なメモリ空間に構築したデータベースに格納して，高速な情報処理により従来は解釈が困難であったデータから意味を見出そうとするところのいわゆるデータマイニング技術は，例えばバイオインフォマティクスやライフログ解析などの分野で成果を挙げている．

ところでこれらの研究においては，個人から採取した遺伝子情報や ID タグのトラッキングなどで得られる行動履歴情報が取り扱われる．プライバシー保護の観点から，これらの情報には個人を識別可能な情報を取り除くことによる匿名化が施される．これによりデータベースの情報が事故で流出したときに，あるいは正規の利用であっても必要以上には，個人のプライバシーが脅かされることがないよう配慮されている．

しかしながら，個票データ (microdata) については，匿名化が施されたものであっても，複数のデータベースへの問い合わせおよび背景知識により匿名性が破られることが既に指摘されている [3]．また，前述のようにデータマイニング技術は解釈が困難なデータから意味を見出すための技術である．現在は遺伝子情報から個人情報と連結されたデータを介することなく個人を同定することはほとんど不可能であるが，今後の技術発展によりデータからより多くの情報を抽出できるようになると，それが匿名性を破るための背景知識を補強するものになってしまう可能性がある．

以上に基づき，本稿ではまず日本における現状の個人情報保護施策および匿名化の手法について概観する．具体的には，個人情報保護法における個人情報の定義と，不正アクセス禁止法における識別符号の定義，および JIS Q 15001 における機微情報の定義を参照して，保護すべき情報について整理する．さらに運用の具体例として，文部科学省，厚生労働省，経済産業省による『ヒトゲノム・遺伝子解析研究に関する倫理指針』で提示されている匿名化手法と，経済産業省による「情報大航海プロジェクト」において採用された匿名化手法について言及する．さらに匿名化の突破が可能になる理論と，考え得る対応策とその副作用について述べる．

2. 保護すべき個人情報

本節では，本稿で対象とする個人情報の定義について述べる．

2.1 個人情報と機微情報

『個人情報の保護に関する法律（個人情報保護法）』第２条において,「個人情報」は次のように定義されている．

> 生存する個人に関する情報であって、当該情報に含まれる氏名、生年月日その他の記述等により特定の個人を識別することができるもの（他の情報と容易に照合することができ、それにより特定の個人を識別することができることとなるものを含む。）

このうちの「特定の個人を識別することができるもの」とは，一般には住民基本台帳の 4 情報（氏名，生年月日，住所，性別）を指すと解釈されるようである．宇治市住民基本台帳漏洩事件（最高裁 H14.7.11）では，これらの情報が漏洩したことについて 1 人あたり 10,000 円の慰謝料が認められた．

一方，プライバシーマークの認証基準であるところの JIS Q 15001 において機微情報（センシティブ情報）として次のような種類の情報が列挙されている．

> １）思想・信条・宗教に関する情報
> ２）人種・民族・出生地・本籍地。身体障害・精神障害・犯罪歴・社会的差別の原因となる情報
> ３）労働運動への参加状況
> ４）政治活動への参加状況
> ５）保健医療や性生活

これは，いわゆるプライバシーに何が含まれるかを規定していると見なして良い．日本においてのプライバシーの概念は未だ固定されたものはないが，個人情報保護法における個人情報よりはこれらの方が明らかにされたときの精神的苦痛は大きいことは間違いないであろう．しかしながら，これらの機微情報も結局は個人と紐づけられない限りは本人に精神的苦痛をもたらさない．例えば「特定の思想を持つある人物がどこかに存在する」という情報のみでは個人のプライバシーの侵害にはならない．機微情報についても，それらが「個人情報」と紐づけられて初めてプライバシーの問題に発展するのである．なお,「2005 年度情報セキュリティインシデントに関する調査報告書」においては，経済的損失レベルと精神的苦痛レベルの 2 つの軸で情報漏洩の重大さをマッピングした Simple-EP 図が提示されている．さらにこの評価に本人特定容易度を乗じて損害賠償額を計算する式が示されている．すなわち，本人特定が不可能な状況であれば情報漏洩の重大さは減ずるのである．以上のことから，個人を識別して一意に同定することが可能な情報が個人情報であり，また日本の社会における基本 4 情報は，ほとんど本人を指し示すと言って良いほど本人特定が容易な情報と見なされていると考えて良いだろう．

2.2 識別符号

なお,『不正アクセス行為の禁止等に関する法律』において「識別符号」が次のように定義されている．

> （定義）第二条
> 当該アクセス管理者において当該利用権者等を他の利用権者等と区別して識別することができるように付される符号であって、次のいずれかに該当するもの又は次のいずれかに該当する符号とその他の符号を組み合わせたものをいう。
> 一　当該アクセス管理者によってその内容をみだりに第三者に知らせてはならないものとされている符号
> 二　当該利用権者等の身体の全部若しくは一部の影像又は音声を用いて当該アクセス管理者が定める方法により作成される符号

三　当該利用権者等の署名を用いて当該アクセス管理者が定める方法により作成される符号

このうち，一はパスワード，二は生体認証情報（バイオメトリクス情報），三は暗号鍵のことを指している．情報処理システムにおける個人認証は，これらと個人に割り当てられた ID とを照合することで行われる．政府機関統一基準ではこの ID を識別コードと呼ぶようである．

識別コードも，個人を識別して一意に同定するものである．しかし本人が特定可能かどうかは（基本 4 情報の意味での）個人情報と，識別コードとが対応づけられているか否かに依存する．その意味では識別コードも識別符号も個人情報ではない．しかしながら，識別符号を知ることでなりすましが可能になり，また個人情報や機微情報のコントロールも可能になるから，識別符号は個人情報よりも上位の機密情報である．

3. 匿名化と実現できる匿名性

結局のところ，機微情報が個人情報に紐づけられる状況を避けることが重要である．そのために行う処理を一般に匿名化と呼ぶ．本節では，現状行われている匿名化手法と，それにより実現できる匿名性の程度について述べる．

3.1 連結可能匿名化と連結不可能匿名化

ここで文部科学省，厚生労働省，経済産業省による『ヒトゲノム・遺伝子解析研究に関する倫理指針』を参照し，具体的に行われている匿名化について述べる．

同指針において「保護すべき個人情報」の定義は，個人情報保護法における「個人情報」と同様のものである．なお，この指針で対象にしているバイオインフォマティクス研究においては，実験協力者の DNA データや病歴等が取り扱われることが多い．これらは機微情報である．同指針では，これらの機微情報について連結可能匿名化もしくは連結不可能匿名化を施すよう示されている．連結可能匿名化は，個人情報と機微情報とで別々のテーブルを構築するすることで匿名化するものである．ただし個人情報と機微情報とに共通の識別コードを割り振ることで，必要に応じて個人情報を機微情報とを紐づけられるようにしている．提供者本人の求めに応じて，機微情報そのものを削除するため等の理由で行われる．一方，連結可能匿名化は，機微情報に対応する個人情報を抹消してしまうことによる匿名化である．同指針では，

個人情報を連結不可能匿名化した情報は、個人情報に該当しない。

と明言されている．もちろん，同指針は機微情報の扱いを無制限に緩和するものではなく，

遺伝情報、診療情報等個人の特徴や体質を示す情報は、本指針に基づき適切に取り扱われなければならない。

と明示しており，第一義としては情報提供者の権利を守るための指針である．一方で，バイオインフォマティクス研究を進めるにあたって過度に神経質な機微情報の取り扱いが発展を妨げる恐れがあり，それを防ぐことも同時に目的としている．

この匿名化が功を奏した事件として，理化学研究所遺伝子多型研究センターからの情報流出事件が上げられる．2006 年 9 月 11 日，同センターで取り扱っていた遺伝子情報がファイル共有ソフト Winny のネットワークに流出したことが発覚した [12]．理化学研究所のプレスリリースによると，流出した情報の中には「患者１４４名分の疾患関連ＳＮＰ」が含まれていたという．さらに,「個人を特定する情報と遺伝子情報がつながることがないよう匿名化し、直接個人を特定できないような仕組みを構築しており、個人名等が直接明記されている訳ではありません」と説明されている．この匿名化が本当に十分であったかは議論の余地があるが，これについては後述する．

もっとも，真に機微情報の漏洩を怖れるのであれば，そもそも機微情報を取り扱わないことが唯一絶対の方策である．匿名化は，機微情報の保護と機微情報の利用の両立を目指して行われる．

3.2 k-匿 名 性

以上で述べた連結不可能匿名化によって，個人情報と機微情報との直接の対応関係は失われる．しかしながら，前述のように，個票データ (microdata) については，連結不可能匿名化が施されたものであっても，複数のデータベースへの問い合わせおよび背景知識により匿名性が破られる可能性がある．Sweeney は，連結不可能匿名化が施された医療記録において，マサチューセッツ州知事に対応する情報を特定できることを指摘した [3]．また石原らは，クエリ結果から推論によりデータベースの全体像を構築して機微情報を得ようとする攻撃を推論攻撃と呼んでいる [4]．2 つ以上のデータベースに共通して含まれる情報を媒介にして，個人情報と機微情報とを対応づけられる場合があるのである．このような情報は準識別子 (Quasi-identifier) と呼ばれる．一方で，前述の連結可能匿名化が施されたデータにおける個人情報と機微情報とを媒介にする識別コードは，データベースの構造から一意であり，このような識別子を正識別子と呼ぶ．

準識別子によって絞り込まれた候補が k 個あるとき，素朴に考えればそのうちの一つが真に対応する情報である確率は $1/k$ である．準識別子をどのように取ったとしても少なくとも k 個の候補が現れることが保証されるとき，そのデータベースは k-匿名性を持つという．さらに，その条件を満たすような処理を施すことを k-匿名化と呼ぶ．

k-匿名化は，経済産業省による「情報大航海プロジェクト」において採用された．このプロジェクトは，蓄積される一方である膨大なデータの中から，ユーザが求める情報を的確に抽出する技術の確率を目的にしたものである．このプロジェクトで取り扱うデータには機微情報も含まれることから，匿名が技術について検討が行われ，k-匿名化が採用されたようである．この他，ライフログ分析のための移動軌跡データなどを k-匿名化する研究が種々行われている [6], [11]．位置情報は特に住所等の個人情報と容易に結びつくので，匿名化が極めて重要なのである．

3.3 候補の偏りを考慮に入れた匿名性

上記のように，k-匿名化によって個人情報と機微情報との対応が一意に特定される危険を減ずることができる．しかしながら，その確率が必ず $1/k$ となることを保証するものではない．その理由の一つは，k 個の候補のうち対応する機微情報の候補が k よりも少ない種類しかない場合があるからである．このような場合の匿名性については，機微情報の種類の数を ℓ として ℓ-多様性という [1]．また，ℓ-多様性があるとしても，候補の出現確率が偏っているときにはやはり期待よりも高い確率で正しく対応づけられてしまう．これについては，分布間の距離に応じて t-closeness という [2]．

3.4 背景知識も考慮に入れた匿名性

以上で述べた匿名性は定量的な指標であるが，攻撃者が背景知識を持つとすると，さらに匿名性の程度は低くなり，また定量的に評価することが困難になる [7]．なぜならば，攻撃者は背景知識に基づいて対応する確率が低い候補を棄却できることで匿名性の程度が低くなるのであるが，データベースに含まれない背景知識についてはその程度を事前に評価することはできないからである．これはデータ解釈の合理性に基づく戦略である．

なお，実は数理的には基本 4 情報と言えども準識別子に過ぎない．そして合理性に基づく戦略は，個人同定について第 1 種の誤りを引き起こしやすい．以下に実例を挙げる．2011 年 10 月 17 日，埼玉県久喜市が同姓同名の別人の財産を差し押さえてしまっていたことを発表した [5]．生年月日も同じであるために混同してしまったのである．これは生命保険の解約返戻金を差し押さえたものであるが，保険会社から住所が違う旨の指摘があったにも関わらず，職員は解約手続きに踏み切ったという．この事件の背後にあるのも合理性である．氏名と生年月日に比べて，住所は変わる可能性が高いので，同姓同名の別人であると考えるよりは転居した同一人物であると見なした方が合理的であるという思考が働くのである．なお 2007 年にも，横浜市旭区が全く同様の状況で同姓同名で生年月日が同じ別人の財産の差し押さえを行っている [10]．この件は，データベース攻撃によるものではないが，個人同定の第 1 種の誤りが攻撃者にとっての不利益よりむしろ誤って同定された個人が被ったという点で典型的な事例である．なお，これらの事件は氏名や生年月日も結局は準識別子に過ぎないことも裏付けている．とはいえ，実際問題としては氏名と生年月日の組み合わせが正識別子として扱えない確率は極めて小さい．遺伝子情報を元にした識別子も厳密には準識別子であるが，やはりほとんど正識別子として扱って構わないものである．

3.5 技術の発達により匿名性が変化する場合

さらに情報処理技術の発達によって，個人情報に対応づけられていない機微情報が個人情報（認証情報）になってしまうことが考えられる．例えば，DNA 認証 [8] は既に実現されているので，将来的にこれが実用化されれば，DNA 情報は識別符号になる．実現していないので意味はない仮定ではあるが，この場合，前述の理化学研究所からの流出事件の重大性は極めて大きくなってしまう．また脳波認証も実現可能性が示唆されている [9]．いずれにせよ，いまは機微情報に過ぎず，個人情報と対応づけられない限り問題にならない生体情報が識別符号そのものとなってしまうのである．

4. おわりに

本稿では，データベースの匿名化手法と匿名性評価指標について概説した．そして，背景知識を利用した攻撃がデータ解釈の合理性を元にした戦略であり，さらにこのような攻撃は第一種の誤りを誘発して関係のない第三者が被害を被ることになることを述べた．最後に，情報処理技術の発達によって今まで個人情報とは見なされないデータが個人情報になり得る危険について言及した．

謝　辞

本研究の一部は，文部科学省脳科学研究戦略推進プログラムにより実施された「ブレイン・マシン・インターフェースの開発」の成果である．

文　献

[1] A. MACHANAVAJJHALA, D. KIFER, J. GEHRKE, , and U. VENKITASUBRAMANIAM. ℓ-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data*, Vol. 1, No. 1, 2007.

[2] Li Ninghui, Li Tiancheng, and S. Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and ℓ-diversity. *ICDE2007*, pp. 106–115, 2007.

[3] Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty Fuzziness and Knowledge Based Systems*, Vol. 10, No. 5, pp. 557–570, 2002.

[4] 高須賀史和, 橋本健二, 石原靖哲, 藤原融. XML データベースへの型推論を用いた攻撃に対する安全性検証. 電子情報通信学会技術研究報告. DE, データ工学, Vol. 106, No. 148, pp. 151–156, 2006.

[5] 産経新聞. 埼玉・久喜市が人違いで生命保険を差し押さえ　同姓同名、生年月日も同じ (2011.10.17 14:39). `http://sankei.jp.msn.com/politics/news/111017/lcl111017144100001-n1.htm`, 参照 Oct. 18 2011.

[6] 村本俊祐, 上土井陽子, 若林真一. データを極小歪曲し k-匿名性を保持したデータに変換するプライバシー保護アルゴリズム. 日本データベース学会 Letters, Vol. 6, No. 1, pp. 97–100, 2007.

[7] 村本俊祐, 上土井陽子, 若林真一. 背景知識を用いた推測を困難にしデータ歪曲度を極小化するプライバシー保護手法. In *DEWS2008*, pp. C1–4, 2008.

[8] 板倉征男, 長嶋登志夫, 辻井重男. DNA バイオメトリックス本人認証方式の提案. 情報処理学会論文誌, Vol. 43, No. 8, pp. 2394–2404, 2002.

[9] 宮本千正, 原秀樹, 中西功, 伊藤良生, 副井裕. 脳波による個人認証の研究. 2007 電子情報通信学会大会講演論文集, pp. S.144–S.145, 2007.

[10] 読売新聞. 同姓同名でうっかり、横浜市旭区が別人の財産差し押さえ (2007.9.28 22:51). `http://www.yomiuri.co.jp/national/news/20070928i115.htm`, 参照 Sep. 29 2007.

[11] 冨山克裕, 川島英之, 北川博之. 移動軌跡ストリームに対する連続的 k-匿名化技法. 全国大会講演論文集, Vol. 2011, No. 1, pp. 93–95, 2011.

[12] 理化学研究所. プレスリリース　ＮＴＴデータによる情報の流出について　流出情報、理化学研究所遺伝子多型研究センターとの共同研究情報を含む (平成 18 年 9 月 13 日). `http://www.riken.go.jp/r-world/info/release/press/2006/060913/index.html`, 参照 Sep. 13 2006.