パーソナルデータに関する検討会　第１回技術検討ワーキンググループ　議事要旨

１　日　時　　平成２５年９月２７日（金）１０：００～１２：００

２　場　所　　中央合同庁舎第４号館　全省庁共用１２３会議室

３　議　事

（１）開会
（２）主査あいさつ
（３）技術検討ワーキンググループについて
（４）これまでの検討や取組事例等について
（５）ワーキンググループでの検討項目（技術面、制度面）について
（６）次回に向けた作業事項について
（７）閉会

４　＜配付資料＞

【資料１－１】パーソナルデータに関する検討会の体制について
【資料１－２】パーソナルデータに関する検討会の運営について（ワーキンググループの設置について規定）
【資料１－３】技術検討ワーキンググループ構成員名簿
【資料１－４】技術検討ワーキンググループの進め方
【資料２－１】匿名化に関する検討資料（総務省提出資料）
【資料２－２】※非公開資料（席上配布のみ）
【資料２－３】匿名化技術の現状について（高橋構成員提出資料）
【資料３　　】匿名化に関する検討項目（技術、制度）案（佐藤主査提出資料）

【参考資料１　　】パーソナルデータに関する検討会の検討予定
【参考資料２　　】パーソナルデータの取扱いルール整備に向けて検討すべき論点
【参考資料３－１】※非公開資料（席上配布のみ）
【参考資料３－２】※非公開資料（席上配布のみ）
【参考資料３－３】※非公開資料（席上配布のみ）
【参考資料３－４】※非公開資料（席上配布のみ）

５　出席者

佐藤（一）主査、森主査代理、伊藤構成員、岡村構成員、菊池構成員、佐久間構成員
佐藤（慶）構成員、高橋構成員、松本構成員

総務省　総合通信基盤局　電気通信事業部　消費者行政課
経済産業省 商務情報政策局 情報経済課
内閣官房情報通信技術（ＩＴ）担当室　二宮参事官、濱島参事官、瓜生参事官、村上調査官、宮田補佐、楠政府ＣＩＯ補佐官、満塩政府ＣＩＯ補佐官

６　概要

○このワーキンググループは親会（パーソナルデータに関する検討会）の下に置かれており、親会は法律家の方を中心に構成され､技術面はこのワーキンググループで見る形になっている。ともかく時間が限られているため、深く議論をさせていただきたいが、ある程度、結論を出すことを前提にして議論していかないとならいことがあることを念頭に置いてほしい。そのため各構成員には、宿題をお願いし、第２回目、第３回目で御報告とご議論いただき、第４回目で報告案を纏めたいと思う。

親会の要求によっては別の技術項目も議論する必要が出てくるかもしれないが、当面、匿名化を議論対象とする。趣旨としては、パーソナルデータに関していろいろ規制のかかる部分、データに関して匿名化をすることによりその枠の外に出すことで利活用を進める部分、国民の安全をちゃんと担保した形で活用することが重要になるかと思っている。

事務局から資料１－１～資料１－４の説明、総務省から資料２－１の説明、経済産業省から資料２－２の説明、高橋構成員から資料２－３の説明、事務局から資料３の説明あり。出席者から以下の発言があった。

○検索エンジンのアルゴリズムは、完全に人間の知能と同じように、140言語を文章として読み取り、70億ページを0.1秒で検索するものであり、既に十数年動いている。対象文章がアルバムや論文や、それに対する文章に関する記述も入る。また、ハッキングのアルゴリズムがありある程度の推論で中のデータを抽出する技術もある。

２つ提案したい。１点目は対象が文章もしくは表になっていない情報について、その匿名化をどう制御するか。２点目は、アクセスに対するある程度の記述みたいなものが必要なのではないか。

○個人情報保護法の適用がない形で利活用するには２つ大きな問題がある。

１点目は、法律に匿名化されていない情報には個人情報保護法は適用しないと書いているわけではない。法律側の問題として、識別性の有無について明確な基準がまだできおらず、その境界線が法律上明確ではない。従って、技術にて匿名化が可能になったとしても直ちに識別性

を失わせたということは非常に難しい。
２点目は、仮に識別性がない形にできた場合、これで個人情報保護法の適用はなくなったため、自由に第三者提供、公表については、ほかのルールが残っているおり、検討が必要である。

○今の検討項目に２点欠けている要素がある。
１点目は、法律屋はどうしてもトゥルー・フォルスの２択に持っていかれたがるが、２次に分けられるものはなく、匿名化にもいろいろなレベルがあり、そのレベルについての議論をする必要がある。例えば集合匿名化一つをとっても、ｋというパラメーターの値を２にするのか、100にするのか、とういうパラメーターの問題とがどこまでも残っており、そこの観点が今は抜けていると思う。
２点目は、いつまでにできないのかという時間に関する議論が欠けていると思う。例えば、共通鍵暗号は寿命があるように、時間さえかければ現在の暗号技術は、必ず解けるようになっている。匿名化も同様に、匿名化されたデータが蓄積されるに従って、やがて再識別されるリスクが上がり、いつまで担保するかという議論が必ず必要と思う。

○資料３の２ページ目について、経済産業省のワーキンググループでもQ1「できない情報か？」を追求したが、結論としては汎用的な処理基準を定めることは無理だという回答をつけてもいいと思う。今回、「できない情報」から「してはいけない情報」に観点を変えた形も枠に入れる必要があると思う。もちろん、個人情報の利活用という観点では、ｋ－匿名化を使える状況では、すぐ使ってもいいが、ｋ－匿名化は、匿名情報と呼ばれていたものが万が一流出したときの危険度をはかる尺度としてむしろ有用だと思う。しかし、特定の用途で継続的に利活用する場合には、匿名化前のデータ母集団の状況によって匿名化後の粒度が変化してしまうデータは実務上は使いにくいことになるものと思っている。
米国は、FTCの勧告で再識別化を禁止するとしている。できるか、できないかではなく、してはいけないというルールに定めることが法律で可能ならば、そちらに駒を進めることがあってもいいと思う。Q2も容易に照合できないではなく、照合してはいけない、という考え方で整理すればいいのではと思う。技術論というよりは管理論のところも含めた形で、この問題を解決したほうがいいと思う。

○このワーキンググループは匿名化が焦点だが、線引きが難しいと思う。
例えばカーナビのデータでは、幾ら自動車の運転者や自動車を匿名化しても、車の出発点と目的地で個人が特定できる。ヨーロッパでは自動車の乗り始めと乗り終わりのデータを捨てており、従来の匿名化かどうかわからない範囲も含めて議論していかなければいけない。

○資料３の匿名情報の第三者提供の類型の例でわかりやすいのは、鉄道会社の話だと思う。鉄道会社の例は、委託ではなく、資料２－３で言うと、高度な匿名データではなく、いわゆる匿名データを第三者提供した例だと思う。

個人情報ではなく、匿名情報として扱いたいという理由については、同意をとりたくないという話だと思う。しかし、これには、匿名性というだけの軸だけの話だけではなく、機微性等の他の軸も考慮する必要がある。

いわゆる個人情報に近い匿名データを扱うとすると、統計的な匿名性のレベルの話だけではなく、組織の信頼性や、その他技術的安全性等も一緒に検討していく必要がある。

○高度な匿名化に関しては今のデータホルダーには手に負えないと思っており、そこの運用に関しては考慮する必要がある。

どんなにいい制度や指針を出しても、結局、現実世界でインプリメントできなければ意味がない。

○一つのポイントは、匿名情報という呼び方はやめたほうがいいと思っている。匿名性は程度の問題であり、ある水準になったら、それが何か匿名情報というよりは無名情報になるのかというと、技術においては不可能である。具体的な加工データをもって、それをｋ－匿名性を使って計算して匿名性の程度を求めることはできるが、それをここまでになったら、これは再識別化が不可能になったとすることは不可能と割り切ってもいいと思う。

この情報は匿名化しましたという言葉を使われてしまうと無名化した雰囲気を醸し出すため、この言葉の使い方は今後整理していく必要があると思う。

もう一つ、利活用の観点においては、NTTドコモのモバイル空間統計もオプトアウトをとっていると思うが、これを事業者がやらなくて済むぐらいのところまで明確にしないといけないと思う。

このNTTドコモのモバイル空間統計は、もはや匿名ではなく統計処理であり、統計処理の元データに対してオプトアウトをとったら統計結果の精度が悪くなる。さらに、一旦とったオプトアウトはめったなことでは解除できないため、データとしては別の統計処理にも使えなくなる。

個人情報の第三者提供を利用目的にしたオプトアウトを日本の事業者が安易な方法としてやり始めたら、日本のいわゆるオープンデータというものの品質がどんどん下がってしまう。いったんオプトアウトされたものが同意に転じることが少ないと考えれば、事業者が安易にオプトアウトをとることが防げるような形で物事を持っていくところまで今回のワーキンググループのところで導ければと思う。

○モバイル空間統計は、個人情報保護法の識別性ということを意識してここが基準ですという境界線はわからないが、ここまでやればセーフというところに持ってくることを目的にすると、その目的を達しているのではないかと思う。

○結論ではないのですが、統計は事実上フリーにしてもいいのではないか。ただ、統計の定義をもう一回考える必要はあるかもしれない。逆にオプトアウトをとる必要もないとは法律には書けないかもしれないが、事実上はオプトアウトをとらずに使うことで統計データをちゃんと

活用することができるようにすべきと思う。

〇統計と匿名データというものの線引きが難しい。例えば、港区に100人いるというのが統計データであるが、港区何番地何々町に何時に３人いるというのも統計データである。

ビッグデータになってきて、データの解像度がどんどん細かくなってきているから匿名性の問題が出てきているので、統計データと匿名データを区別することが難しいということを留意する必要がある。

Q1について例えばこの自動車は安全かと言われたら、100％安全はないため、安全ではないと工学者としては言わなければならない。こういう聞き方をされるとやはり不可能となってしまう。ため、例えばｋ－匿名も、識別できなくする技術ではなく、識別できる確率がｋ分の１以下であることを保証する技術であり、識別できないことを保証するわけではない。

そのため、用語をきちんと定義して、同じことを、同じ概念について皆が議論できるような環境をつくりたいと思う。

もう一つ、モバイル空間統計の報道が出たとき、まとめサイトとかをよく見ていたのが、みんな盛大に勘違いをしていた。要するにデータに電話番号がくっつき、どこにいるかという情報がどんどん売られていくのだと勘違いをしていた。

うまくデータ提供者のマインドをコントロールできれば、そんなに激しく反応することはないと思う。技術的には問題ないはずなのに、マインドのコントロールがうまくいっていないから正論が変な方向に行ってしまうということがあると思う。データ提供者への説明をどうやって適切にやるかとか、という議論は結構重要なのではないかと思う。

〇静的データでインターネット、もしくはコンピュータに入っている表やグラフのイメージを検索するのではなく、対象物がテレビのニュースから音声から映像といったものをランダムに探していく非常に高度な検索エンジンを使うことで、全ての情報は全て検索がでる。

もう一つは、それが匿名性へ変化させたときにもログがきちんと取られているため、ほとんどのケースは可逆性を持ち、戻すことがでる。例えば、Q1にあるような識別できない情報を本当にデッドに特定するという言い回しはかなりこれからも難しくなってくると思うため、してはいけないとか、法律論的な言葉にしたほうが今後も使いやすいと思う。

〇政府統計における集計された結果表である統計表と、ここで議論されている匿名データは明確に区別をする必要があると考えている。

政府統計については、匿名データが何種類か提供されている。調査票情報に対して匿名化処理を施すことによって、匿名データが作成されている。

ただし、この匿名データは誰でも利用できるのではなく、利用申請をして利用が認められたものでなければ使えない匿名データである。その点を今回の議論で整理をしておく必要があると思った。

１点目は、匿名化の技術について、さまざまな匿名化技法があり、幾つかの組み合わせによ

って匿名化処理がなされているが、どの技法の優先度を高くするかについて整理をする必要がある。

攪乱をしたデータについては、ユーザーサイドに立った場合に、そのデータはそもそも使いやすくなるのか。これについても攪乱の仕方や手法にもよると思うので、整理が必要と考えている。

また、匿名化を行う場合、どういう変数に匿名化を行うかは、Q2の照合可能性と関係する。どういったデータを照合可能性の対象になる外部情報と捉えるかによって、マッチングする変数が変わってくると思う。それでもどのような外部情報にも共通する変数がキー変数として存在すると思うので、そのあたりのマッチングの対象になるキー変数の整理を行う必要があると思った。

もう一点は、匿名化された情報を考えたときに、どこまでリスクを下げればオーケーなのか、要するに閾値をどういうふうに考えるかというのが、実は秘匿を考える、匿名化を考える場合に非常に難しい論点であり、ユーザビリティーすなわち有用性とのトレードオフの観点からも考える必要が出てくるのかもしれない。

○どこまで匿名化をするか、どこまで再識別するかというのは、多分、例えば何％とか、数字で出せる話ではないと思う。

データによって本当に100％に近い再識別の可能性を言わなければいけないデータと、緩く見てもいいデータで、例えば医療データとか政治思想といったデータに関しては絶対守らなければいけないというデータもあり、逆に位置情報のようなものは場所が動けばキャンセルできる。

そういうデータがあるので、ある程度データを考えて、リスクや再識別の可能性を見るという方向に御議論いただければと思う。

○親会で示した「パーソナルデータの取扱いルール整備に向けて検討すべき論点」の３ページ目「②匿名化されたパーソナルデータの扱い」に「合理的な水準まで匿名化を施されたパーソナルデータについて、法的に通常の個人情報とは異なる取扱い（例：第三者提供に関する同意を不要とする一方、提供先事業者に対して法的な責任を課す等）とすることの可否について」とある。また、※にあるとおり、規制改革会議の中で、合理的な匿名化措置の内容をガイドラインにとある。

まず、現行法の個人情報保護法の中で匿名化データというものを一定の整理をし、どう取り扱えばできるのかというのをガイドラインに落とすことができないかというのがある。答えとして、技術的に整理するのが難しく、統計化情報であればそういう取り扱いができまるというのは一つの答えかもしれない。

それから、制度の見直し方針を作っていくが、そのときに通常の平文の個人情報とは違う取り扱いができる匿名情報、あるいは匿名化の技術を適用した情報というものを取り扱うとき、技術と制度と運用といったFTCの３要件のようなものと組み合わせて使えるようにすべきではないか。

現行法でできることと、現行法では無理なため新たな制度的な手当てが必要なところ、２つの持っていき方があると思う。

構成員に対し、ご検討いただきたい事項の依頼があった。出席者から以下の発言があった。

〇非構造化データは、時間的余裕があればとても興味深いが、そこに拡大すると逆に抽象化するという気がする。特に非構造化データのうち、事業者が取得するものではなく、本人が発信しているものまで含めて取り扱うのは、世の中的にはどこかで議論が必要だと思うが、このWGでやるには拡大し過ぎという気がする。

〇非構造化データは扱わなくてもいいというと、今の時代を考えると、非常にバランスの悪い報告になってしまう。大体のデータが非構造化データとくっついているので、そこのところは判断してほしい。

〇私たちの仕事でも、まさに非構造化データのような定性データが増えてきているため、ある程度、個々の検討の範囲の中でそれを含むという程度でいいと思うが、全く排除するというのはちょっと難しいかもしれない。

〇資料３の１ページ目の第三者提供の類型には、いわゆる鉄道会社ケースあるかと思う。

〇含んでおり、議論いただきたい。

〇非構造化データは、現在の個人情報保護法の個人データの関係で、体系化されたデータが議論の対象かと思うため、まずはそこを主眼としつつ、非構造化データについても引き続き御相談させていただきたい。

〇合理的な水準は、作れないという気がしており、日本の状況を考えると業界団体で何か匿名化なり再識別の可能性の指針のようなものを作る形になると思う。
そういう結論に持っていくのであれば、実は合理的な水準というものは一般論の水準ではなく、ある程度、個々のケースごとに作っていくしかないかもしれない。

〇統計データまで落としてしまうというすごく消極的な解もあることはある。それを高度な匿名化まで入れるのか、もっと高いレベルまで入れるのかというところは、まだ検討する余地もあるように思う。
このままふわっとした状態でまたいろいろな会社がそういうビッグデータを解析して、たたかれて、戻りを繰り返すことになるのがすごく心配である

〇どういう情報に識別性がなくなっても権利侵害が残る情報であり、どういう情報がそうでないのかは、実は裁判所でやったり、そういう民事の紛争とは別に、個人情報の問題として、この程度の機微性であれば、この程度でいいのではないのかといった話ができるのか、どうなのかという問題もある。一概にここまで識別性をなくせばいいというのはなかなか、そもそも問題の立て方として難しいと思っている。

〇合理性があるかどうかは、そもそも我々が結論づけることなのかということもある。むしろこういう状況ではこういうリスクがあるということや、こういう方法を使うと再識別ができてしまうというように、ある程度、列挙することというのが我々のミッションかもしれない。

〇我々の業界で論文を書くときに、ｋ－匿名のｋというのはソーシャルチョイスという言い方をする。社会が選択するのだという形で、それ以上は追求しない形で論文を完結させてしまう。例えば、ｋを３にしたとき、３人に１人はHIVであるというのと、３人に１人は新宿にいるというのでは全然意味が違う。したがって、ユニバーサルな線引きというのは絶対にできなく、少なくとも、何か言うのだったら分野を絞らないと言えないというのが強く思うところ。

以上