Vol.2016-DPS-166 No.34
Vol.2016-CSEC-72 No.34
2016/3/4

匿名加工・再識別コンテストを通じた情報流通プラットフォームの検討

小栗 秀暢[†1]　黒政 敦史[†2]　松井 くにお[†2]

2015 年に成立した改正個人情報保護法により，新たに匿名加工情報という情報の類型が加わった．だが，現状において，匿名加工情報の加工基準，評価基準に加え，安全に流通させるためのシステム要件が確立されていない．その状況下，匿名加工・再識別コンテスト(PWSCUP)は，匿名加工情報の利活用とプライバシー保護に関する研究開発を活性化するため 2015 年 10 月に実施された．我々は PWSCUP の運用システムとして，匿名加工情報の登録，評価を行うためのプラットフォームを構築した．本稿では，改正個人情報保護法の条文を確認した上で，参加者の独立性を保ち，かつ公正に大会を進めるために必要となるプラットフォームの構造と，その評価方法に求められた要件を整理し，社会実装に適した匿名加工情報の流通プラットフォームの要件について検討する．安全な環境下にて匿名加工情報を作成，流通，評価する仕組みが確立されることで，匿名加工情報の利活用が促進されることを期待する．

1. はじめに

個人情報の保護に関する法律に，2015 年 9 月に成立した同法の改正法[1](以後，改正後の同法を「改正法」という)により，匿名加工情報という新たな情報の類型が定義された．匿名加工情報とは個人情報保護委員会規則で定める基準に従い，個人情報を加工して特定の個人を識別することができないようにするとともに，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(改正法三十六条)．

以下，本稿においては，後に述べる PWSCUP 内での用法に沿い，法律上の定義を意識することなく，単に何らかの匿名加工処理を行った情報を匿名加工情報などと呼ぶ場合もあるが，法律用語の場合は***匿名加工情報***などと区別してイタリック体で記載することとする．

匿名加工情報は，一定の条件の下で，本人の同意がなくても第三者に提供することが可能となる．それによって情報共有やマーケティング分析，機械学習の教師データ等への活用が期待できる．

匿名加工情報の加工基準については，今後，個人情報保護委員会の規則において定める基準に合わせて，業界ごとの認定個人情報保護団体によって指針が整備される予定である．しかし，各業界で利用されるデータ種類，利用方法，またはそのデータに適用される匿名加工処理アルゴリズム等の議論は進んでいない．

世界的に見ても，EU データ保護規則の承認や，ISO/IEC での匿名加工方式の国際標準化などが進むなか，日本国内における匿名加工処理に関する研究の振興が求められている．だが，技術の進歩や社会の変化，業種業態などによってプライバシーの保護水準は異なるため，データを取り扱う事業者は，その都度最新の規則や指針(以下「規則等」という)を参照しつつ民事責任を負うリスクも踏まえながら，作業内容を決定する必要が出てくる．そのような状況に対応するため，最新の規則等に準拠した軽量なソフトウェアとプラットフォームが必要である．

このような状況下において，コンピュータセキュリティシンポジウム 2015 では，匿名加工技術の開発と再識別に対する公平な安全性評価手法の確立を目的とした世界初の大会『PWSCUP -匿名加工・再識別コンテスト”Ice & Fire”』[2](以後 PWSCUP)が開催された．その中で，我々は，PWSCUP の参加者/実行委員に対して，匿名加工処理を行った情報を流通，評価するプラットフォームを提供した．

本稿では，PWSCUP を通じて得られた事例をもとに，匿名加工情報とその再識別リスクを評価し，安全な情報の流通を促進させるプラットフォームの構成について検討する．

2. PWSCUP 概要説明

PWSCUP は最大知識攻撃者モデル(maximum-knowledge attacker[3])を想定し，匿名加工・再識別の参加者双方が元データセットを共有した状態で行われ，匿名加工と再識別の 2 つのフェイズが存在する．

PWSCUP の運用の流れを図 1 にまとめる．

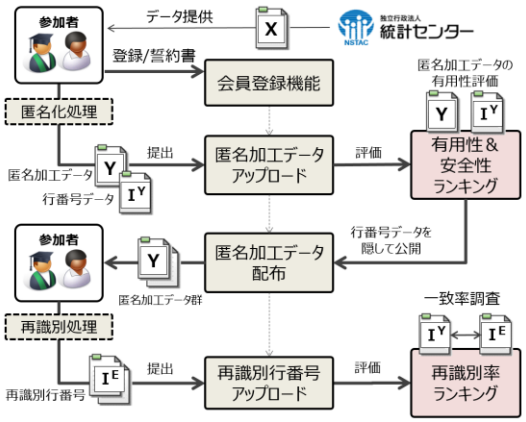

図 1　PWSCUP 運営の流れ

PWSCUP で使用したデータは，教育機関などの演習用として独立行政法人 統計センターが作成した疑似ミクロデータ[4]である．これを個人情報 X とした時，処理された匿名加工データを Y と表す．

1) プラットフォームにアクセスし，チーム登録と参加者リスト，誓約書を送付する．
2) 事務手続き完了後，独立行政法人 統計センターの WEB サイトから擬似ミクロデータ X を取得する．
3) 匿名加工フェイズ：参加者が擬似ミクロデータを匿

†1 総合研究大学院大学 複合科学研究科 情報学専攻/ニフティ株式会社
†2 ニフティ株式会社

名加工処理した結果 Y と，その行番号データ I^Y を提出する．

4) 実行委員が作成した 13 項目の評価指標によって，Y の有用性と安全性のスコア(Rank)を算出．

5) 再識別フェイズ：提出された Y を配布し，参加者は再識別行番号 I^E を提出する，I^Y と I^E を対照して正解数を算出．

これら PWSCUP で利用した評価指標とその意義，及び順位を定めるためのルール等は菊池らが発表している．[2]

PWSCUP の想定するユースケースを図 2 に示す．

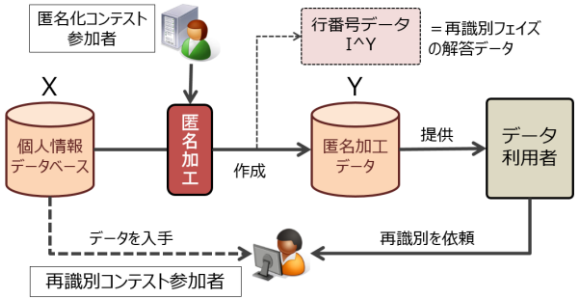

図 2　PWSCUP の想定ユースケース

1) 匿名加工フェイズ：データ提供者が個人情報 X の匿名加工データ Y を生成．連結可能キーI^Yを廃棄してデータ利用者に提供する．

2) 再識別フェイズ：X を入手可能な人物が，Y の再識別を依頼されたため，X と Y を対照して再識別処理を行う．

本ユースケースは，ある機関や組織の中で作成された匿名加工情報を外部に提供する際に，同組織内の技術者によって，提供データによる再識別リスクを評価し，安全管理指標として活用できる．

我々は，PWSCUP を運営する中で，匿名加工情報の公開と評価のシステムを構築した．その運用と改善の中で，改正法の趣旨に即した情報流通のプラットフォームの要件を検討し，部分的に機能として搭載した．

本稿は，過去のプラットフォームにおける要件と改正法が求める要件の違い，及び大会を通じて得られた知見をまとめる．

3. 従来研究

匿名加工処理(単に匿名化，匿名加工などということもある)とは，個人情報を含めた個人に関する情報（パーソナルデータ）を加工して，特定個人の識別リスクを減少させることをいう．特に k-匿名化 [5]や，Pk-匿名化[6]など，元情報に再識別される可能性を確率的に減少させたものを指す．

匿名加工処理された情報を流通させるシステムは，欧米では政府機関や医療機関において定着しており，オランダの μ-Argus 方式 [7]や，カナダの CHEO における匿名加工情報の提供[8]などが知られている．

日本においては，2010 年に経済産業省のプロジェクトとして「行動情報活用型クラウドサービス振興のためのデータ匿名化プラットフォーム技術開発事業」[9]が行われた．

その結果として，永井らがデータ匿名化プラットフォームのアーキテクチャの提案[10]を行っている．その中で，匿名化処理は，一次事業者，委託第三者機関，第三者評価機関の 3 者が関係し，対象データが 1 種類の場合と複数種類の場合でそれぞれの役割が異なる点に着目し，マルチテナント型のアーキテクチャを実現した．

だが，その課題を千田ら[11]がまとめており，匿名加工処理は，マージナル[12]等を用いた集合型匿名化アルゴリズムと，撹乱･再構築型アルゴリズムでは手法が異なるため，それぞれ個別の評価指標が必要となり，一律に利用する事は難しいとの課題を報告している．

特に，集合型匿名化アルゴリズムは，Datafly[3]等の手法が知られているが，複数の属性値を匿名化することで識別リスクは低下するが，次元の呪い[9]と呼ばれる有用性の低下が発生する．そのため，Pk-匿名化処理などの撹乱手法を併用して再識別リスクを減少させる必要があるが，その際に結果データの有用性と安全性を評価する方式が定まっていないという問題点があった[11]．

これらの技術的な課題に加え，現在においては，改正法の成立によって，新たに要件が追加されている．4 章では，改正法の条文から必要とされる要件を検討し，プラットフォームに求められる機能を整理する．

4. 改正法条文から読み取れる要件

まず，システムに求められる要件を検討するため，改正法の条文を参考に要件，及び，問題点を整理する．

本稿においては，個人情報取扱事業者における定義を行う．個人情報を保持し，匿名加工処理を実行して***匿名加工情報***を提供する側をデータ作成者，それを受領し分析等に用いる匿名加工情報取扱事業者をデータ利用者と定義する．

まず，基本的な要件として改正法からの要件を検討する．改正法より，***匿名加工情報***の取り扱いに関する技術的要件が記載されているのは第三十六条である．

> ***改正法　第三十六条より***
> ***1.個人情報取扱事業者は，匿名加工情報(略)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.個人情報取扱事業者は，匿名加工情報を作成したときは，その作成に用いた個人情報から削除した記述等及び個人識別符号並びに前項の規定により行った加工の方法に関する情報の漏えいを防止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，これらの情報の安全管理のための措置を講じなければならない．***

まず，条文より，データ作成者は，個人情報を匿名化して利用する場合「***個人情報保護委員会規則に定める基準に従い***」以下の義務を負う．

1)　特定個人の識別をできないようにするために必要な加工を行う義務

Vol.2016-DPS-166 No.34
Vol.2016-CSEC-72 No.34
2016/3/4

2) 個人情報の復元をできないようにするために必要な加工を行う義務

3) 削除した記述等の漏えいを防止するための安全管理措置を講じる義務

4) 個人識別符号の漏えいを防止するための安全管理措置を講じる義務

5) 情報の加工方法の漏えいを防止するための安全管理措置を講じる義務

1 と 2 を***匿名加工措置***, 3 乃至 5 を***安全管理措置***という．その 5 要件に加え，第三十七条において「***第三者に提供される匿名加工情報に含まれる個人に関する情報の項目及びその提供の方法について公表する***」ことが義務付けられている．

6) 提供項目と提供方法の公表義務

また，データ利用者に対しては，第三十八条において***識別行為の禁止***が定められており，データ利用者が 1) 特定の個人を識別するために 3)もとの個人情報から削除された記述等もしくは 4)個人識別符号もしくは 5)情報の加工方法を取得しようとすること，又は「***当該匿名加工情報を他の情報と照合してはならない***」と記載されている．

また，6)は，データ作成者が公表するものであるが，データ利用者が更なる加工をせず提供する場合には，内容は同一であると考えることができる．[表 1]

表 1 データ作成者とデータ利用者の義務

要件	データ作成者	データ利用者
1)特定個人の識別	できないようにするために必要な加工	他の情報と照合してはならない
2)個人情報の復元	できないようにするために必要な加工	(記載なし)
3)削除された記述	漏えいを防止	取得を禁止
4)個人識別符号	漏えいを防止	取得を禁止
5)情報の加工方法	漏えいを防止	取得を禁止
6)公表義務	提供項目と提供方法	再提供時の公表義務

改正法で定められた要件について，データ作成者が利用契約を結んだデータ利用者に対して，***匿名加工情報*** Y を提供する場合を考えたものが図 3 である．

まず個人情報データベース X に対して，匿名化アルゴリズム（情報の加工方法）A_i を用いて処理し，***匿名加工情報*** Y，個人識別符号 I^Y，加工に伴う中間データや削除された情報 X_i を出力した．また，それらの情報を，安全管理措置＋提供基準 R_i に沿って漏えいを防止する措置を講じた．この場合における提供基準とは，k-匿名性などの再識別リスクの定量基準や倫理基準などの定性基準を指す．

このとき，R_i，I^Y，X_i，A_i の基準を，個人情報保護委員会規則等で定める基準に合致させ，提供された個人に関する情報の項目と提供方法の公表を行うことでデータ利用者に Y を提供できる．これら R_i，I^Y，X_i，A_i の基準は，個人情報保護委員会，又は認定個人情報保護団体が作成する規則等に沿って決定される．だが，これらの値は常に一定ではなく，学術向けのデータ提供やグループ会社での共同利用の場合など，提供先の信頼性に基づいた提供基準の変化も発生する．

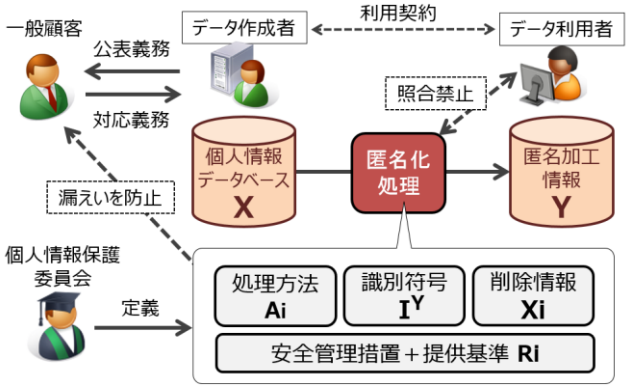

図 3 改正法で定められた要件の概念図

そのため，匿名加工情報の提供と受領に関するシステムの要件定義を行う場合，データの利用法や社会背景等のコンテクストに応じた安全性基準やアプリケーションを柔軟に組み入れるシステムが求められる．

だが，これらの改正法が求める要件をシステムとして展開する際に，解釈が難しい箇所が複数存在する．

まず，**5)情報の加工方法の漏えい** を防止しながら，**6)情報の項目と提供方法を公開** する必要がある点がある．

これは，加工方法を類推されない範囲での情報提供と考えることが出来る．本要件は，条文的には分離している概念だが，実際に運用した場合には情報開示レベルのコントロールが必要となる．

例えば，企業 A が企業 B に対し，個人情報を匿名加工し，データベース形式で提供した，という情報が公開された場合を考える．

これを企業 A が HP やサービス利用規約などで公開した場合，顧客から企業 A へ問い合わせが来ることが予想される．例として「提供した情報は安全なのか」「安全性基準は何か」「匿名加工ソフトウェアは何か」等が挙げられる．

公開範囲を広くしたり，丁寧に質問に対応したりすると **5) 情報の加工方法の漏えい** につながる可能性がある．現在，多様な個人情報の匿名化アルゴリズムや評価指標が提案されており，それらの顧客からの質問，情報開示要求全てに対して，合理的な回答を行うことは難しいであろう．

しかし，規則等に即して「加工方法の詳細を伝えることは出来ない」という対応を行った場合，顧客の不信感の増大，オプトアウト希望ユーザの増加，風評被害など，データ作成者のリスクが大きく，自社データを提供するモチベーションが低下する．

ちなみに，改正法第三十六条 6 項には，***当該匿名加工情報の作成その他の取扱いに関する苦情の処理その他の当該匿名加工情報の適正な取扱いを確保するために必要な措置を自ら講じ，かつ，当該措置の内容を公表するよう努めなければならない*** と，データ作成者側が匿名加工情報の取り扱いに関する苦情処理等を行うことが義務とされて

おり，データ作成者の負担は重い．

この他の問題として，**1)特定個人の識別** について，データ作成者が特定の個人を識別できないようにするための匿名加工措置を行う必要があるが，同時にデータ利用者側による照合を禁止している．そのため，データ作成者が匿名加工措置を行っているという前提で受け取った情報が，識別可能な個人情報である場合もあるが，それを検証することができない．そのため，悪質な情報提供者による偽った匿名加工情報の提供や，匿名加工情報と間違えて個人情報を送付してしまった場合に対処できないという問題がある．

宇賀[13]は『個人情報保護指針で定めた加工方法を遵守しない結果，特定の個人が識別されるものが「匿名加工情報」として作成され利用・提供されていないかを注意深くモニタリングすること』が望ましいと述べているが，現時点では，有効なモニタリング手段が存在しない．

これらの問題に対して，我々はデータ作成者とデータ利用者の間に情報流通プラットフォームを介することで解決する方法を検討した．

5. 匿名加工情報の流通プラットフォーム検討

第 4 章にて問題提起した内容をまとめる．

1) 匿名加工処理とその評価指標への柔軟な対応
2) 一般顧客に対する公表の方法とその根拠
3) データ利用者に対する匿名加工措置の保証

この 3 項目が改正法に対応するためにシステム側が対応するべき課題だと設定する．

現実的にはこれ以外にも検討するべき安全性基準は存在するが，それらは規則等，及び契約等で条件を変更することが可能であり，改正法で禁止されている要件ではない．そのため，本稿では言及しない．

5.1 プロトタイプ ver1.0 の検討

そこで，我々は匿名加工情報を流通する上での課題を解決できる仕組みとして，データ作成者とデータ利用者の間をプラットフォーム事業者がつなぐ方式を検討し，図 4 のプロトタイプ（Ver.1.0）を開発した．

プラットフォームは，主に以下の要素で構成されている．

1) サービス利用者管理・履歴保存システム
2) 匿名加工システム
3) 安全性評価システム
4) 利用者管理/安全性基準 管理システム

まず前提として，図 4 のプロトタイプ ver1.0 は，プライベートクラウド内に構築され，限られたユーザしかアクセス出来ない．かつ，その環境下でデータ作成者ごとに，独立したハード/データベース領域を生成し，処理を行う．

1)サービス利用者管理システムは，システムの根幹である．まず，利用者同士はネットワーク上のアクセス管理権限とセッション上の権限情報の二重の制限によって，他ユーザのデータベース環境を閲覧することは出来ない．

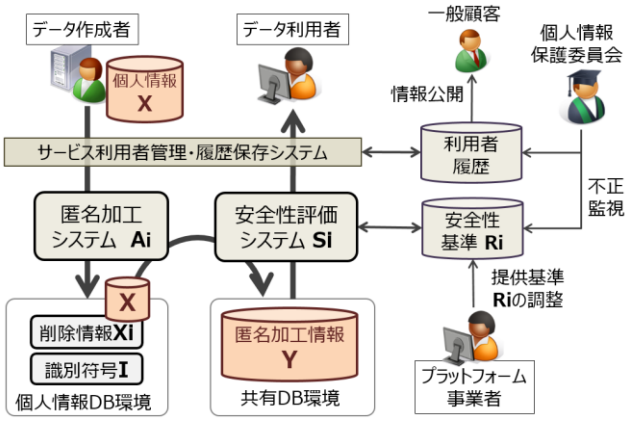

図 4 プロトタイプ ver1.0 概念図

同環境下において，ユーザは 1 人に 1 つずつ別環境における個別データベース領域を設置し，識別リスクが評価されていない個人情報 X や，削除・中間ファイル X_i，個人識別符号 I^X も含め，全てが個別の環境下で安全に管理される．

特に，過去に発生した匿名加工措置の問題点である，識別子の管理方法と消去のタイミングを検討し，ユーザが操作する中で[識別子の消去→新 ID への書き換え]を意識する必要が無いシステムを実現した．

また，各人には権限が指定されており，使用できる機能と取り扱い可能な情報の k-匿名性をコントロールすることで，高い権限を持つユーザは低い匿名性，低い権限を持つユーザは高い匿名性を実現しないと，共有 DB 環境にデータを公開できない仕組みとなっている．

ユーザは，個人の権限に合致した範囲で **2)匿名加工システム**にデータを投入する．匿名加工システムは，データの種類に合わせて複数存在し，ユーザ自身が選択する．

作成された匿名加工情報は**3)安全性評価システム**を通過する際にデータの評価を行い，問題ないものだけがデータ利用者に対して公開される．また，公開する匿名加工情報のデータベースを共有にすることで，検索エンジンを搭載することが出来るため，公開情報を横断で検索し，そのまま複数情報ソースを比較検証することが可能である．

データ作成者が投入した R_i，I^Y，X_i，A_i 等のデータを **4)利用者管理/安全性基準 管理システム**にて，正式な手順を踏んで匿名加工措置を行っていることを検証する．一般顧客に情報公開する場合は，このシステムと履歴データベースから適切なもののみを出力する．

プロトタイプ ver1.0 の問題点は大きく 2 点存在する．

1 つ目は，データ型式の多様性を受容するためにデータの型式とエラーチェック処理が多く必要となる点である．

データ作成者は，常に投入するべきデータのプロパティについて正確に認識している訳ではない．具体的には，マスター型かトランザクション型の選択，又はカテゴリ属性と数値属性の区分，又は言語情報や位置情報における辞書データ等を多く必要とし，システムが肥大化した．

もう 1 つは評価システムである．評価システムは既存の

プラットフォーム実験[7][9]等で行われた定性評価と定量評価を含めた総合評価システムを参考に複数搭載し，データの総合的な評価を試みた

だが，全てのデータに一律に適応できる評価方式が存在せず，結局は審査する側の定性評価が主になってしまった．

そこで，プロトタイプ ver1.0 では，最終的に全てのデータを識別子がユニークに存在するマスターデータ型に変換し，k-匿名性を共通評価指標として，関係者によるデータ流通実験等に利用してきた．

結果として，1 番目の問題点は，匿名化処理プログラムを自由に設置，改良可能なプラットフォームとして構成することで，軽量化と運用性を高めることに成功した．だが 2 番目の問題である評価システムの問題に関しては，解決する方法が見つからず，検討を続けていた．

その評価・検討にあたり，本プロトタイプの評価と改良を目的に，PWSCUP 運営用システムとして改修を行った．それが PWSCUP プラットフォームである．

5.2 PWSCUP プラットフォームの検討

PWSCUP のルールとシステム要件を検討する中で，ベースとなった ver1.0 から変更点が多く発生した．代表的なものをまとめる．

1) **大会は個人単位ではなくチーム単位で参加**
2) **匿名加工/再識別処理は参加者のローカル環境**
3) **使用データを擬似ミクロデータ[6]に固定**
4) **評価指標が 13 個あり，各値の順位で評価**

匿名加工・再識別コンテストは，世界でも行われた例が無いため，これらの問題に対してプラットフォームを随時改良しながら対応を行った．

表 2　PWSCUP 実施イベント日程

実施イベント	期間
予備戦（匿名加工フェイズ）	2015/08/24～9/24
予備戦（再識別フェイズ）	9/25～10/9
本戦（匿名加工・再識別）	10/21

今大会では，予備選と本戦の期間を分離し，予備選にてルール，システム，評価指標の問題点を洗い出し，本戦までの期間に改良し，反映させる形で行われた[表 2].最終的な概念図を図 5 にて示す．

まず **1) チーム単位での参加システム**を開発した．PWSCUP では，参加チームに対して，実行委員が ID/Pass を発行し，内部でチームとして結合するシステムを追加して解決した．近年では，開発ツールや機械学習などの共同作業ツール等の要件でも同様の機能が求められており，ロール毎の権限設定については今後も改良が必要である．

2)匿名加工/再識別処理が参加者のローカル環境で行われることに関して，システム上の対応は軽微であった．だが，匿名加工処理に関する問題点を多く現出させた．この点は 7 章にて記載する．

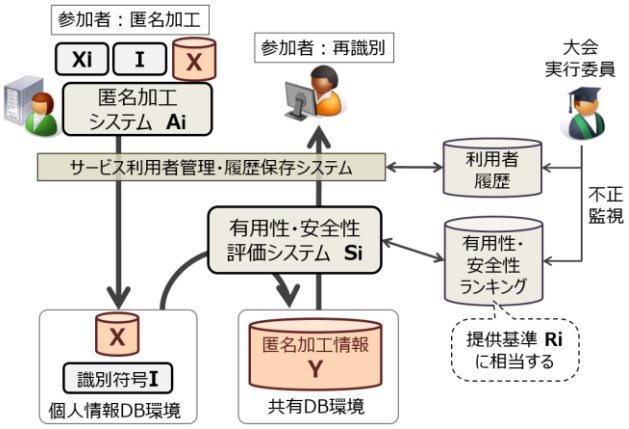

図 5　PWSCUP プラットフォーム概念図

3)使用データが擬似ミクロデータに固定されたことは，システムの簡素化に貢献した．だが，ベースとなる ver1.0 において入力データの正規化チェック等が行われている箇所と，大会における提出データのレギュレーションが大きく異なったため，システムの対応が必要となった．

PWSCUP で用いられた擬似ミクロデータは，QID と SA 属性が明確に区分されており，各属性値の表示バイト数も指定されていたことから，内部ロジックとして持っていたデータの正規チェックシステムの多くが不必要となった．

4) 評価指標が 13 個あり，各値の順位で評価 については，最も改修が多いものであった．本来，ver1.0 では，各匿名加工情報が 1 つの評価指標によって評価される仕組みで設計されており，複数，それも 13 個の指標による評価は想定しておらず，評価システムは全て再設計された．これらの問題点は 7 章にて記載する．

6. システム構成

PWSCUP プラットフォームの簡易機能構成図を図 6 にて示す．本システムは ver1.0 時代の基本設計を踏襲しており，評価システム，ファイル管理システム，共有 DB，個別 DB 領域等は 1 サーバ～全て独立した環境に展開することも可能な形で構築されている．

PWSCUP プラットフォームは，ニフティクラウド内に設定されたプライベートクラウド領域における SaaS を志向して開発された．全体のセキュリティ要件として，Firewall やアクセス権限管理があるが，それらの要件は事業者ポリシー，サーバ構成，アクセスするユーザの役割等によって異なるため，本稿ではスコープ外とする．

通常の SaaS の場合，利用者数が増加するとサーバの CPU コア数やメモリ量を増加する形でのスケールを志向するが，本システムでは，各利用者による独立性を重視するため，個人環境のサーバ台数を増加する方向に設計されている．

クラウドサービスの発達により，近年ではハイスペックなサーバ性能を更に向上させるより，追加のサーバ台数を増加させる方が，コストパフォーマンスが良い場合がある．

Vol.2016-DPS-166 No.34
Vol.2016-CSEC-72 No.34
2016/3/4

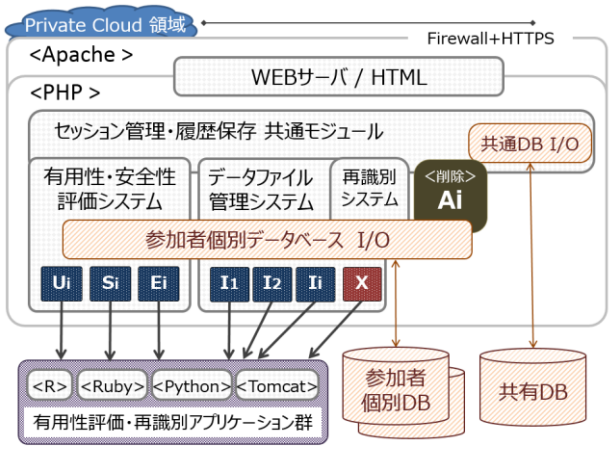

図 6　PWSCUP プラットフォーム簡易機能構成図

各 WEB サーバ環境や DB 環境がセキュアな状態で独立している場合，結合可能キー（行番号データ I^Y）や中間・削除ファイル Xi 等の管理は各 WEB サーバ内で行われている．個人情報に類する情報を KVM やファイルサーバ等の複数事業者の共有ストレージに設置することはセキュリティ上好ましくないため，この構成とした．だが，WEB サーバ上に設置することもセキュリティ上問題がある．Ver1.0 では全てのデータは個人情報として DB に格納しており，消去や撹乱，ハッシュ化等の操作が可能であった．これらファイル管理方法は今後も検討が必要である．

また，増加するデータベースと WEB サーバの制御も共有 DB 内で行われており，ユーザは個人のサーバ環境を意識することなくセキュアな匿名化処理が可能である．

処理サーバが分散される場合，各アプリケーションは軽量で運用性が高い方が望ましい．近年では企業向け HP の作成には WordPress が多く用いられており(w3techs.com 調べ 2016 年 1 月 CMS シェア 58.8%)，運用性の高い Apache + PHP + Mysql の，所謂 LAMP 環境での開発需要が高まっている．本プラットフォームも LAMP 環境で作成されており，多くのサービスとの親和性が高い．

表 3:　システム仕様

種別	詳細
サーバ	ニフティクラウドサーバ
システム	ニフティ匿名化処理プラットフォーム
CPU	Intel(R) Xeon(R) 3.00GHz
Memory	4 GB (負荷に応じて可変)
Apache	Apache/2.2.3
PHP	PHP 5.5.30
R 言語	R version 3.2.0
Java	Java(TM) SRE 1.8.0 45
Ruby	ruby 2.2.2p95
Python	Python 2.7.10
Mysql	mysql　Ver 14.14 Distrib 5.6.15

だが，評価システムの実装にあたり，必要とされるミドルウェアの数が非常に多くインストールされている．今後もデータ型式の多様性に伴い，多くの評価用プロトタイプアプリケーションが搭載されることが予想されるため，今後は評価プログラムとプラットフォームを抽象化するための共通 API 層が必要になるだろう．

最終的なサーバスペックとインストールされたミドルウェア等の情報を表 3 に示す．予備選は WEB＋DB を含めた 1 台での構成．本戦は WEB1 台＋DB1 台の 2 台構成で運営された．

今回の大会における参加者は 18 チーム 84 名，その内，本戦に参加したのは 30 名程度であったことから，このスペックにて通常の動作に支障は無かった．だが 7 章にてあるように，データクレンジングの失敗とミドルウェアのバグにより，中断が発生したことは今後の課題である．

7.　PWSCUP プラットフォームの課題

PWSCUP プラットフォームの課題は多く存在するが，本稿では改正法から導いた要件について主に記述する．

7.1　評価指標に関する課題

今大会の運営において，最も多くの時間議論されたのは評価指標のあり方についてである．

表 4　PWSCUP で利用した指標

No	指標	指標説明(使用言語)	必要情報
U1	meanMAE	SA 平均絶対誤差(R)	X,Y
U2	crossMean	クロス集計値の平均絶対誤差(Ruby)	X,Y
U3	crossCnt	クロス集計数の平均絶対誤差(Ruby)	X,Y
U4	corMAE	SA の相関係数の平均絶対誤差(R)	X,Y
U5	IL	データ各値の平均絶対誤差(Ruby)	X,Y,I^Y
U6	nrow	データのレコード数(システム)	Y
S1	k-anony	k-匿名性指標の最小値(R)	Y
S2	k-anonyMean	k-匿名性指標の平均値(R)	Y
E1	IdRand	QI からランダムな再識別率(Ruby)	X,Y
E2	IdSA	QI から SA15 列による再識別率(Ruby)	X,Y
E3	Sort	SA 総和ソートによる再識別率(Python)	X,Y
E4	SA21	SA21 列について再識別率(Ruby)	X,Y
E5	AYA	山岡攻撃の検知による再識別率(Ruby)	X,Y, I^Y

PWSCUP 本選では Ui : 有用性評価，Si : 安全性評価(k-匿名性)，Ei : 安全性評価(再識別率)，計 13 指標が設定された[表 4]．各指標に対応した評価アプリケーションは，実行委員が作成し，プラットフォーム側からの処理要求によって稼動する．

本指標は法律要件側から見ると 3 種類に分類できる．

1)　匿名加工情報 Y だけで評価する指標

2)　元情報 X と匿名加工情報 Y を比較した指標

3)　元情報 X，匿名加工情報 Y，行番号 I^Y を利用する指標

PWSCUP は個人情報を扱わないため問題ないが，本システムが社会実装された場合，上記区分は大きな意味を持つ．

匿名加工情報 Y を評価し，流通させるだけの場合，個別のデータ作成者と個人情報取扱契約を結ぶ必要はない．だが，2) 3)の指標が必須になる場合，プラットフォーム事業者は複数の個人情報取扱事業者との間で契約を交わす必要があり，情報の管理と授受に関わる要件が大きく異なる．

本大会では，数値の絶対評価を用いず，最終的なスコアの決定に相対的な順位(Rank)を用いた．だが，これらの評

価システムを社会実装する際には，他者との相対評価よりも絶対的な数値評価が必要となる．データを利用するコンテクストと合わせて評価数値を設定し，今後の評価システムに適用する必要があるだろう．

また，各指標は実行委員が提案，評価アプリ開発，課題発見を繰り返すため，実データでの検証工数が非常に多く必要となった．これは，ver1.0 において，投入データが多様であることから，データの正規化処理とアルゴリズムが肥大化し，運用性が低下した問題と同様である．

今大会などを通じて有効性が実証された評価指標やアプリケーション等は標準パッケージとして，広く一般的に用いることができる仕組みを実現すべきである．

7.2 データ不正検知の限界

最も ver1.0 と比較して変更された要件として，匿名加工アルゴリズムを内部に設定せず，各参加者の環境に任せた部分にある．本来的には，本システムは匿名加工処理の実行中に発生した不正要件を検知するための仕組みであった．

例えば，匿名加工情報を提供する際に，あらかじめ定められた安全性基準を達成しているか，又は，データに対して過度な改ざん等を行っていないかを確認することを意図してシステム履歴機能を設計してあった．

PWSCUP では，各参加者の開発環境にて匿名加工処理を行われた後の，結果データのみを投入する仕組みであるため，システム上から不正を検知することが非常に難しい．

その例として，ユーザによる山岡攻撃[2]についての対応を紹介する．山岡攻撃とは，PWSCUP ルール制定時に課題となった匿名加工方法である．匿名加工データ Y と行番号データ I^Y を作成する際に，I^Y を故意にランダム化して並び替える処理を指す．本処理によって，匿名加工情報の平均絶対誤差や k-匿名性を変化させずに，再識別リスクだけを減少させることが出来る．

実行委員内で本問題に対する検知方法について検討し，U5(IL 指標)にて検証し，明らかな問題が発生した場合には本指標で投稿されたデータを判定し，別途対応する方針とした．だが，予備選終了後に各指標にて応募データを検証したところ，山岡攻撃を行っているデータに対して特徴的な傾向を見出すことが出来なかった．(図 7)

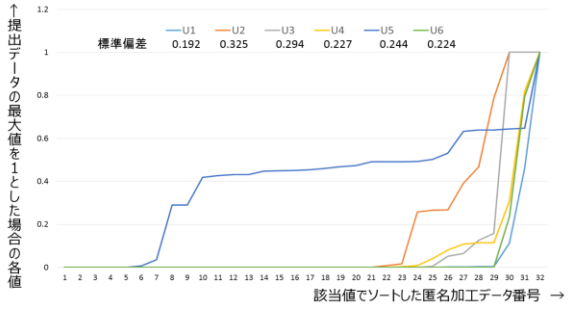

図 7 予備選における各参加者の有用性指標分布

実行委員側では，山岡攻撃に対する完全な検知が不可能と判断し，ルールを改正し，山岡攻撃が総合順位等への影響が無いことを明記した．だが，このルール改正によって，安易なランダム化処理が増加することを懸念し，新たな指標 AYA を作成し，検出を試みた．

結果として，本戦では IL 指標，AYA 指標ともに機能し，最終順位の判定に大きく寄与したが，今後もルールと指標の不備による問題が多く発生する可能性がある．

本問題は，提出された匿名加工データについて，データ作成者が偽装を行うことによって，プラットフォーム側が検出することなく市場に流れる可能性を示唆している．

本来ならば，匿名加工データは個人情報保護委員会又は認定個人情報保護団体が作成する規則等が定める基準に従い適切に加工されることが前提となっているが，もし指定の処理を採用しなかった場合，プラットフォーム側での検知を行う仕組みを必要とする．その際，プラットフォーム側に処理の手順等が示されていない場合，検知が非常に難しい．各業種業態に特徴的なデータが存在するように，その検知にも独自のアルゴリズムが必要になるだろう．

また，本大会では，再識別処理も各参加者の環境にて行われたことから，その手法の正当性が証明できない．例えば複数チームの結託によって結果データを共有し，再識別のスコアを増加させることが可能である点は事前に指摘されていたが，それらの問題への対応もしていなかった．

今後も，新しい匿名加工・再識別方法が開発される可能性がある．大会後の目視判定では結託や不正の形跡が無いことから，今回は参加者のモラルによって維持されていたと考える．これも今後の課題である．

7.3 再識別アルゴリズムの共有

今大会において，再識別率という共通基準によって安全性を評価したことは大きな成果である．実際に再識別された数による評価基準によって，k-匿名性処理後の識別リスクを，理論的ではなく，実数に表すことが可能となった．

だが，これらの再識別処理については，第 4 章で定義した **6) 提供項目と提供方法を公表**について課題となる箇所が見えてきた．PWSCUP 本戦終了後，処理手法についてのプレゼンテーションが実施されたが，再識別の手法については大きく 3 種類に区分された．1 つ目は，実行委員が作成した再識別アルゴリズムと，その改良による再識別試行である．これは，実行委員側での想定の範囲である．

2 つ目は，参加者に公開された有用性・安全性指標を用いるやり方である．匿名加工コンテストは，その性質上，有用性と安全性の結果数値を公表する必要がある．

だが，それらの結果数値が詳細に公開されている場合，各種のアルゴリズムによって匿名加工処理を行い，最もその数値に近づいたパターンを探索し，処理方法を類推するという手法が多く見られた．今回の大会では小数点 8 桁(予備選)，6 桁(本選)の数値が公表されたが，これは，今後の社会実装に向けて，匿名加工データを公開する際に留意するべき点である．

3 つ目に機械学習による再識別処理が存在する．今回の PWSCUP に参加したチームは 18 チーム，最終的に予備選には 32 データが投稿された．この匿名化処理結果を用いて機械学習にかけた結果によって，本選での再識別率が大きく向上することが判明した．今後，匿名加工情報が一般的に流通する場合，処理担当者等の特徴から機械学習によって元データへの再識別や復元が可能になるリスクが考えられる．今後も，継続的に新しい技術によるプライバシー侵害の可能性を検証する場が必要とされるだろう．

7.4 匿名による参加の可否

今回の PWSCUP では，匿名によるチーム参加が認められていたため，多くの企業参加者はチーム名のみ公開され，その所属企業・研究機関やメンバー名などは伏せられている場合が多かった．

この制度は功罪の両面を持つ．一つは，企業名等の公開がされないことから，参加者は企業への申請作業等が簡便となり，参加者の間口が広がること．だが，その反面，参加者が作成したアルゴリズムや手法に関する説明等には，企業としての特許や研究成果が含まれる可能性があり，公表されづらいという問題がある．

大会終了後にアンケートを取得し，採用された手法を比較，要件毎に優れた手法を検証する予定だったが，上記の問題によって，分析できない状態である．このようなイベントを開催しながら，学術的な成果として共有されないことは大きな損失であるため，今後の方針を検討したい．

7.5 想定外データへの対応

本大会で利用した多くの評価用ソフトウェアは，プロトタイプ的に作成されていることから，データの正規化チェックアプリケーションを専用に作成し，厳格に管理した．

そのため，ユーザからはデータ生成コストが重い旨の意見を頂いた．投入データセットと評価アプリケーションの不具合に起因する問題は何度か発生した．最も重大な問題は，PWSCUP 当日の大会中に評価システムがトラブルによって停止した点にある．

大会終了後に理由を検証した所，評価システムに想定外のデータが投入された場合，メモリオーバーフローが発生しサーバ処理が停止することが判明した．

これはデータ処理前のデータクレンジングの問題でもあるが，異常なデータについての検証を厳しくすることは事前処理の肥大につながることから対応が難しい．

Ver1.0 のシステムでは，セキュリティ上の対応として，DB サーバと WEB サーバを他ユーザのシステムと分離して稼動するシステムを構築した．だが現実的には，不具合影響範囲の限定を目的として再設計した方が良いと考える．今後の大会運営時に構成の変更を検討したい．

8. まとめ

PWSCUP は世界初の匿名加工・再識別コンテストではあったが，匿名加工情報の利活用に向けて，実行委員，及び参加者の皆様と様々な課題を共有できたことと，情報プラットフォームがデータ利活用とプライバシー保護において重要な役割を担っていることを確認できたことに大きな意義があったと考える．

今後はデータ型式の範囲を広げ，利用目的に応じたプラットフォームの役割や機能を評価する必要がある．是非，今後も PWSCUP のような取り組みを通じて，安全性と有用性を両立したデータ利活用が振興することを期待する．

謝辞

コンテスト実施に向けたプラットフォーム開発にあたり、菊池先生，佐久間先生，山口様，濱田様，山岡様[2]をはじめ PWS 実行委員の皆様からの助言，激励を頂き，厚く御礼申し上げます．また、指標やルールの周知問題，当日のシステムトラブルなど運営の不手際にもかかわらず，辛抱強くご対応いただいた参加された皆様に感謝いたします．

参考文献

1) 個人情報の保護に関する法律及び行政手続における特定の個人を識別するための番号の利用等に関する法律の一部を改正する法律(平成 27 年法律第 65 号)
2) 菊池 浩明,山口 高康,濱田 浩気,山岡 裕司,小栗 秀暢,佐久間 淳, "匿名加工・再識別コンテスト Ice & Fire の設計",コンピュータセキュリティシンポジウム 2015 論文集, 2015(3), pp.363-370,2015-10-14
3) J.Domingo-Ferrer,S.Ricci,J.Soria-Comas, "Disclosure Risk Assessment via Record Linkage by a Maximum-Knowledge Attacker", 2015 Thirteenth Annual Conference on Privacy, Security and Trust (PST), IEEE,2015.
4) 秋山 裕美,山口 幸三,伊藤 伸介,星野 なおみ,後藤 武彦, "教育用擬似ミクロデータの開発とその利用～平成 16 年全国消費実態調査を例として～", 統計センター製表技術参考資料,16,pp.1-43,2012
5) L.Sweeney, “k-anonymity: a model for protecting privacy”, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, pp.557-570, 2002
6) 五十嵐 大,千田 浩司,高橋 克巳,"k-匿名性の確率的指標への拡張とその適用例",コンピュータセキュリティシンポジウム 2009(CSS2009) 論文集,2009,pp.1-6,2011-10-12
7) A.Hundepool, L.Willenborg, Statistics Netherlands, “Record Linkage Techniques μ and t-ARGUS: Software for Statistical Disclosure Control”,pp.142-149,1997
8) K.Emam and L.Arbuckle, "Anonymizing Health Data Case Studies and Methods to Get You Started", O'Reilly, 2013
9) 経済産業省, (株)日立コンサルティング, “行動情報活用型クラウドサービス振興のためのデータ匿名化プラットフォーム技術開発事業」事業報告書”,2012-03
10) 永井 康彦,五十嵐 亮基,羅川 広行,松岡 健,藤田 麻里央,佐藤 祥太郎,美馬 正司,"行動情報活用型クラウドサービス振興のためのデータ匿名化プラットフォーム「匿名化クラウド」のアーキテクチャ提案",研究報告情報セキュリティ心理学とトラスト(SPT),2011-SPT-1(26),pp.1-8,2011-07
11) 千田 浩司,五十嵐 大,高橋 克巳,濱田 浩気,冨士 仁,"集合匿名化クラウドの課題と対策",研究報告インターネットと運用技術(IOT),2011-IOT-13(21),pp.1-6,2011-05
12) D. Kifer,J. Gehrke, “Injecting Utility into Anonymized Datasets”, Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data, pp. 217-228, 2006.
13) 宇賀克也,"個人情報・匿名加工情報・個人情報取扱事業者", ジュリスト 2016 年 2 月号,No.1489, pp.36-41, 2016-01-25