

海外における公的大規模データの利活用の現状

伊藤 伸介

(中央大学教授)

本稿では、公的大規模データに関する提供形態とデータの提供にあたって適用されている秘匿措置の特徴を述べた上で、海外における公的大規模データの作成・提供の動向について明らかにした。また、イギリスを例に、個人情報保護される形でリンケージされたデータの作成の動向を概括し、わが国における公的大規模データの利活用をめぐる課題についても考察した。海外では、安全性・秘匿性および有用性のバランスを取りながら、各種の公的大規模データの提供を行っており、それは、個票データと秘匿処理済みマイクロデータに類別される。前者は、法制度や統計技術による秘匿措置、さらに組織的な対応によって、リモートアクセスを含む安全な分析環境を担保することで、リンケージデータのアクセスを含む利用者のさまざまなニーズに対応した利用サービスを可能している。一方で、後者に関しては、秘匿性の強度とデータに含まれる情報の粒度、データの利用目的に合わせて、学術研究用ファイル（Scientific Use File＝SUF）や一般公開型ファイル（Public Use File＝PUF）が作成・提供されている。わが国では、学術研究を指向した公的大規模データにおけるリンケージは、統計作成者側と利用者側との信頼関係を前提としつつ、統計法制度を踏まえて議論する必要がある。また、攪乱の手法の適用可能性を追究した上で、オンデマンド集計によるデータの提供や PUF の作成・公開のあり方を検討することも求められよう。

目次

- I はじめに——わが国における公的統計や行政記録情報の二次利用の現状
- II 公的大規模データの提供形態と秘匿措置に対する考え方について
- III 海外における公的大規模データの提供状況
- IV 公的機関によるデータリンケージの可能性
- V おわりに——わが国における公的大規模データの利活用の今後の課題

- I はじめに——わが国における公的統計や行政記録情報の二次利用の現状

わが国では、近年さまざまな経済政策の分野に

おいて、EBPM（＝Evidence Based Policy Making、客観的な事実に基づいて政策立案を行うこと）の推進の必要性が高まっている。そのために、公的統計データ、民間のビッグデータ、さらには行政記録情報の広範な利活用を可能にするための基盤整備が求められている。

わが国で EBPM の推進に向けたデータの整備に関する議論がなされたのは、2016 年 12 月に開催された経済財政諮問会議に遡ることができる。2017 年 5 月に刊行された『統計改革推進会議最終取りまとめ』（平成 29 年 5 月 19 日統計改革推進会議決定）（2017 年 5 月刊行）では、EBPM を推進する体制を構築するために、「統計を始めとする各種データ（統計、統計マイクロデータ及び統計的な

利活用を行うために用いられる行政記録情報が含まれる)」の整備・改善の必要性が議論された。そこで、「現行の統計関係法制について、総合的に見直しを行う」ことを目指して、特に、公的統計の調査票情報等の提供における法律面での課題について審議がなされた。その後、統計法の改正をめぐる議論がなされ、2018年には、「統計法及び独立行政法人統計センター法の一部を改正する法律」（以下「改正統計法」と呼称）が公布され、2019年に本格施行された。それによって、わが国では、公的統計のマイクロデータの利活用のさらなる展開が図られた（伊藤 2022）。

改正統計法が施行されて以降のわが国の公的統計のマイクロデータの特徴は、公表される統計表の元データとしての調査票情報（個票データ）が、法第33条（調査票情報の提供）に基づいて、磁気媒体の形で提供されるだけでなく、オンライン施設で調査票情報が利用可能になっていることである。なお、調査票情報は、2025年3月31日からリモートアクセス¹⁾による利用サービスも開始されていることは重要な動きであると言える²⁾。また、わが国では、法第36条（匿名データの提供）に基づいて、調査票情報に秘匿処理を行うことによって作成された匿名データが、総務省統計局所管の「就業構造基本調査」「労働力調査」等の8種類の基幹統計調査を対象に提供されているだけでなく、一般からの委託に応じて調査票情報から集計するオーダーメイド集計の仕組みも存在する（法第34条）。さらに、わが国では、公表された統計表の分布から乱数の生成等を施すことによって作成した擬似的なデータとして、「全国消費実態調査」「就業構造基本調査」および「国勢調査」の一般用マイクロデータが提供されており、それらは（独）統計センターのHPから入手することができる。

最近では、学術研究目的のために、行政記録情報の利活用を可能にする仕組みが存在する。例えば、国税庁は、わが国の税・財政施策の改善・充実に資する統計的研究という形で、2021年6月より国税庁保有行政記録情報（税務データ）を用いた税務大学校との共同研究を展開している。これについては、対象となる研究プロジェクトの

研究者が税務大学校の職員という国家公務員の身分を付与され、税務大学校の中の安全な分析環境の下で、税務データにアクセスすることが可能になっている。さらに国税庁は、2025年4月より、税務データの研究用匿名データの提供も開始している。

他方、欧米諸国においては、各国の法制度に基づきつつ、個体が特定されないための秘匿措置を施した上で、統計作成部局やデータ提供機関が、公的なマイクロデータの作成・提供を行ってきた。具体的には、データに対する安全性・秘匿性³⁾と利用者のさまざまなニーズ（有用性）を考慮し、利用者の利用目的も踏まえた形で、粒度（データの詳細さ）が異なる各種のマイクロデータの利用サービスが展開されてきた。また、学術研究目的のために、公的統計の個票データだけでなく、行政記録情報さらには医療健康データの提供が進められてきた（伊藤 2020a, 2020b；木村ほか 2020；佐々木・大寺・木村 2020；伊藤・木村・佐々木 2024）。さらに、社会人口の属性や経済的属性に関する公的統計のマイクロデータと医療・健康・福祉等の他分野のマイクロデータとのリンケージを行うことによって、広範な分野を対象にした実証研究が展開されている（伊藤・木村・佐々木 2024；伊藤 2025）。こうした公的統計マイクロデータや行政記録情報の個票データ（以下「公的大規模データ」と呼称）の提供に関する基盤整備の進展は、EBPMを指向するデータの利活用の動きと符合していると言える。

本稿では、最初に公的大規模データの提供形態を概括した上で、提供対象となる公的大規模データに対する秘匿処理の方法についてその特徴を論じる。つぎに、欧米諸国における公的大規模データの提供状況を明らかにする。さらに、イギリスの事例をもとに、近年注目が高まっている公的大規模データを対象にしたデータリンケージの方法について述べる。最後に、わが国における公的大規模データの利活用をめぐる将来的な課題を議論する。

Ⅱ 公的大規模データの提供形態と秘匿措置に対する考え方について

海外では、各種の公的大規模データは、利用者のニーズに応じる形で、さまざまなチャンネルによって提供されている。本節では、公的大規模データの提供形態についての特徴を概説した上で、データの提供にあたっての秘匿措置に関する考え方を整理する。

最初に、公的大規模データの提供形態に関しては、以下のように大別することが可能である（伊藤 2018, 2022；伊藤・木村・佐々木 2024）。

第1は、個票データ（非識別データ、deidentified data）の利用サービスである。個票データについては、学術研究目的で利用する研究者等を対象に、原データ（raw data, confidential data）から直接的な識別子を削除するが、さらなる秘匿処理が施されず、非識別化がなされたデータとして研究者等が利用することが可能である。個票データに関しては、安全な（secure）分析・利用環境の下での利用が求められている。それは、（1）統計作成部局やリサーチデータセンター（Research Data Centre=RDC）の内部におけるオンサイト施設での個票データのアクセスと（2）リモートエグゼキューション（remote execution）における利用サービスに類別される。後者についてはさらに、①大学・研究機関の研究室等からの個票データのリモートアクセス、②プログラム送付型のリモートエグゼキューション（利用者が分析用のプログラムコードを統計作成部局に送付し、統計作成部局から分析結果が提供されること）、および③オンデマンド集計（ウェブサイト上で利用者が変数を選択することによって集計（分析）結果を取得すること）が含まれる。

オンサイト施設での個票データのアクセスでは、インターネットとの接続ができないだけでなく、オンサイト施設内にカメラが設置され、個票データの利用に対する不正がないかを監視することによって、その安全性が物理的に担保されている。その一方で、近年では、リモートアクセスシステムによる個票データの利用サービスが広範に

普及することで、オンサイト施設の運営が停止される国も存在する。

個票データには直接的な個体識別子の削除以外の秘匿処理が行われていないことから、一般には機微な個体情報が含まれている。それゆえ、個票データのアクセスにあたっては、利用資格に関する法制度的な措置が取られることが少なくない。具体的には、法律に基づき、利用者に対してアクセスの資格を与える（「承認された研究者（approved researcher）」ないしは「認可された研究者（accredited researcher）」）、あるいは研究者が所属する大学や研究機関が、「認可された機関（authorized institution, recognized organization）」であることを要件とする仕組みが、海外では整備されている⁴⁾。これらの法制度的な仕組みは、データの提供者と利用者との信頼関係を保証するために制度化されたものと考えられることができる。

第2は、秘匿処理済みのマイクロデータの提供である。これに関しても、（1）匿名化マイクロデータ（個票データに秘匿処理が施されたデータ、anonymized microdata）の提供、および（2）一般公開型マイクロデータ（個体情報に関する秘匿性の程度を相対的に高めることによって作成された一般公開用のデータ、public use microdata）の作成・公開に大別することができる。前者については、学術研究目的のために作成され、利用申請が必要な学術研究用ファイル（Scientific Use File=SUF）が知られている。後者に関しては、一般には利用目的の制限がなく、誰でも統計作成部局等のウェブサイトから入手可能であるが、教育目的やプログラム作成のための利用が想定されている一般公開型ファイル（Public Use File=PUF）が該当する。

匿名化マイクロデータの作成にあたっては、統計作成部局は、統計調査の調査設計やデータ特性を踏まえつつ、調査票に含まれる調査事項から把握される調査客体の外観識別性や個体情報の特定化を可能にする外部情報の入手可能性も検討した上で、個票データに匿名化技法を適用する。具体的には、2つの方法が考えられる。第1は、非攪乱的手法（non-perturbative method）であって、レコードあるいは属性の削除（suppression）、量的属性を対象に分布における上位（あるいは下位）

の属性値を統合するトップ・コーディング (top coding) (ボトム・コーディング (bottom coding)), 質的属性を対象にカテゴリーを統合するリコーディング (recoding, 区分統合), 対象となるレコード群の中からランダムに一部抽出を行うサンプリング (sampling) (標本調査データの場合はリサンプリング, resampling) 等が該当する。第2は, 攪乱的手法 (perturbative method) であって, これは, ノイズの付加 (ノイズ付与, noise addition), 地域間で類似するレコードの入れ替え (スワッピング, data swapping) を行う等の攪乱 (perturbation) を施すことを意味する。

統計作成部局は, 個人情報特定されるリスクやそれに伴うセンシティブな属性値の露見リスク (disclosure risk) をコントロールすること (統計的開示制御, Statistical Disclosure Control) を指向する形で, 非攪乱的手法と攪乱の手法のメリットとデメリットも勘案しつつ, 適用可能な秘匿処理の方法を選定する。そのために, 秘匿処理済みのマイクロデータを対象に, 秘匿性 (あるいは安全性) と有用性についての定量的な評価も行われてきた。この場合の定量的な評価値は, 統計作成部局が秘匿処理の方法を採用する上での参考情報として位置付けられる (伊藤 2022)。

ところで, 公的大規模データの提供に関する安全性の考え方については, イギリス国家統計局 (Office for National Statistics=ONS) によって考案された法制度的・統計技術なフレームワークである「5つの安全モデル (Five Safes Model)」が, ヨーロッパ諸国だけでなく, カナダやオーストラリアでも適用されていることが知られている。最近ではアメリカにおける公的大規模データの利活用推進のためのモデルとしても注目されている⁵⁾。

5つの安全モデルは, 以下の5つの基準から構成される。すなわちそれは, (1) 妥当な統計目的に対して個票データのアクセスを行うためのプロジェクトであることを示す安全なプロジェクト (safe projects), (2) 個票データを適切に利用することに関して信頼されている安全な利用者 (safe people), (3) 個人情報露見しないデータを表す安全なデータ (safe data), (4) 技術的な管理措置によって, 容認されていない個票データのア

クセスが回避可能な安全な施設 (safe settings), および (5) 個人情報露見する結果が含まれない安全な分析結果 (safe outputs) である (Ritchie 2008; Desai, Ritchie and Welpton 2016; 伊藤 2016, 2020a; 伊藤・木村・佐々木 2024)。

5つの安全モデルの考え方では, 上記の5つの基準の組み合わせによって安全性の強度を高めることで, 「安全な統計 (safe statistics)」が作成される。各種の公的大規模データの提供形態についても5つの安全モデルで説明することが可能である。例えば, 個票データの利用サービスにおいては, 安全なプロジェクト, 安全な利用者, 安全な施設と安全な分析結果が重視される。個票データから直接識別子のみが削除された非識別データが, オンサイト施設やリモートアクセスのような安全な施設の中でのみ利用可能になることによって, 安全な統計の作成が担保される。それに対して, 秘匿処理済みのマイクロデータに関しては, 安全なデータであることが重視されており, そのために各種の秘匿処理の方法が用いられる。この場合, 安全な施設や安全な分析結果は, 安全な統計を作成するための要件にはなっていない (伊藤・木村・佐々木 2024)。なお, 前者の個票データの利用サービスに関しては, 5つの安全モデルに基づき, 利用者にとっての法制度面, 組織面および統計技術面で安全な分析環境が担保されている。このような環境は, 信頼される研究環境 (Trusted Research Environment) と呼ばれている (伊藤 2025)。

公的大規模データを対象にした秘匿処理に関する近年の興味深い動きは, 情報工学の分野で発展してきた差分プライバシー (differential privacy) (Dwork 2006) の方法論が, 公的統計の分野でもその有効性をめぐって議論されていることと, アメリカ人口センサスにおいて差分プライバシーの実現方式が適用されたことにある。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(寺田ほか 2015: 1803)。

アメリカセンサス局は、公表されている統計データから攻撃者が複数の集計表を組み合わせて元となる個票データに含まれる個人情報暴露する、再構築攻撃(reconstruction attack)(Dinur and Nissim 2003; 伊藤ほか 2022)に基づくリスクを回避するために、差分プライバシーの方法論の公的統計への適用可能性に関する実証実験を行った上で、2020年センサスに対する差分プライバシーの実現方式を採用した。具体的には、アメリカセンサス局は、差分プライバシーが与える安全性を保証するためのパラメータであり、統計表の公表によって消費されるプライバシー損失予算(privacy loss budget) ϵ を設定し、地域レベルで階層的に ϵ を割り当て、それにしたがってノイズが付与された統計表として、2021年に区画改定データ(PL94-I71)を作成・公表した。さらには、差分プライベートな一般公開型マイクロデータであるプライバシー保護済マイクロデータファイル(Privacy-Protected Microdata Files=PPMFs)(伊藤・寺田・加藤 2024)も 2024年に公開された⁶⁾。

その後、2023年5月~2024年9月に、年齢、性別、人種等の詳細な人口社会的な特性値と、住宅の所有関係(housing occupancy)、居住歴(housing tenure)を含む住宅に関する特性値を対象にした集計データとして、複数の人口・住宅ファイル(Demographic and Housing Characteristics File=DHC)が公表されている。これらのいずれに関しても、差分プライバシーの方法論が適用されている。このことから、アメリカセンサス局は、2020年のセンサスデータの提供において、差分プライバシーの実現方式を積極的に採り入れてきたことが窺える。

ところで、秘匿処理済みのマイクロデータに対する有用性については、主として①記述統計量やクロス表等を用いた元データとの分布特性の差異の把握、および②情報量損失(information loss)に関する指標を用いて、定量的な評価を行うことができる(伊藤 2019)。差分プライバシーを例にすると、前者に関しては、差分プライバシーの実現方式が適用されたデータの元データからの分布特性の差異を有用性に関する指標として設定した上

で、平均絶対誤差(MAE)、2乗平均平方根誤差(RMSE)等、セル単位での平均誤差を用いた評価がアメリカセンサス局によって行われたことを指摘することができる。そして、2010年センサスを用いて作成したPPMFsの有用性の評価結果が公開されることによって、差分プライベートなセンサスデータの有用性を利用者側が確認・検証することを可能にした。なお、後者については、スワッピングや差分プライバシーの実現方式によって攪乱が施されたデータの元データに対する有用性に関する評価指標として、変数のヒストグラムを対象に平滑化されたカルバック・ライブラー情報量(μ -smoothed Kullback-Leibler divergence)を用いて検証を行った Christ, Radway and Bellovin (2022)等の事例がある。

こうした中で、人口センサスの利用者側(例えば、ミネソタ人口研究センターの Steven Ruggles 教授等)より、差分プライベートなセンサスデータの有用性に関する疑問が呈された(Ruggles et al. 2019)。その一方で、統計作成部局側として、Hawes (2020)は、差分プライバシーの方法論が適用されることで、プライバシーとデータの精度に関する定量的な関係性が明示され、データの有用性とプライバシー保護に関するオープンな議論が可能になったことを指摘している(伊藤ほか 2022)。

2020年センサスに対する差分プライバシーの方法論の採用に関しては、例えば、Muralidhar and Ruggles (2024)が、公表されている差分プライベートなセンサスデータに対しては、再構築攻撃が可能であり、その結果として、個人情報識別リスクに対して脆弱であると批判している。こうしたこともあり、Hawes et al. (2025)は、露見回避(disclosure avoidance)のために求められることとして、「公的統計の三重のトレードオフ(triple tradeoff of official statistics)」(Abowd and Hawes 2023)、すなわち公的統計の秘匿性(confidentiality)、精度(accuracy)と利用可能性(availability)と間のバランスの必要性を指摘している。このことは、差分プライバシーの実現方式を含む秘匿処理の実施によって公的統計データの安全性・秘匿性が担保されたとしても、提供される公的統計

データの粒度，さらには利用者にとって実証研究等で利用可能となるかという有用性の視点を持つことの重要性が，統計作成部局側に求められていることを示唆している。

Ⅲ 海外における公的大規模データの提供状況

本節では，海外で展開されている公的大規模データの作成・提供の現状について見ていくことにしたい。

表1は，海外の主な統計作成部局における公的大規模データの提供形態を整理したものである。国によって統計作成部局やデータ提供機関による公的大規模データの提供状況は異なることがわかる。

例えばフランスでは，INSEE（フランス国立統計経済研究所）が，SUFとPUFを作成しているだけでなく，学術研究のために個票データをリモートアクセスシステムによってのみCASD（＝The Secure Data Access Centre）を通じて提供している。ドイツでは，ドイツ連邦統計局（Federal

Statistical Office）が，匿名化されたマイクロデータとしてSUFとPUFを作成・提供しているだけでなく，個票データが，オンサイト施設を備えたリサーチデータセンター（Research Data Centre＝RDC）内部で利用可能になっている。カナダでは，カナダ統計局が一般公開型マイクロデータファイル（Public Use Microdata Files）を作成・公開するだけでなく，RDC内のオンサイト施設による個票データの利用サービスが，主な提供のチャネルとして展開されている。さらに，カナダ統計局は学術目的のための行政記録情報の二次利用システムであるSDLE（＝Social Data Linkage Environment）の仕組みも備えている（伊藤 2021）。オーストラリアの場合，オーストラリア統計局が，CURF（＝Confidentialised Unit Record File）と呼ばれる匿名化マイクロデータを作成・提供しているだけでなく，リモートアクセスによる個票データの提供，さらにはTableBuilderというオンデマンド集計システムによる利用者のニーズに応じた集計表の提供サービスを行っている（伊藤・谷道・小島 2018）。また，ニュージーランドも，PUFの公開は行われていないが，ニュージーラ

表1 海外の主な統計作成部局における公的大規模データの提供形態の一覧

データの種類／提供形態	個票データ（非識別データ）のオンサイト施設による提供	個票データ（非識別データ）の磁気媒体による提供	個票データ（非識別データ）のリモートアクセスによる提供	プログラム送付型のリモートエクゼキューションによる提供	オンデマンドシステムによる提供	匿名化マイクロデータの提供	一般公開型マイクロデータによる提供
各国の統計作成部局							
欧州連合統計局（Eurostat）	○		○			○	○
イギリス国家統計局	○		○		○	○（UKDSから提供）	○
ドイツ連邦統計局	○		●（SUFとして提供）	○	○	○	○
フランス INSEE	○		○			○	○
オランダ統計局			○	△	○	○	○
デンマーク統計局			○		○		
フィンランド統計局	△		○			○	
アメリカセンサス局	○		○		○		○
カナダ統計局	○		●	○	○		○
オーストラリア統計局			○		○	○	
ニュージーランド統計局	○					○	
ノルウェー統計局	○	○		○			○（NSDから提供）
総務省統計局	○	○	○			○	○

注：UKDSとはイギリスエセックス大学のUK Data Service，NSDとはNorwegian Centre for Research Dataの略である。

○……運営されている。●……計画中である。△……運営中であるが，活動を休止している（2023年12月時点の情報をもとに作成。一部の国々については2025年4月時点の情報に更新）。

出所：伊藤・木村・佐々木（2024），伊藤（2024）における表1を一部加筆・修正

ンド統計局が、CURFを作成・提供しているだけでなく、統計局内のオンサイト施設やリサーチデータセンター内の安全な分析環境の下で個票データの利用サービスが行われている。

表1を見ると、統計作成部局がリモートアクセスによる個票データの利用サービスを行っている国々が少なくないことが確認できる。デンマーク、フィンランドといった北欧諸国、イギリス、フランス、オランダ等で、個票データのリモートアクセスによる利用サービスが展開されている。また、デンマーク、フランス、オランダとオーストラリアでは、リモートアクセスでのみ個票データの提供を行っている点は興味深い（伊藤 2024）。さらに、ドイツやカナダにおいては、リモートアクセスによる個票データ提供の仕組みは存在しないが（2025年3月時点）、ドイツ連邦統計局は、現在作成・提供されているSUFよりも粒度が細かいSUFを対象にしたリモートアクセスを計画している（Brenzel et al. 2023）。カナダ統計局は、CRDCN（=Canadian Research Data Centre Network）との共同開発によって、リモートアクセスシステムであるvRDC（=virtual Research Data Centre）の開設の準備を進めている（Cranswick 2019）^{7）}。

つぎに、イギリス、アメリカとデンマークを例に、公的大規模データの提供状況について具体的に見ていくことにしたい。

1 イギリス

イギリスでは、個票データの利用サービスが広範に展開されている。また、公的統計や社会調査データを対象にした匿名化マイクロデータの作成・提供と人口センサスの一般公開型マイクロデータの作成も進められてきた（伊藤 2020a；伊藤・木村・佐々木 2024）。

イギリス国家統計局（ONS）には、Secure Research Service（SRS）と呼ばれる、個票データを利用するためのオンサイト施設を管理・運営する部局が存在する。イギリス統計法（The Statistics and Registration Service Act 2007）およびDigital Economy Actに基づき（伊藤 2020a）、研究者はSRSを通じて、個票データにアクセスすることが可能になっている。また、SRSはリ

モートアクセス施設も運営しており、近年では、ONSはリモートアクセスによる個票データの提供を積極的に行っている（伊藤・木村・佐々木 2024）。そのため、研究者は、大学・研究機関からの研究室からSRSのサーバに対して個票データをリモートでアクセスすることもできる。なお、リモートアクセスについては、ONSのSRSだけでなく、エセックス大学に付置されたデータサービス提供センターであるUK Data Service（UKDS）が管理・運営するthe Secure Labも利用可能である。

また、UKDSでは、公的統計や社会調査に関する匿名化マイクロデータを提供している。これは、安全な利用に関する誓約書を提出し、ライセンス（End User Licence）を取得した上で、IDとパスワードを入力することによって、UKDSのウェブサイトですぐ入手可能な「ライセンス型マイクロデータ」である。

ライセンス型マイクロデータが作成される公的統計については、人口センサスだけでなく、「労働力調査」のような標本調査データ（サーベイマイクロデータ）が含まれる。人口センサスに関しては、1970年代からマイクロデータの提供に対する研究者のニーズは存在したものの、匿名化マイクロデータという形で作成・提供されたのは、1991年センサス以降である（伊藤 2011）。イギリスの場合、人口センサスは、利用者のさまざまなニーズに応える形で、個人単位で抽出された個人ファイルと世帯単位で抽出された世帯ファイルという属性の粒度が異なる複数のファイルが作成されている。例えば、1991年人口センサスでは、世帯と個人を抽出対象とする2種類のSARs（=Sample of Anonymised Records）が提供されている。また、2001年センサスでは、地域区分が詳細なマイクロデータの利用のニーズに応じる形で、SAM（=Small Area Microdata）が作成・提供されている。2021年センサスでも、個人単位で抽出されたSafeguarded Microdata Sampleだけでなく、Safeguarded Household Sampleと呼ばれる世帯単位で抽出され、階層構造を持つライセンス型マイクロデータが作成されている。また、一般公開型マイクロデータとしては、2011年センサスに関す

る教育用のマイクロデータが、ONS のウェブサイト
で公開されているだけでなく、2021 年センサスに
関しても、Public Microdata Teaching Sample と
して作成・公開がなされている（伊藤・寺田 2023）。

なお、ONS は、2021 年人口センサスを対象
に、攪乱的手法の 1 つで、個票データから作成し
た集計表の全てのセルにノイズを付与する cell
key method に基づくオンデマンド型公表システ
ム（Flexible Dissemination System）を開発した⁸⁾。
ONS は、オーストラリア統計局の TableBuilder
を参考にした上で、「Create a custom dataset」
と呼ばれるウェブサイト上で、人口センサスにお
いて把握される変数群の中から、利用者が選択し
た変数やそれに含まれるカテゴリーの数に関する
利用者の要求に応じて多次元集計表を作成し、オ
ンデマンドで利用者に提供している。

2 アメリカ

アメリカでは、人口センサスの PUMS や経常人
口調査（Current Population Survey）の PUF がア
メリカセンサス局のウェブサイトで、無料でダウ
ンロードすることができる。さらに、アメリカセ
ンサス局の経済研究センター（Center for Economic
Studies）が管轄し、全国 35 カ所の大学等に設置
されている連邦統計リサーチデータセンター
（Federal Statistical Research Data Centers）（2025
年 4 月時点）の内部にあるオンサイト施設におい
て、世帯・人口系および事業所・企業系の個票
データのアクセスも可能である。そのために、ア
メリカセンサス局は、利用者が宣誓職員（Special
Sworn Status）の資格を有することを個票データ
の利用の要件としている（伊藤 2018）。また、
ResearchDataGov が個票データの検索システム
である SAP（Standard Application Process）Portal
を運営していることから、利用者は、16 の統計
作成部局が管理する個票データの検索を行うこと
ができる。なお、リサーチデータセンターでの個
票データの利用経験がある研究者に関しては、リ
モートアクセスによる個票データの利用が可能に
なっている（2025 年 4 月時点）。

アメリカにおけるマイクロデータの提供に関する
法的な規定は、アメリカセンサス法（Title 13）に

基づいている（石田 1999；伊藤 2018）。アメリカセ
ンサス局が、人口センサスのマイクロデータの公開
を開始したのは、1960 年センサスに遡ることが
できる。1962 年に、センサスの詳細（long form）
調査票の対象世帯（全世帯の約 1/4）に対してサ
ンプリングが施されるだけでなく、直接的な識別
子の削除や人口規模が 25 万人以上の地域区分の
設定によって、全人口の 0.1% のレコードが含ま
れるセンサス最初の一般公開型マイクロデータサン
プル（Public Use Microdata Sample=PUMS）が作
成・提供された（McKenna 2019）。その後、1970
年から 2010 年までの人口センサスにおいては、
PUMS の作成のために、サンプリング、トップ・
コーディング、リコーディングといった各種の非
攪乱的手法が適用された。また、2000 年以降の
人口センサスでは、PUMS の公開に対して攪乱
的手法が新たに用いられた。具体的には、2000
年と 2010 年の人口センサスでは、非攪乱の手
法だけでなく、公表統計表の元になる個票データ
にスワッピングやノイズ付与といった攪乱の手
法を施すことによって、PUMS が作成されている。
それに対して、2020 年の人口センサスにおい
ては、スワッピング技法では、再構築攻撃への対
応が困難であるという理由から、アメリカセンサ
ス局内部に設置された DSEP（=Data Stewardship
Executive Policy Committee）が差分プライバシー
の方法論の実現可能性を追究した。その結果とし
て、II で述べたように、2020 年センサスにおい
ては、秘匿処理のための新たな方法論として、差
分プライバシーの実現方式が採用された PPMFs
が、アメリカセンサス局のウェブサイトで取得可
能となっている（伊藤ほか 2022；伊藤・寺田・加
藤 2024）⁹⁾。

3 デンマーク

デンマークでは、1988 年にデンマーク統計局
内に、学術研究目的に対するマイクロデータの利
用を可能にするためのオンサイト施設が設置され
た後に、2001 年より、デンマーク統計局がリモ
ートアクセスによる利用サービスを行っている。オ
ンサイト施設が 2008 年に閉鎖されてからは、デ
ンマーク統計局は、リモートアクセスによる公的

大規模データの利用サービスのさらなる展開を図ってきた（伊藤 2020b, 2024）。

個票データの利用申請が行われると、デンマーク統計局は、「知る必要性（need to know）」原則に基づき（伊藤 2020a, 2020b）、人口社会情報、労働・所得情報、教育に関する情報、企業経済情報、医療健康情報等、各種のレジスターで管理されている行政記録情報から、利用者の分析目的に対して必要な属性群のみを選定し、各種の行政記録情報に付与されている共通 ID を仮名化し、さらに仮名化された ID に基づきデータリンケージを行った上で、利用者がアクセス可能なデンマーク統計局内のサーバにリンケージデータを移送する。それによって、利用者は、個票データの申請書類に記載した使用予定の変数群を含むリンケージデータをリモートでアクセスすることができる。なお、デンマーク統計局は、個票データの利用者が行う実証研究が、社会に還元する、公共の利益に資する研究であることを要求しているが、近年ではマーケティングリサーチのための個票データの利用も、社会に還元する研究という条件のもとで容認している（伊藤 2024）。なお、デンマークでは、デンマークデータ保護法、1985 年行政機関法（Public Administration Act 1985）およびデンマーク刑法（Danish Penal Code）が、デンマークにおける行政記録情報の二次利用さらには秘密保護に関する法的根拠となっている。

リモートアクセスによって得られた分析結果のチェックに関しては、デンマーク統計局が管理・運営しているスキャンシステム上で研究者自身が分析結果に関する自動的なチェックを行う仕組みが備わっており、そのチェック済みの分析結果を研究者はサーバから外部へ持ち出すことが可能になる。なお、デンマーク統計局の職員が、事後的に分析結果の一部をランダムに選定し、マニュアルでチェックを行うことが知られている（伊藤 2020b）。

こうしたリモートアクセスによる各種のリンケージデータの提供とその利活用がデンマークで広範に展開されたことについては、デンマーク統計局がデータの提供者と利用者との間の信頼関係を重視している点を強調しておきたい。

IV 公的機関によるデータリンケージの可能性

海外のいくつかの国においては、学術研究のために、統計作成部局あるいはデータ提供機関によってさまざまな社会人口の属性や経済的属性を含む公的大規模データ間のリンケージが可能になっている。このことは、実証研究の対象範囲を拡大させることから、EBPM の推進に寄与することができる。

デンマーク等の北欧諸国では、個人に付与されている共通 ID（個体識別番号）に基づいて、公的大規模データ間のリンケージが可能である。それに対して、イギリス、アメリカ、カナダ、オーストラリア、ニュージーランド等、個人に関する共通 ID の仕組みは存在しないが、リンケージ用の ID が生成され、それを用いて公的大規模データのリンケージが行われている国々が存在する。例えばオーストラリアとニュージーランドでは、公的大規模データや医療健康データの二次利用促進のために、オーストラリア統計局とニュージーランド統計局がそれぞれ、データリンケージを行っている。それによって、学術研究目的に対するリンケージデータ（統合データ integrated data）の提供が広範に展開されている（伊藤 2025）。

イギリスについて見ていくと、2018 年に Digital Economy Act 2017（DEA）が施行されることによって、公的大規模データの中でも、特に行政記録情報の個票データ同士のリンケージを法制度面で可能にする仕組みが整備されてきた。さらに、社会経済研究協議会（Economic and Social Research Council=ESRC）が実施している ADR UK と呼ばれるプロジェクトの資金提供が、とりわけ公的大規模データのリンケージに基づくデータの利用サービスに寄与してきた（伊藤 2020a）。

イギリスの場合、ONS が行政記録情報を含む公的大規模データのリンケージ業務を担っている。ONS における公的大規模データのリンケージの手順は、以下の①～④にしたがって進められる（伊藤 2025）。

- ①ONS 内 部 の Data Access Platform (DAP) という部署のデータリンケージチームが、リンケージの対象となる複数の公的大規模データから名前や住所といった直接的な個体識別子 (Direct Identifier) のみを切り離す。この情報は、DAP で個体識別子の仮名化に直接携わっている職員以外は閲覧することができない。
- ②データリンケージチームの中の仮名化担当チーム (indexing team) が、個体識別子の仮名化を行った上で、リンケージ用の仮名化 ID を生成する。この仮名化 ID の生成に関する作業においては、異種の公的大規模データにある仮名化 ID 同士で、個体識別子によるマッチングを行った後に、作成されたリンケージ用の仮名化 ID の対応表のみが保存される。
- ③データリンケージチームの中のマッチング担当チーム (matching team) が、リンケージ用の仮名化 ID の対応表をもとに、複数の公的大規模データ間のリンケージを実行する。
- ④データリンケージが行われた後に、リンケージ用の仮名化 ID が削除された (非識別化された) 個票データは、リンケージデータとして SRS のサーバ上でアクセスすることができる。そのために、利用者は、DEA に基づき、認可された研究者の資格取得の手続きを行い、対象となる研究プロジェクトが UK Statistics Authority の研究認定委員会 (Research Accreditation Panel) の審査で容認されることが求められる。

ONS では、個体識別子に関する情報と属性情報を切り離した上で、個体識別子の仮名化によってリンケージ用 ID を生成するチームとリンケージ用 ID で属性情報を含む複数の公的大規模データ間のリンケージを行うチームに業務が組織的に分化している。それによって、仮名化担当チームは属性情報を確認することはできず、マッチング担当チームは、仮名化 ID から個々のレコードにおける個体識別子の対応づけを行うことが困難になっている。こうした方法を採用することによっ

て、公的大規模データにおける個体情報の安全性が、統計組織のかつ技術的に担保されている。

V おわりに——わが国における公的大規模データの利活用の今後の課題

本稿では、公的大規模データを対象に、その提供状況とデータの提供にあたって実施されている秘匿処理の特徴を明らかにした上で、海外におけるリンケージデータを含む公的大規模データの作成・提供の動向について述べた。最後に、こうした海外における公的大規模データの利活用の現状を踏まえて、わが国における今後の課題を論じることしたい。

本稿で見てきたように、海外においては、安全性・秘匿性および有用性のバランスを勘案した上で、統計作成部局が、公的大規模データの提供をどのように進めているかが論点となっている。その上で、個票データと秘匿処理済みマイクロデータへの対応を見ていくと、前者は、統計作成部局と利用者との間の信頼関係を前提とした上で、安全な分析環境を法制度的組織的技術的に担保することによって、学術研究を指向した利用者のさまざまなニーズへの対応を可能にしていることから、利用者は、各種の公的大規模データの利用サービスを楽しむことができる。その中には、リンケージされた個票データの利活用も含まれる。それに対して、わが国の統計法では、調査票情報の対象はあくまで統計調査であるから、行政記録情報は調査票情報の中に含まれないことが考えられる。こうした点を踏まえると、わが国における統計法の下での公的統計の二次利用の促進という観点から見た場合、公的統計マイクロデータや行政記録情報におけるリンケージに関する議論が法制度面で十分に展開されてきたとは言いがたい。これに関しては、広範な分野における学術研究を可能にするために、データリンケージにおいて求められる ID の仮名化に関する手続きをどのように整備するかという論点もあるだろう。このように公的大規模データの利用可能性については、法制度と統計技術の両面から具体的に検討することが求められよう。

その一方で、海外では個票データの提供形態の1つとして、オンデマンドで多次元集計表の作成・提供が行われている国々が存在するが、わが国で運用されているe-Statは、オープンデータである公表統計表の利用サービスの拡充に寄与しているものの、それは、オンデマンドで集計表を作成・提供するシステムとは異なる。こうしたオンデマンド型の集計システムをわが国で展開しようとするれば、cell key methodといったノイズが付与された公的統計データの提供可能性について実証的に追究することが必要である。

他方、後者の秘匿処理済みマイクロデータに関しては、海外の一部の国では秘匿性の強度が高い一般公開型マイクロデータであるPUFが作成・公開されている。それに対して、わが国の公的統計では、PUFという形での秘匿処理済みマイクロデータの提供はなされていない。学術研究のために実証分析を行う上でのプログラムコードのチェック等、個票データを利用する前段階としてのPUFの公開に対する利用者のニーズは小さくないと言える。また、個票データの利用にあたってのテストデータとして、合成データ(synthetic data)の作成の方向もありうる。合成データも、わが国の統計法制度において明確には位置付けられていないことから、海外の合成データの作成に関する研究、およびその提供状況を見据えることが求められる。こうした状況を踏まえると、差分プライバシーの実現方式の適用可能性も探究しながら、攪乱的手法の実用性に関する検討が必要となるだろう。

- 1) 本稿におけるリモートアクセス(remote access)とは、大学の研究室等の研究を行う場所で使用するPCを通じて、個票データが保管されているサーバにリモートでアクセスすることが可能な仕組みを意味している(伊藤 2020b)。
- 2) 2025年3月31日より、調査票情報の利用申出の窓口が、マイクロデータ利用ポータルサイト(miripo)に一元化されたことも、公的統計の二次利用における大きな進展だと言えよう。
- 3) 本稿における安全性とは、法制度的・組織的・技術的な安全管理措置が施されることによって、個人情報に関する漏洩のリスクから保護されている状態を表す。また、秘匿性については、個人情報に対する秘匿処理に基づく露見リスクの低減の程度が定量的に評価される。
- 4) ヨーロッパにおいては、認可された研究者に個票データのアクセスの資格を付与するよりも、むしろ認可された機関に個票データのアクセス権限を与えて、所属する研究者が個票データの利用が可能になるケースが少なくない。前者についてはイ

ギリスが該当し、後者に関しては、デンマークや欧州連合統計局(Eurostat)等が該当する(伊藤 2020b; 伊藤・木村・佐々木 2024)。

- 5) アメリカ予算管理局(Office of Management and Budget)のアメリカ連邦統計方法協議会(Federal Committee on Statistical Methodology)は、公的大規模データの保存・共有およびデータの利活用に関する方法論として、階層的なデータ提供モデル(Tiered Access Models)を提唱しているが、その基盤となるモデルとして、5つの安全モデルが採用されている点は興味深い。詳細については、下記のURLを参照。Data Protection Toolkit Tiered Access Models <https://nces.ed.gov/fcsm/dpt/content/4> [2025年4月27日アクセス]
- 6) 以下のURLを参照。<https://www.census.gov/newsroom/press-releases/2024/2020-microdata.html> [2025年4月27日アクセス]
- 7) カナダのvRDCについては、下記のURLを参照されたい。<https://crdcn.ca/virtual-research-data-centre-vrdc/> [2025年4月27日アクセス]
- 8) ONSが開発したオンデマンド集計システムの特徴については、伊藤・寺田(2023)を参照。
- 9) 差分プライバシーの実現方式が採用された2020年センサスのPPMFsは、ノイズ付与等と同様に、元データからの変数の攪乱が実行されたという意味では、事実上合成データ(synthetic data)と同じタイプのデータと位置付けられる(伊藤・寺田・加藤 2024)。なお、合成データとは、「元になるデータからその分布特性が近似するように属性値を新たに生成することによって作成され、個人情報秘匿性が確保されたマイクロレベルの擬似的なデータ(Templ 2017: 157)と定義されている。なお、わが国の公的統計の個票データを用いた合成データの生成技法に関する実証研究については、例えば伊藤・横溝(2024)を参照。

参考文献

- 石田晃(1999)「アメリカ、カナダにおけるマイクロデータの現状について」『研究所報』No. 25, pp. 1-34.
- 伊藤伸介(2011)「わが国におけるマイクロデータの新たな展開可能性について——イギリスにおける地域分析用マイクロデータを例に」『明海大学経済学論集』Vol. 23, No. 3, pp. 36-54.
- (2016)「諸外国における政府統計マイ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, pp. 233-249.
- (2018)「公的統計マイ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, pp. 77-101.
- (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, pp. 39-62.
- (2020a)「諸外国における公的統計と行政記録データの二次利用に関する展開方向」『経済学論纂(中央大学)』第61巻第2号, pp. 1-16.
- (2020b)「デンマークとオランダにおける医療健康データの二次利用について」『日本統計学会誌』第50巻第1号, pp. 109-138.
- (2021)「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」『経済学論纂(中央大学)』第61巻第3・4合併号, pp. 235-251.
- (2022)「マイクロデータの匿名化と統計情報の秘匿可能性について」『経済学論纂(中央大学)』第63巻1・2合併号, pp. 1-23.

- (2024)「海外におけるマイクロデータの利活用推進に向けた取り組み状況——リモートアクセスを中心に」『経済学論纂 (中央大学)』第 65 巻 1 号, pp. 199-218.
- (2025)「海外における公的大規模データのリンケージの動向——イギリスを例に」『経済学論纂 (中央大学)』第 65 巻 5・6 合併号, pp. 187-200.
- 伊藤伸介・木村映善・佐々木香織 (2024)「イギリスにおける大規模データの二次利用の新たな展開——スコットランドの事例を中心に」『統計研究彙報』第 81 号, pp. 89-108.
- 伊藤伸介・谷道正太郎・小島健一 (2018)「オーストラリアにおける公的統計の二次利用について——オンデマンド集計システム TableBuilder を中心に」『経済学論纂 (中央大学)』第 58 巻第 2 号, pp. 187-208.
- 伊藤伸介・寺田雅之 (2023)「海外における公的統計に関するプライバシー保護の現状——アメリカとイギリスの事例をもとに」『統計研究彙報』第 80 号, pp. 117-136.
- 伊藤伸介・寺田雅之・赤塚裕人・北井宏昌 (2022)「海外における公的統計に対する攪乱的手法の新たな取り組み——アメリカセンサス局による差分プライバシーの適用を中心に」『統計研究彙報』第 79 号, pp. 131-150.
- 伊藤伸介・寺田雅之・加藤駿典 (2024)「公的統計に対する差分プライバシーの適用と有効性の評価に関する検討——国勢調査を例に」『統計研究彙報』第 81 号, pp. 69-88.
- 伊藤伸介・横溝秀始 (2024)「事業所・企業系の公的統計を対象にした合成データの生成技法に関する検討——経済センサスを例として」『統計数理』第 72 巻第 2 号, pp. 217-231.
- 木村映善・大寺祥佑・佐々木香織・黒田知宏 (2020)「フィンランドにおける医療分野レジスタとデータ提供の状況」『日本統計学会誌』第 50 巻第 1 号, pp. 47-80.
- 佐々木香織・大寺祥佑・木村映善 (2020)「より包括的で正確な医療統計を可能とする社会・制度基盤に向けた一考察——イギリスの England における医療情報二次利用に関する調査・事例研究から」『日本統計学会誌』第 50 巻第 1 号, pp. 81-108.
- 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』第 56 巻第 9 号, pp. 1801-1816.
- 統計改革推進会議 (2017)『統計改革推進会議最終取りまとめ』.
- Abowd, J. M. and M. B. Hawes (2023) “21st Century Statistical Disclosure Limitation: Motivations and Challenges,” Working Paper ced-wp-2023-002. <https://www.census.gov/library/working-papers/2023/adrm/ced-wp-2023-002.html> [2025 年 4 月 17 日アクセス]
- Brenzel, H., Cramer, K., Güttgemanns, V. and Mathes, M. (2023) “Remote Access for Scientific Use Files – A New Pathway for German Official Statistics Microdata Access,” Paper presented at UNECE Expert Meeting on Statistical Data Confidentiality 2023, pp. 1-6.
- Christ, M., Radway, S. and Bellovin, S. M. (2022) “Differential Privacy and Swapping: Examining De-Identification’s Impact on Minority Representation and Privacy Preservation in the U.S. Census,” Paper Presented at Conference: 2022 IEEE Symposium on Security and Privacy, pp. 457-472.
- Cranswick, K. (2019) “Virtual Data Labs - A More Flexible Approach to Access Statistics Canada Microdata,” Paper presented at Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, pp. 1-13.
- Desai, T., Ritchie, F. and Welpton, R. (2016) “Five Safes: Designing Data Access for Research,” Economics Working Paper Series 1601, University of the West of England.
- Dinur, I. and Nissim, K. (2003) “Revealing Information while Preserving Privacy,” in *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, ACM, pp. 202-210.
- Dwork, C. (2006) Differential Privacy, ICALP.
- Hawes, M. B. (2020) “Implementing Differential Privacy: Seven Lessons from the 2020 United States Census,” *Harvard Data Science Review*, No. 2.2, Spring 2020.
- Hawes, M. B., E. M. Brassell, A. Caruso, R. Cumings-Menon, J. Devine, C. Dorius, D. Evans, K. Haase, M. C. Hedrick, S. H. Holan, C. D. Hollingsworth, E. B. Jensen, D. Kifer, A. Krause, P. Leclerc, J. Livsey, R. A. Rodríguez, L. T. Rogers, M. Spence, V. Velkof, M. Walsh, J. Whitehorne and S. A. Keller (2025) “Towards a Principled Framework for Disclosure Avoidance,” Working Paper ced-wp-2024-005. <https://www2.census.gov/library/working-papers/2024/adrm/ced/ced-wp-2024-005.pdf> [2025 年 4 月 17 日アクセス]
- McKenna, L. (2019) “Research and Methodology Directorate: Disclosure Avoidance Techniques Used for the 1960 through 2010 Decennial Censuses of Population and Housing Public Use Microdata Samples,” U.S. Census Bureau.
- Muralidhar, K. and Ruggles, S. (2024) “Escalation of Commitment: A Case Study of the United States Census Bureau Efforts to Implement Differential Privacy for the 2020 Decennial Census,” in J. Domingo-Ferrer, and M. Önen, (eds.) *International Conference, PSD 2024 Antibes Juan-les-Pins, France, September 25-27, 2024 Proceedings*, Springer, pp. 393-402.
- Ritchie, F. (2008) “Secure Access to Confidential Microdata: Four Years of the Virtual Microdata Laboratory,” *Economic and Labour Market Review*, Vol. 2, No. 5, pp. 29-34.
- Ruggles, S., Fitch, C., Magnuson, D. and Schroeder, J. (2019) “Differential Privacy and Census Data: Implications for Social and Economic Research,” *AEA Papers and Proceedings 2019, Vol. 109*, pp. 403-408.
- Templ, M. (2017) *Statistical Disclosure Control for Microdata: Methods and Applications in R*, Springer International Publishing.

いとう・しんすけ 中央大学経済学部教授。最近の主要な論文に「デンマークとオランダにおける医療健康データの二次利用について」『日本統計学会誌』第 50 巻第 1 号, pp. 109-138 (2020 年)。経済統計学専攻。