大規模データの利活用のあり方と匿名加工に関する一考察*

伊　藤　伸　介

1．はじめに
2．海外における行政記録データの利活用の現状―アメリカとカナダの事例を中心に―
3．パーソナルデータに対する匿名加工の可能性
4．統計目的と統計情報について
5．おわりに

1．はじめに

諸外国では，各国の法制度に基づきながら，公的統計データだけでなく，行政記録データや民間が保有する個人情報（パーソナルデータ）といった大規模なデータの利活用が行われている．こうした大規模なデータに関しては，行政記録データやパーソナルデータのリンケージを通じたさらなる広範な利活用が展開されている．

行政記録情報（データ）については，諸外国では，レジスターベースの統計を作成している北欧諸国を中心に，その利活用が進められている（伊藤（2017））．わが国でも，第Ⅲ期「公的統計の整備に関する基本的な計画」（平成30年３月６日閣議決定，一部変更後令和２年６月２日閣議決定，以下『基本計画』と呼称）において，第Ⅱ期基本計画に続いて，行政記録情報等の利活用の推進が明記されている．具体的には，「行政記録情報等の活用」の中で，総務省に対して，「行政記録情報等の統計作成への活用に係る実態調査を充実させるとともに，諸外国の取組状況も踏まえつつ行政記録情報等の活用に係る研究を基礎・実用の両面から推進する」だけでなく，「関係府省と連携し，報告者の同意を得て行政記録情報を調査票への記入に代えるなど統計の作成に活用すること」を検討することを求めている．さらに，『基本計画』では，「行政記録情報から作成する業務統計について，ユーザーのニーズを踏まえた提供情報の充実等に取り組むことにより，行政記録情報等の利活用の推進を図り，その利活用状況や課題等に関して，統計委員会や各府省との間で情報共有・横展開を進める」ことを明記している．

さらに，民間が保有する大規模なパーソナルデータの利活用も注目されている．営利目的のた

＊　本稿の一部は，伊藤（2020）に基づいている．

めの個人情報の第三者提供だけでなく，民間が持つ大規模なデータを用いた公的統計（政府統計）の作成可能性が追究されている．『基本計画』は，「民間企業等が保有するビッグデータの活用」の中で，「民間企業等が保有するビッグデータを新たな統計指標や分析に活用するための検討が進められている」ことに言及している．こうした状況において，個人情報の利活用に関しては，法制度面から個人情報のプライバシーを守った上で，個人情報の活用が技術的にどのように可能になるかについての議論がなされてきた．その結果，2015年 9 月に改正された「個人情報の保護に関する法律」（以下「個人情報保護法」と呼称）が公布され，2017年 5 月から施行されている[1)]．この法律では匿名加工情報に関する条文が明記されただけでなく，この条文に基づいて匿名加工情報に関するガイドラインが整備されてきた（個人情報保護委員会（2016））．

このように，公的統計データだけでなく，行政記録データや民間の個人情報といったデータの利活用においては，個人に関する情報についての安全性の確保と大規模なデータに対するニーズ（有用性）を勘案しつつ，個人に関する情報の秘密保護と利活用のバランスをどのように図っていくかが重要な論点になると言える．

本稿では，海外における行政記録データの利活用の動向を明らかにした上で，行政記録データやパーソナルデータといった大規模データの公的統計への利活用に向けた論点の整理を行う．さらに，統計目的の観点から見た大規模データから作成される統計情報の特徴についても議論する．

2．海外における行政記録データの利活用の現状
―アメリカとカナダの事例を中心に―

北欧諸国（フィンランド，スウェーデン，ノルウェー，デンマーク），さらにはオランダといった国々では，行政記録データに基づいた統計作成システムを確立してきた（森（2009a, 2009b），伊藤（2017））．これらの国々は，出生あるいは移住した時点で，個々人に識別子としての ID が付与されており，この ID を介して，各種の行政記録データのリンケージが可能な状況になっている．それによって，基本的な人口社会情報だけでなく，税務，社会給付，教育，雇用，医療・健康等に関する情報を行政記録データとして把握することが可能になっている．

デンマークやノルウェーといった北欧諸国においては，様々なレジスターに含まれる行政記録データが統計局に自動的に集められており，社会保障番号（social security number）が仮名化された形で統計局に保管されている．仮名化された ID を用いて行政記録データのリンケージを行うことによって，各種の公的統計が作成されるだけでなく，利用者には，こうした仮名化された個票データ（非識別データ，deidentified data）の学術目的のための広範な利活用が可能になっている．

1） 2015年に改正された個人情報保護法から 5 年が経過した2020年 6 月 5 日に，現行の個人情報保護法の一部が改正された「個人情報の保護に関する法律等の一部を改正する法律案」が成立している．

それに対して，レジスターベースではなく，統計調査を実施することによって公的統計を作成しているイギリスでは，個々人の様々な社会経済的属性を連結するための共通の ID が存在しない．そうした状況の中で，イギリスにおいては，名前や住所といった直接的な識別子に基づいて仮名化された ID の作成と，仮名化された ID を用いたリンケージの方法論に関する研究が進められている．また，イギリス国家統計局において，行政記録データのリンケージに基づいた人口センサスデータの作成に関するプロジェクトが展開されてきた（伊藤（2017））．

本節では，イギリスと同様に調査票ベースの統計作成を行っているアメリカとカナダを対象に，北米諸国における行政記録データの利活用の状況を述べることにしたい．

2–1　アメリカセンサス局における行政記録データの利用状況

本節においては，アメリカセンサス局における行政記録データの活用状況を概観する．アメリカセンサス局は，MAF（＝Master Address File）と呼ばれる住所情報と TIGER（＝Topologically Integrated Geographic Encoding and Referencing）と呼ばれる地理的情報を有している．MAF と TIGER は，アメリカセンサス局内部において連結されており，MAF/TIGER として一体化されている（森（2007））．具体的には，それは，geocode によって住所情報と地理的情報がリンクされていることを意味する．

MAF/TIGER に含まれる各レコードは，郵便局（U.S. Postal Service）が郵便物を送付するための住所ファイルとして，DSF（＝Delivery Sequence File）と呼ばれるファイルを備えている．そして，郵便物を受け取る者は誰でも DSF に含まれる．郵便局から DSF が無料でアメリカセンサス局に送付される．それによって，LACS（Locatable Address Conversion System）ファイル等が DSF に連結されている．

MAF/TIGER の作成に関する 3 つの情報源として，上述した DSF，実査による確認（American Community Survey 等），および LUCA（Local Update of Census Addresses）プログラムを指摘することができる．LUCA プログラムは，州，郡や都市の地方政府当局とセンサス局がパートナーシップを持った場合に可能になるプログラムである．アメリカセンサス局は，地方政府当局に MAF の住所リストを送付し，地方政府当局は住所の追加や削除を行うだけでなく，geocode を付与する．地方政府当局は地図や境界線を更新し，更新された情報をセンサス局に送付する．それによって，アメリカセンサス局は，MAF/TIGER のデータベースを更新する．なお，近年では，地方・州政府が住所情報を共有することが可能になる地理的支援システム（Geographic Support System＝GSS）が開発されている．

MAF は，世帯や企業の住所情報を含んでいるが，個別主体の名前を含んでいない．ゆえに，個人が引っ越しても追跡することができないようになっている．こうした措置をとることによって，MAF/TIGER における個体情報が管理されている．また，住所と geocode は合衆国法典第13

編（Title 13 of U.S. Code）で保護されている．

1990年までは，アメリカセンサス局は毎回ゼロから住所リストの作成を行っていた．民間部門から住所リストを購入し，リストを確認した上で，欠損している住所を追加する．1990年人口センサスまでは，センサスが終了するたびに，住所リストを廃棄していた．しかし1990年人口センサス以降，住所リストは保管されており，そのリストに新規の住所が追加されている．この住所リストは，システム上で管理されている．6か月ごとに，郵便局から住所ファイル（DSF）を取得し，1年を通して新規の住所があればファイルに追加されている．

アメリカセンサス局は，税務データといった行政記録データを利用しようとする場合，行政記録データに含まれている住所とMAFの住所データを照合した上で，一意な識別子であるMAF（Master Address File）IDを作成し，行政記録データに付与している．これによって行政記録情報と人口センサスのミクロデータに関しては，MAF IDによるリンケージが可能になる．

さらに，アメリカセンサス局の行政記録研究・分析・利活用センター（Center for Administrative Records Research and Analysis and Applications＝CARRA）では，センサス局の担当部局から提供されたMAFを行政記録データとマッチングした上で，社会保障番号や税務データに関する識別番号が暗号化されたIDである個人識別コード（Personal Identification Code＝PIC）を新たに生成している．MAF IDが住所情報のIDであるのに対して，PICは，個人レベルでのIDと言える．このPICを用いることによって，行政記録データや公的統計データのリンケージを行うことが可能になる．なお，2010年センサスの場合，データの90％に対して個人の識別IDの付与が可能であることが確認されており，PICとMAF IDを用いて行政記録データとリンケージを行うことが可能な状況になっている．

PICは，CARRAとアメリカセンサス局経済研究センター（the Center for Economic Studies）によって管理されている．したがって，例えば，内国歳入庁（Internal Revenue Service＝IRS）が管理する税務データがアメリカセンサス局によって入手されると，税務データにPICとMAF IDが付与される[2)]．

アメリカでは，北欧諸国のような人口社会的な情報に関するレジスターは存在しない．そのため，アメリカセンサス局では，行政記録データの利活用の方向として，北欧諸国のように行政記録に基づくセンサスに関する統計の作成は，2020年センサスでは指向されていない．2010年センサスの場合，不在となっていた未回答世帯については，調査員が再訪問を行うという形で再調査を実施していたために，未回答者への再調査（事後調査，follow up）に伴うコストが発生してい

2） 税務情報は，合衆国法典第26編（Title 26 of U.S. Code）に基づいて管理されており，研究者による行政記録データへのアクセスも制限されている．税務情報の利用にあたっての審査・承認システムも存在する．

た．こうしたコストの低減を図ることを目的として，アメリカセンサス局では，2020年人口センサスにおいて税務データやメディケア（Medicare）等の行政記録データを活用することが追究されている．図１は，アメリカセンサス局における未回答者の再調査に関するモデルを示したものである．未回答者世帯の再調査（Non-response Follow Up）のための住所を確認するにあたって，自宅に不在の世帯（vacant）の住所や存在しない世帯（non-existent）の住所を定めるために行政記録データが活用されている．行政記録に記載されている住所に調査票を郵送し，配送可能だった場合には，未回答者に対して調査対象者から調査票の回答がなされるか，あるいは調査員の訪問による聞き取り調査が行われる．その一方で，行政記録に記載されている住所に対して調査票が配送不能だった（Undeliverable-As-Addressed）場合には，未回答者は，「行政記録データにおいて自宅に不在の世帯」かあるいは「行政記録データにおいて存在しない世帯」に類別され記録される．また，行政記録データの利用の対象外となった未回答の世帯に対しては，再度の面接調査が実施されるが，未解決の場合には，在室である（occupied）世帯の住所を定めるために行政記録データが利用され，行政記録に記載されている住所に調査票が郵送される．それによって，調査対象者からの回答が得られることが期待できるが，場合によっては行政記録による補完が行われる．

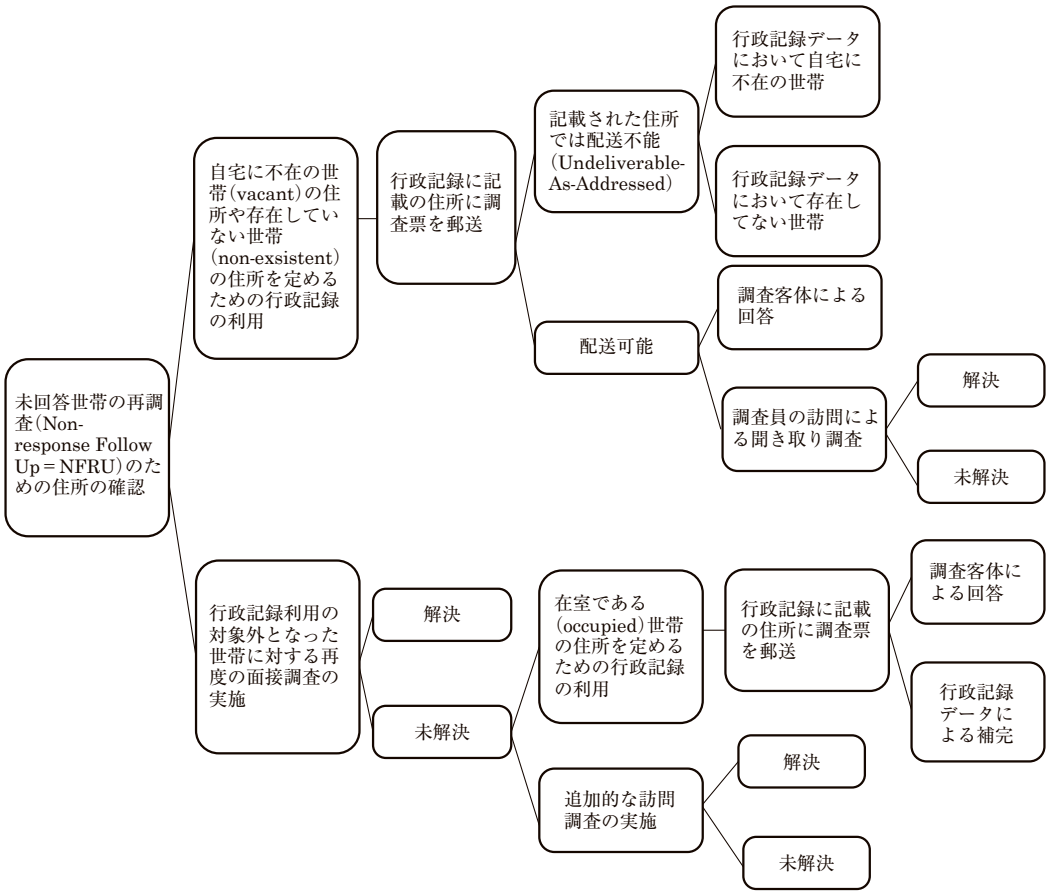
このように，アメリカセンサス局では，行政記録データに用いて未回答者に対する属性情報を把握することが指向されている．このことは，アメリカセンサス局において行政記録データが統計調査の効率的な実施のための補助情報として機能していることを示唆している．

2-2　カナダ統計局における行政記録データの利活用について

本節では，カナダ統計局を例に，カナダにおける行政記録データの利活用の状況を明らかにする．人口センサスデータの作成における行政記録データの利活用の観点から見た場合，カナダ統計局は，税務データ等の行政記録を用いて人口センサスの統計表の作成可能性を追究してきたと言える（Lebel and Denis（2016））．カナダ統計局では，住所レジスターを保管しているが，住所レジスターの更新においては，15種類（2017年８月時点）の行政記録データが用いられている．具体的には，税務データ，電話サービスの料金明細ファイル（telephone billing file），カナダ郵政公社（Canada Post Corporation）の住所ファイル等の最新データを用いて，３か月に１度住所レジスターが更新されている．そして，人口センサスの実施において，部分的に住所レジスターが利用されている．

カナダ統計局においては，様々な行政記録情報に含まれる社会保険番号（Social Insurance Number＝SIN），運転免許証番号，健康保険証番号等の一意になっている直接的な識別子が，リンケージを可能にするために保管されていることが知られている．さらに，カナダ歳入庁（Canadian Revenue Agency）からの行政記録情報，移民，出生，死亡に関するレジスターの利用の可能性

図 1　アメリカセンサス局における未回答者の再調査に関するモデル



注）アメリカセンサス局におけるヒアリング調査（2017年 8 月24日）の資料に基づき作成

も議論されている．

カナダ統計局における行政記録データの場合，行政記録データのリンケージにおいては，確率的ではなく，探索的な（heuristic）決定論的マッチング（deterministic matching）が用いられている[3)]．すなわち，名前，住所，出生年，性別，社会保険番号，移住に関する識別子といった直接的な識別子を用いたマッチングが行われている．

カナダ統計局には，人口レジスター（statistical population register）は創設されていない（2017年 8 月時点）．しかしながら，カナダ統計局の担当者によれば，個人の統計数値（individual statistical number）が縦断的に連結された構造が想定される．このことから，カナダ統計局が，人口レ

3）　2017年 8 月時点のインタビューでは，カナダ統計局の担当者は，イギリス国家統計局において検討が進められたハッシュ化に基づく確率的なリンケージについては懐疑的な反応であった．

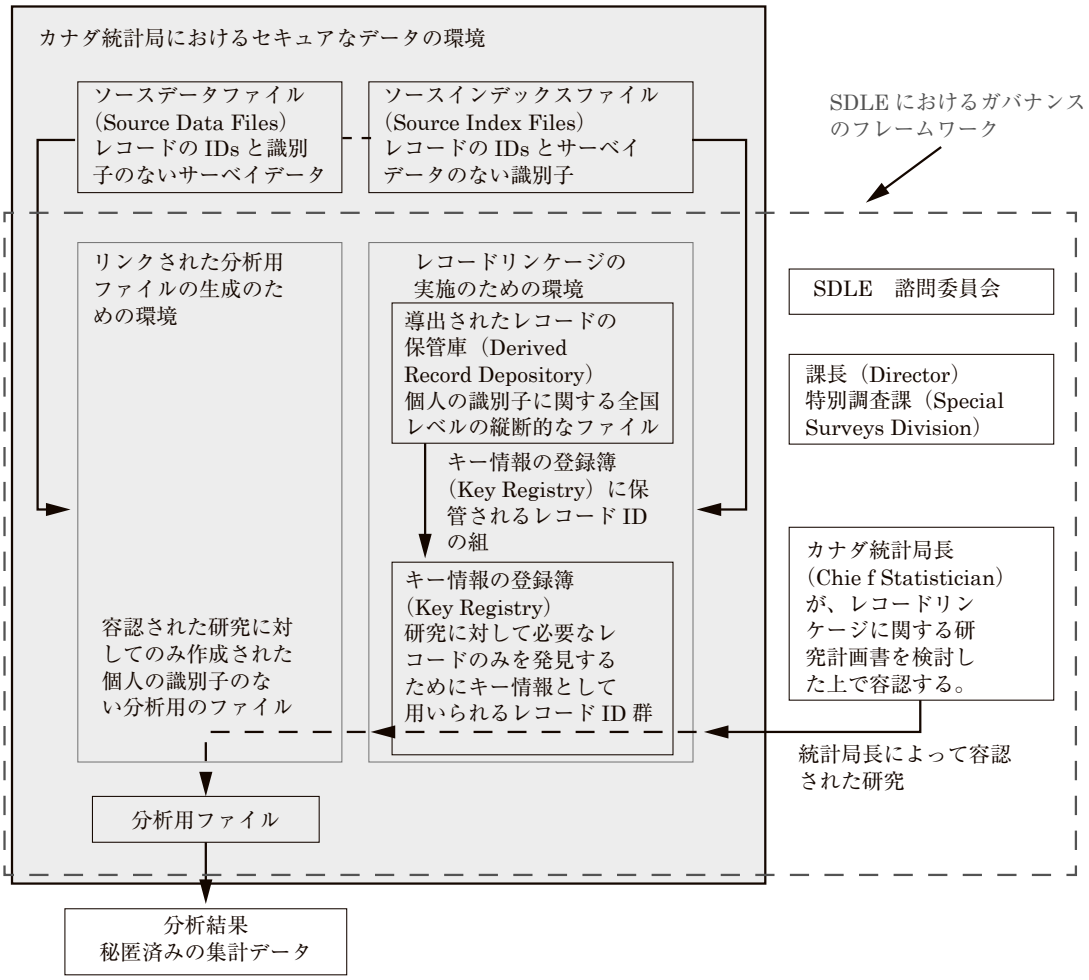
ジスターを構築するためのレジスターベースの統計システムを指向していることが推察される．

カナダ統計局では，学術目的のための行政記録データの二次利用として，Social Data Linkage Environment（SDLE）と呼ばれる二次利用のシステムが展開されている．SDLE とは，レコードリンケージを行うためのセキュアな環境であり，人口社会分野を対象にしたレジスター（social register）と類似したシステムを備えている．図 2 は，SDLE の概略図を示したものである．SDLE においては，サーベイデータが含まれないが，識別子とレコード ID が保管されたインデックス変数用のファイル（ソースインデックスファイル，Source Index File），および識別子が削除されたサーベイデータとレコード ID が含まれている分析用変数のためのファイル（ソースデータファイル，Source Data File）が別々に保管されている．インデックス変数用のファイルには，性別，出生年月日，郵便番号，社会保険番号といった直接的な識別子が含まれる．さらに，SDLE においては，個人の識別子に関する全国レベルの縦断的なファイルである「導出されたレコードの保管庫（Derived Record Depository）」が備わっている．この導出されたレコードの保管庫に含まれる個人の識別子とソースインデックスファイルから移管されるレコード ID と識別子に基づいて，キー情報の登録簿（Key Registry）にレコード ID の組が移管される．このレコード ID の組は，キー情報の登録簿において研究に対して必要なレコードのみを発見するためのキーとなる情報として用いられる．そして，行政記録データのリンケージが行われた上で，個人の識別子が含まれない形で，研究に必要な分析用ファイルが生成される．

SDLE におけるガバナンスのフレームワークとしては，SDLE の担当部局だけでなく，カナダ統計局長（Chief Statistician）と SDLE 諮問委員会が，SDLE における統計組織上の役割を果たしている．カナダ統計局における SDLE プロジェクトでは，申請者は，行政記録データのリンケージされたデータを用いた研究計画書を提出する必要がある．カナダ統計局長がレコードリンケージに関する研究計画書を検討した上で申請を承認すると，その調査研究は，統計局長から容認された研究として位置付けられる．レコードリンケージを実施するための環境では，リンケージの対象となるデータ源からデータを抽出した上で，ID だけでなく，ソースインデックスファイルに含まれる直接的な識別子も用いながら，行政記録データのレコードリンケージが実施される．つぎに，リンクされた分析ファイルを生成するための環境において，レコード ID を用いて，サーベイデータとのリンケージを行われた上で，個人の識別子を削除することによって，容認された調査研究に対してのみ必要な分析用ファイルが生成される．

リンケージが施された分析用の個票データは，個票データの利用のためのセキュアな環境を備えるリサーチデータセンターのみでアクセスすることが可能である．こうしたリンクされた個票データに関しては，一般公開用のミクロデータ（Public Use Microdata File）は作成されておらず，あくまで学術研究目的のための個票データとして利用が可能になっている．

図２　カナダ統計局における SDLE の概略図



注）カナダ統計局におけるヒアリング調査（2017年８月28日）の資料に基づき作成

3．パーソナルデータに対する匿名加工の可能性

2015年に改正された個人情報保護法が成立して以降，個人情報の利活用と保護に関する社会的な関心が高まる中で，事業者が持つパーソナルデータを匿名化した上で，匿名加工情報として営利目的で第三者に提供する動きに大きな注目が集まっていた．

パーソナルデータには多種多様なデータが存在するが，パーソナルデータに含まれるレコードは，基本的には直接識別情報（氏名，住所，ID，個人情報保護法で定義されている「個人識別符号」），間接識別情報（性別，年齢等の自然的社会的属性），および履歴情報から構成される．パーソナルデータの対象となる履歴情報には，主として，地域（地点）情報（緯度・経度情報）の移動履

歴や購買履歴といった情報が該当する．このような移動履歴や購買履歴は，単位，標識，時間と場所の制約を受ける静態的な（static）情報としての統計情報とは異なり，イベントヒストリー的な性格を持つ動態的な（dynamic）情報だと言うことができる．

こうしたパーソナルデータの匿名加工において，公的統計の分野で議論されてきたミクロデータに対する匿名化手法が，匿名加工の方法を検討する上での先行的な事例となってきたと考えられる．公的統計における匿名化ミクロデータの作成では，リコーディング（global recoding, local recoding），データの削除（record suppression, attribute suppression），トップ（ボトム）・コーディングといった非攪乱的な手法が，統計実務において適用可能な匿名化技法として主に用いられてきた．さらに，ノイズ（加法ノイズ（additive noise），乗法ノイズ（multiplicative noise）），スワッピング（data swapping），ラウンディング（丸め）（rounding），ミクロアグリゲーション（micro aggregation），PRAM（Post RAndomisation Method）のような攪乱的手法（perturbation）に関しても，公的統計の分野ではその適用可能性が検討され，データ特性を踏まえて，スワッピングやノイズ等の攪乱的手法が適用されてきた（Zayatz（2007), Lauger *et al.*（2014），伊藤（2018a），伊藤（2018b），伊藤（2019））[4)]．

パーソナルデータに対してどのような匿名化技法を適用すべきかについて検討する上で，匿名加工が施されたパーソナルデータの有用性や秘匿性に関する定量的な評価（Yancey *et al.*（2002），Shlomo（2010）等）を行うことも求められる．こうした定量的な評価によって，各種のパーソナルデータに対して適切な匿名化技法を判断するための材料として有益な情報を提示することが可能になる．

パーソナルデータに対して，秘匿性を高めるための１つの方法は，直接識別情報以外の変数から準識別子（quasi-identifier）を選定した上で，k 匿名性（k-anonymity）のような基準を設けて，準識別子（あるいはキー変数）が同じ値を有するレコードが k 個以上存在するように匿名化の処理を行うことである．これについては，パーソナルデータにおけるサンプリング（レコードの一部抽出）の適用可能性が論点になることが考えられる．そこで，購買履歴情報や移動履歴情報を含むパーソナルデータにおいてサンプリングの有効性を明らかにするには，パーソナルデータに対してサンプリングを適用した場合の標本一意と母集団一意の関係を定量的に把握する必要があろう．パーソナルデータでもサンプリングが適用可能であることが実証的に示されれば，サンプリ

4）非攪乱的手法については，そのメリットとして，変数値に攪乱（ノイズ）を入れていないことから，匿名化ミクロデータの利用者側にとっては，実証分析における分析結果の有効性を担保しやすいことが指摘されるが，必要以上の非攪乱的な処理を行った場合には，情報量損失がより大きくなるリスクがあることがデメリットだと言える．一方で，攪乱的手法に関しては，作成者側にとっては，変数値に攪乱（ノイズ）を入れることによるブラフ効果を期待できる点が大きいが，攪乱的手法における有効性を作成者側が保証する必要性がある点が，デメリットとみなされる（伊藤（2019））．

ング以外の他の匿名化技法（例えばｋ匿名化等）を組み合わせることによって，有用性と秘匿性の両面からパーソナルデータの匿名加工のためのより望ましい匿名化措置も可能になるだろう．

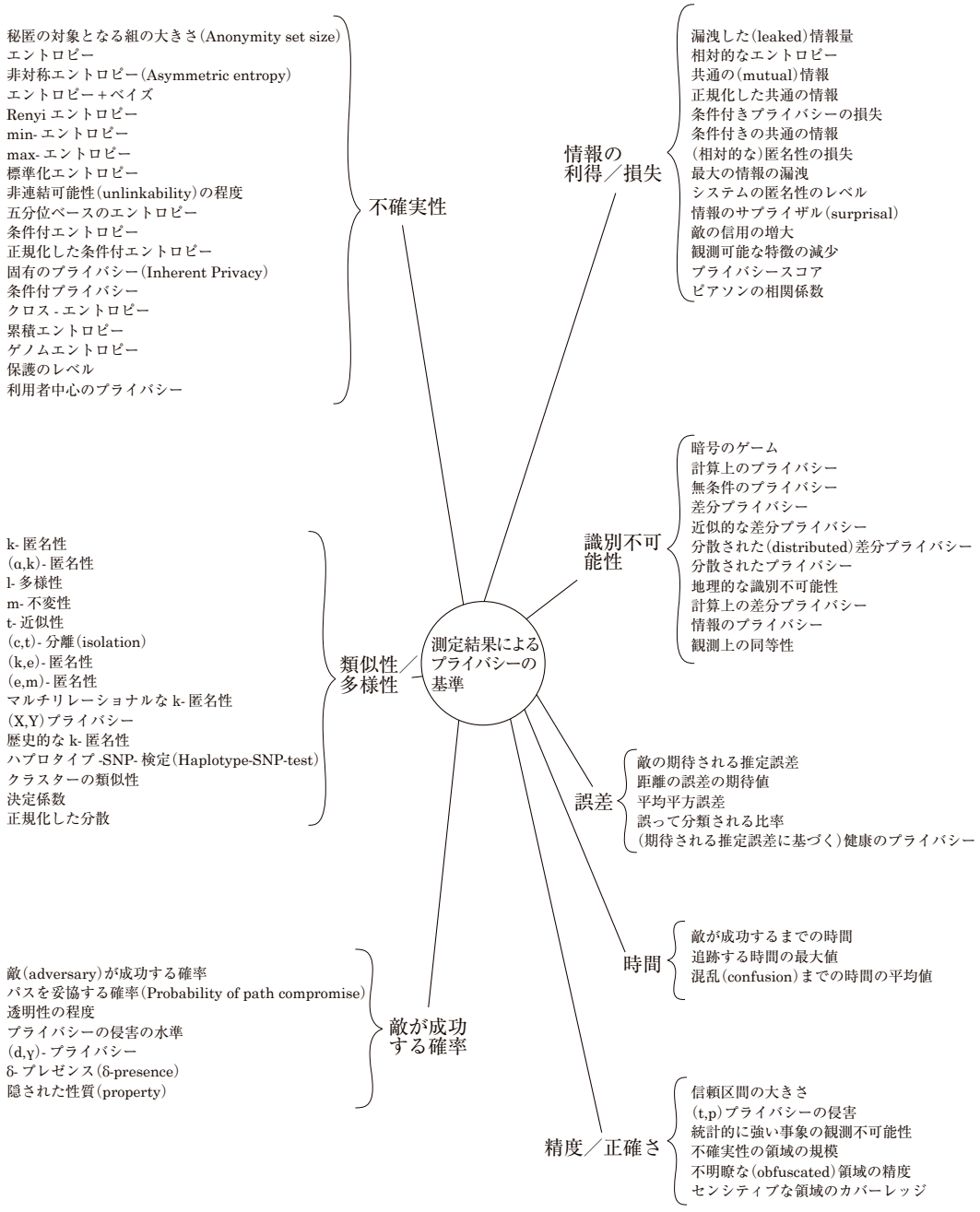
ところで，秘匿性の定量的な評価基準に着目するコンピュータサイエンスの分野では，プライバシーに関する様々な基準（metrics）が存在しており，それは，図３のように類型化されている．図３では，測定結果に基づくプライバシーの基準が，「不確実性」，「類似性／多様性」，「敵（adversary）が（個体の特定に）成功する確率」，「情報の利得／損失」，「識別不可能性」，「誤差」，「時間」，「精度／正確さ」の８つの基準に類型化されることを示している．「不確実性」については，主にエントロピーに基づく評価基準が含まれている．「類似性／多様性」に関しては，例えば，k- 匿名性（k-anonymity）といった個人情報の保護においてこれまでも議論されてきた評価基準が該当するだけでなく，回帰分析でモデルの説明力の指標として用いられている決定係数も含まれている．また，識別不可能性については，差分プライバシー（differential Privacy）が含まれる．差分プライバシーとは，「「ある個人のデータを含むデータベースに対する問い合わせ結果が，その個人のデータを含まないデータベースへの問い合わせ結果と区別できないなら，その問い合わせは安全である（個人に関するプライバシーを開示しない）」という考え方によりプライバシーを規定する」基準である（寺田他（2015，1803頁））．

パーソナルデータの利活用においては，統計情報に加工した上で第三者提供を行う可能性も議論されている．秘匿性の程度を高めるために，位置情報といったパーソナルデータを統計情報（統計表）という形で提供することも考えられる．統計表が単位，標識，時間と場所によって規定されることを勘案すれば，パーソナルデータを統計表の形で提供可能にするには，単位，標識，時間と場所に関する秘匿措置を具体的に検討する必要があろう（伊藤（2018b））．

パーソナルデータの場合，氏名や住所のような ID の仮名化が議論されてきた（内閣官房 IT 戦略本部パーソナルデータに関する検討会（2013））．具体的には，パーソナルデータにおける仮名化の方法として，①乱数の生成，②ハッシュ関数の利用，③一連番号と ID（直接識別情報）との対応表の活用等が指摘されている（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」（2017））．そのような仮名化は，個人情報保護法の条文に照らし合わせてみても，個人情報から匿名加工情報を作成するための要件となっている（個人情報保護委員会事務局（2017））．

公的統計の場合，仮名化はつぎのように考えることができる．記入済みの調査票には名前や住所といった直接的な識別子が含まれているが，統計表の元になる調査票情報は，記入済の調査票に含まれる情報を反映させた原データ（original data）にのみ付与される直接的な識別子（氏名，住所等）が削除された非識別データ（deidentified data）と位置づけられうる．その意味では，公的統計の作成において，直接的な識別子の削除および直接的な識別子と一意に対応しない一連番号への置き換えは，匿名化の一手法としても議論されうるものの，仮名化という形では論じられない．その意味では，公的統計のミクロデータは統計作成のプロセスにおいて「匿名性」を有す

図３　プライバシーの評価基準に関する類型化



出所）Wagner and Eckhoff（2015, p. 6 ）

る非識別データである．公的統計における匿名化されたミクロデータは，こうした「匿名性」を持った非識別データを対象に，間接識別情報の組み合わせによる個人の特定化の可能性を低減させる観点から，各種の匿名化技法を追加的に適用することによって作成される．

一方，行政記録データの場合，行政記録に含まれる氏名や住所といった個体識別情報は，個別的な行政記録情報の識別を行う上で必要不可欠なものとして備えられている．統計情報として集団特性を把握しようとすれば，氏名や住所といった個体識別情報は必要でないことから，個体識別情報は，個票データの中のレコードから切り離される．そして，公的統計を作成するために行政記録データに含まれる個体識別情報の仮名化が行われる（行政レジスター（administrative register）から統計レジスター（statistical register）への転換（浜砂（2010））．その場合，例えば，デンマークやオランダにおける統計作成部局では，他の行政記録データとのリンケージを可能にするために，統計作成部局内のセキュアな環境で，個体識別情報と仮名化された ID の対応表に関する管理が行われている．このことが，レジスターベースのシステムを備える国々において，行政記録データから公的統計の作成を可能にしていると言える（伊藤（2017））．

パーソナルデータにおいては，個人に付与される ID を用いて，様々なパーソナルデータのリンケージが可能になる．そうしたリンケージされたパーソナルデータをもとに作成された統計情報の第三者提供や外部への公開を行おうとすれば，パーソナルデータに含まれる個人情報の仮名化あるいは削除は，統計情報（統計表）を作成するための基本的なプロセスの１つに位置付けられるだろう．その上で，個体が特定されないような秘匿処理を施すことによって，個人情報とは切り離された形で統計情報（統計表）に加工することが求められよう．

4．統計目的と統計情報について

本節では，パーソナルデータから作成される統計情報の特徴を考察するために，公的統計における統計情報を対象に，調査目的あるいは利用目的の観点からその特徴をさらに追究してみたい．

一般に，公的統計における統計情報（統計表）の元になる統計単位情報[5)]には，個別主体に関するあらゆる情報は含まれていない．その理由は，公的統計ミクロデータに含まれる変数（および変数値）は，調査計画において事前に設定される集計表の集計事項（調査事項）に限定されるからである（伊藤（2018b））．それは，調査計画に関して大屋が指摘した「逆順の原則」が作用していることを意味する．すなわち，「調査計画では「逆順の原則」が強調される．「逆順」とは，調査の

5） 大屋によれば，統計単位情報は，「調査票に記入された個々の記録すなわち記入済みの調査個票」であり，「調査目的に適合するように設計された一連の統計表の様式に従って集計され，表示あるいは表章される」単位情報である（大屋（1995, 77–78頁））．

実施順序すなわち調査票の作成→実査→集計・分類→表示の順序とは逆に，表章計画→集計・分類計画→実査計画→調査票の設計の順に，調査計画は具体化されなければならないという意味である」（大屋（1995，78頁））．

ところで，統計調査の調査設計は，調査目的によって規定される．大屋は，調査目的について以下のように論じている．

> 調査対象を表章する統計表の様式とそこに表示される数字の種類とを統一的にとらえて統計形態（statistical form）とよぶならば，一連の具体的な統計形態が予想ないし期待される調査目的の実現の姿である（大屋（1995，78頁））．

> 政府の統計調査もそのような情報（筆者注 政府が必要とする「各種の情報」）の一形態としての統計の獲得を目的とするものであるから，政府統計の調査目的は国家目的の統計における現われとみなさなければならない（大屋（1995，58頁））．

また，大屋によれば，調査目的には，「社会的顕著事項にたいする国家目的に沿う統計需要の形式」と「関心がもたれる特定の集団と部分集団について，それらの数量的特徴を特定の統計的表章によって獲得すること」の２つの側面が存在することを述べている．

統計単位情報は，公的統計の「調査目的に適合するように設計された」上で収集された単位情報であるから，統計単位情報を集計することによって，様々な結果数値が表章可能になる．ゆえに，それが「国家目的」にそぐわない場合には，公表される統計表から捨象されることが考えられる．また，統計表は，「国家目的の統計における現われ」である調査目的に適合する形で集計計画に設定される．このことは，統計表に含まれる集計事項の数やそのカテゴリーについても，調査目的の観点からその妥当性が規定されていることを示唆している．

さらに，わが国の統計法制度では，統計作成部局が基幹統計調査の承認を得るために，集計計画の中で対象となる統計表が，調査票で捕捉されるすべての調査事項に基づいて作成・公表されるかどうかについて，総務省政策統括官室（統計基準担当）からの審査を受ける．これは，現行の統計法第９条第２項に対応しているが，統計作成の観点からは，調査目的→集計計画→統計表→調査事項の対応関係が，法制度的には統計表の審査のプロセスの中でも確認されていると言うことができる[6)]．

6）　統計法第９条第２項に基づく審査のプロセスは，わが国では統計作成部局が作成する統計表の公表において，統計表における個体情報の露見リスクという観点からの審査を統計委員会のような機関で行わないことにも符合している（伊藤（2015））．

ところで，統計作成部局が各種の統計表を包含しうる高次元の集計（結果）表を想定した場合，集計計画においてそうした高次元の統計表を設定することは概念上は可能であろう．その場合，その統計表の公表可能性は，調査目的に適合的かどうかに依拠する．高次元の集計表が形式的技術的には作成可能であったとしても，統計表のセルに含まれる統計数値の秘匿の観点だけでなく，統計作成部局がそのような高次元の集計（結果）表，さらには高次元の集計表に含まれる集計事項の範囲において変数における分類区分を再統合することによって新たに作成される統計表の公表が調査目的に対して適合的だとみなさなければ，統計作成部局がそうした統計表を作成・公表するのは困難だと言えよう．

それに対して，公的統計の二次利用の側から見れば，公的統計の個票データの利用者は，利用目的の範囲内であれば個票データから高次元の集計表を作成することは可能だと考えられる．しかしながら，利用者が高次元の集計表を持ち出そうとする場合，公表する前に利用者が作成した集計表に関する秘匿性のチェックが必要になる．そのために，ヨーロッパ諸国では議論された個票データに基づく「安全な」統計に関するチェックリスト（Brandt *et al.*（2010），伊藤（2016））が整備され，わが国でもチェックリストに基づく分析結果の事後的な秘匿性のチェックが具現化されてきた．このように利用者の側でも，分析結果の公開においては，秘匿の観点から個票データに基づいて作成された高次元の集計表のチェックが求められるだけでなく，こうした集計表が利用者がミクロデータの利用申請において設定した利用目的と妥当するかどうかという観点も求められよう．

なお，工藤は，「統計目的」の原則と「秘匿性」の原則との相剋を指摘している（工藤（1986））．すなわち，統計目的とは，「一般的知識を増大させること，すなわち，人またはその他の実在物の集まりについて，その大きさ，傾向および諸関係を知ることにある．統計記録とその内容を個別に識別することは，集計されるデータを集積し編整する者以外には，すべての者に対して秘匿されている．個人の記録は，その記録を統計的な集計値，平均値あるいは関係の尺度に用いる以外は，個人に影響を及ぼすいかなる決定にも利用されない」（American Statistical Association（1977））．工藤が指摘する「統計目的」には，統計作成部局にとっての統計の作成目的，および利用者にとっての公的統計データの利用目的の両方が含まれると思われる．そして，個別主体の特定化を行うことなく，「人またはその他の実在物の集まりについて，その大きさ，傾向および諸関係を知る」ために統計情報を作成するのであれば，個体情報の秘匿性の担保を図ることが可能であることが示唆されている．

5．おわりに

本稿では，行政記録データや民間のパーソナルデータのような大規模データの利活用のあり方

を追究するために，最初に海外の事例として，アメリカセンサス局とカナダ統計局における行政記録データの利活用の現状を明らかにした．アメリカやカナダは，レジスターベースではなく，統計調査を実施することによってセンサスに関する統計を作成している国々である．とくに，カナダ統計局は人口センサスの実施において，調査対象者のインターネットによる回答の促進を図ってきた．アメリカセンサス局やカナダ統計局において，行政記録データに対する取り組みの方向性や利活用の目的は異なっていても，統計作成部局内部で行政記録データの利用が積極的に展開されようとしていることは，非常に興味深いと考える．

つぎに，民間のパーソナルデータにおける匿名加工および統計情報としての第三者提供における論点整理を行った．公的統計の匿名化ミクロデータの作成にあたって適用される非攪乱的な手法や攪乱的手法は，パーソナルデータに対する匿名加工の方法としても適用可能である．ただし，公的統計ミクロデータとは異なり，パーソナルデータでは，匿名加工データの作成目的やそのデータ特性が異なることから，適用可能な手法も異なってくる．例えば，パーソナルデータに対して，高度な攪乱的手法としての差分プライバシーの方法論が広範に展開されている．しかしながら，公的統計の分野においては，海外でもアメリカセンサス局の人口センサスの事例（Abowd（2018 ）を除けば，公的統計の実務における本格的な議論はまだ始まったばかりだと言える（Ito and Terada（2019））．

また，パーソナルデータの場合，個人情報の直接的な識別子に対する仮名化のプロセスが求められるが，公的統計の場合，個体識別子が削除された調査票情報は公表される統計表の元データである統計単位情報として機能している．ゆえにパーソナルデータを統計情報として第三者提供しようとすれば，こうした統計単位情報から構成される集団の特性が，パーソナルデータから作成された統計表に備わっていることが求められるだろう．

公的統計における統計目的（調査目的，利用目的）は，作成・公表する集計表を規定する．このことは，集計表に含まれる集計事項やカテゴリーの数にも制約をもたらす．こうした公的統計の議論を踏まえると，パーソナルデータに基づく統計情報の作成目的が，そのような統計情報のデータ構造，さらには第三者への提供可能性に影響を与えることがわかる．

行政記録データやパーソナルデータのような大規模データの公的統計への利活用の可能性は，わが国でも今後さらに模索されるだけでなく，パーソナルデータを統計情報という形で作成し，第三者提供を行うこともより一層追究されるであろう．その意味では，行政記録データやパーソナルデータと公的統計データとの間のさらなる相互連関が図られていくと同時に，これらの大規模データの特性の違いを踏まえた上で，大規模データの利用可能性に関する議論を行う必要性がさらに高まっていくのではないだろうか．

付記：本稿は，「2018年度中央大学特定課題研究費」における成果の一部を発表したものである．

参 考 文 献

伊藤伸介（2015）「公的統計データの匿名化について―パーソナルデータの利活用における基盤整備との関連を中心に―」『中央大学経済研究所年報』第46号，457–478頁.

伊藤伸介（2016）「諸外国における公的統計ミクロデータの提供の現状とわが国の課題」『中央大学経済研究所年報』第48号，233–249頁.

伊藤伸介（2017）「公的統計における行政記録データの利活用について―デンマーク，オランダとイギリスの現状―」『経済学論纂（中央大学）』第58巻第1号，1–17頁.

伊藤伸介（2018a）「国勢調査における匿名化ミクロデータの作成可能性」『経済志林』第85巻第2号，241–277頁.

伊藤伸介（2018b）「公的統計ミクロデータの利活用における匿名化措置のあり方について」『日本統計学会誌』第47巻第2号，77–101頁.

伊藤伸介（2019）「公的統計データにおける秘匿性と有用性の評価のあり方に関する一考察―スワッピングを中心に―」坂田幸繁編『公的統計情報―その利活用と展望』，中央大学出版部，39–62頁.

伊藤伸介（2020）「海外における公的統計ミクロデータと行政記録情報の利活用の動向について」中央大学経済研究所 Discussion Paper No. 331，1–23頁.

大屋祐雪（1995）『統計情報論』九州大学出版会.

工藤弘安（1986）「統計調査における情報提供（I）―諸概念の考察とその周辺―」『成城大學經濟研究』92号，73–95頁.

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」（2017）『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』.

個人情報保護委員会（2016）『個人情報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）』.

個人情報保護委員会事務局（2017）『個人情報保護委員会事務局レポート：匿名加工情報 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』.

寺田雅之・鈴木亮平・山口高康・本郷節之（2015）「大規模集計データへの差分プライバシーの適用」『情報処理学会論文誌』Vol. 56, No. 9, 1801–1816頁.

内閣官房 IT 戦略本部パーソナルデータに関する検討会（2013）『技術検討ワーキンググループ報告書』.

浜砂敬郎（2010）「2007年統計法といわゆる「基本計画」について」『経済学研究』（九州大学）第77巻第1号，27–44頁.

森博美（2007）「合衆国における人口センサスの新展開」法政大学日本統計研究所『研究所報』29–48頁.

森博美（2009a）「オランダの virtual census について」熊本学園大学『経済論集』第15巻第3・4号合併号，35–58頁.

森博美（2009b）「オランダの社会統計データベース SSD について」『経済志林』第76巻第4号，5–28頁.

Abowd, J. M.（2018）“Staring-Down the Database Reconstruction Theorem”, Presented at Joint Statistical Meetings, Vancouber, BC, Canada. https://www.census.gov/content/dam/Census/newsroom/press-kits/2018/jsm/jsm-presentation- database-reconstruction.pdf （2020年7月13日アクセス）

American Statistical Association（ASA）（1977）“Report of Ad Hoc Committee on Privacy and Confidentiality”, *The American Statistician*, Vol.31, Vol.2, pp.59–78, 法政大学日本統計研究所『統計研究参考資料』No. 4.

Brandt, M., L. Franconi, C. Guerke, A. Hundepool, M. Lucarelli, J. Mol, F. Ritchie, G. Seri, R. Welpton（2010）Guidelines for the Checking of Output Based on Microdata Research, Final Report of ESSnet Sub-Group on Output SDC, Eurostat.

Ito, S., M. Terada（2019）“The Potential of Anonymization Methods for Creating Detailed Geographical

Data in Japan", Paper presented at Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, The Hague, Netherlands, pp.1–14.

Lauger A., B. Wisniewski, L. McKenna（2014）"Disclosure Avoidance Techniques at the U.S. Census Bureau: Current Practices and Research", Research Report Series（Disclosure Avoidance #2014–02）, U.S. Census Bureau, pp.1–13.

Lebel, A. and J. Denis（2016）"Assessing the usability of a statistical population register for the Census of Population in Canada", Paper Presented at the United Nations Economic Commission for Europe, Conference of European Statisticians, Group of Experts on Population and Housing Censuses, Geneva, pp. 1–16.

Shlomo, N., C. Tudor, P. Groom（2010）"Data Swapping for Protecting Census Tables", J. Domingo-Ferrer and E. Magkos（eds.）Privacy in Statistical Databases UNESCO Chair in Data Privacy International Conference, PSD 2010 Corfu, Greece, September, 2010 Proceedings, Springer, pp. 41–51.

Wagner, I., D. Eckhoff（2015）"Technical Privacy Metrics: a Systematic Survey", https://arxiv.org/pdf/1512.00327.pdf（2020年７月13日アクセス）

Yancey, W. E., W. E. Winkler, R. H. Creecy（2002）"Disclosure Risk Assessment in Perturbative Microdata Protection", Research Report Series（Statistics #2002-01）, Statistical Research Division U.S. Bureau of the Census. https://www.census.gov/srd/papers/pdf/rrs2002-01.pdf （2020年７月13日アクセス）

Zayatz, L.（2007）"Disclosure Avoidance Practices and Research at the U.S. Census Bureau: An Update", *Journal of Official Statistics*, Vol. 23, No. 2, pp. 253–265.

（中央大学経済学部教授　博士（経済学））