第７回　統計委員会委員と統計利用者との意見交換会　概要

１　日　　時　平成 25 年３月 28 日（木）13：00～14：00

２　場　　所　中央合同庁舎第４号館 12 階　共用第 1208 特別会議室

３　出　席　者

【委　員】

樋口委員長、深尾委員長代理、縣委員、北村委員、西郷委員、白波瀬委員、椿委員、中村委員、廣松委員

【統計利用者】

神林　龍　　一橋大学経済研究所　准教授

伊藤　伸介　　明海大学経済学部　准教授

【国または地方公共団体の統計主管部課の長等】

内閣府経済社会総合研究所総括政策研究官、総務省統計局統計調査部長、財務省大臣官房総合政策課調査統計官、厚生労働省大臣官房統計情報部長、経済産業省大臣官房調査統計審議官、日本銀行調査統計局参事役、東京都総務局統計部長

【事務局等】

村上内閣府大臣官房統計委員会担当室長、清水内閣府大臣官房統計委員会担当室参事官、若林内閣府大臣官房統計委員会担当室参事官、平山総務省政策統括官（統計基準担当）、白岩総務省政策統括官付統計企画管理官

４　議　　事

（１）統計利用者からのプレゼンテーション

神林　龍　　一橋大学経済研究所　准教授

「データアクセスの行方」

伊藤　伸介　　明海大学経済学部　准教授

「イギリスにおける政府統計データの二次的利用の現状」

（２）意見交換

５　資　　料

資料１　事務局資料「第７回 統計委員会委員と統計利用者との意見交換会について」

資料２　神林龍　一橋大学経済研究所准教授資料「データアクセスの行方」

資料３　伊藤伸介　明海大学経済学部准教授資料「イギリスにおける政府統計データの二次的利用の現状」

６　議事概要

（１）統計利用者からのプレゼンテーション

事務局から、資料１に基づき意見交換会の趣旨及び論点等について説明が行われた後、神林准教

授から、資料２に基づき「データアクセスの行方」について説明が行われた。
・ミクロデータへのアクセスに関する国際的な集まりが進展しつつあり、例えばWDA（Workshop on Data Access）という、北米と欧州を中心にした各国の統計部局・研究機関が集まる会合（p3）の場で、各国のミクロデータへの取組が共有されるようになってきている。このような会合が開催される背景としては、ミクロデータを国際的に共有するという動きが、欧州を中心に強くなっているということが挙げられる。
・ミクロデータへのアクセスについては現在、p4 に示しているように、オンサイト利用と匿名化データという、大きく分けて２つの方法に収れんしつつある。匿名化データと呼ばれる方法については、対象となる調査によって匿名化の度合いを変えるという工夫が行われている一方で、オンサイト利用については、DRC（Data Research Center）において利用する際のプロセスに関する各国共通の土台ができつつある。
・例えば p5 に示しているように、オンサイト施設は物理的に外部と遮断することが必要であり、ネットワーク的にも分離する必要があるので、データはオンサイト施設に格納することになり、そのために施設に計算機を備え付ける必要がある。そして出力結果を外部に持ち出す際には、オンサイト施設に常駐する統計部局の施設の職員が確認し、その許可を得てから持ち出すことができるという手順が大体確立している。各国の主要な違いは、出力結果をチェックする方法にある。2000年代にオンサイト施設が出てきた当時は、出力結果の持ち出しについて、常駐する職員が目視でその都度結果をチェックしていた。現在では、これに工夫が加えられていて、外部に持ち出すファイルを中央の統計部局に送り、許可され次第、それを研究室で使用するという手順になりつつあり、リモートでチェックするという方法が普及してきている。
・オンサイト施設に常駐職員がいることが運営費用の中の大きな部分を占めており、リモートで出力をチェックできる場合には、その負担を軽減することができるという利点がある。リモートでチェックできる場合には、オンサイト施設は、物理的に管理された空間だけということになる（p6）。
・最近、欧州を中心に、個別の研究室をオンサイト施設とみなしてリモートアクセスを認めるということが始められている（p7）。リモートでチェックできるのであれば、物理的な環境の制御に論点が集約されるので、例えばオランダ、スウェーデン、フランス、デンマーク等の国々では、各研究室にウェブカメラや指紋認証を導入することで、オンサイト施設と同様とみなすということが行われている。このような形でリモートアクセスを行う場合、データは統計部局の中央サーバにしかなく、研究室のシンクライアントのシステムでアクセスをするので、出力は画面でしか見ることができず、写真に撮るなどしない限りデータは流出せず、国によってはそれで問題ないという理解をしていることになる。
・オンサイトシステムが現在非常に普及しているもう一つの理由として、各国で利用可能なデータが政府統計の個票に限らず、行政データにまで及びつつあることがある（p8）。代表的な例では、ドイツやアメリカの雇用保険のデータ、デンマークやフィンランドの医療保険のデータなどがあり、ミクロデータとして研究者に利用されるようになってきているが、このためには、やはりオンサイトのようなかなり制御された空間で扱う必要があると思う。ここで付け加えると、欧州では日本と状況が異なり、ユーロスタットという共通の統計部局に通常の調査を移し、調査を一元化する動きがある。そして、各国の統計部局が所管するデータの中で、行政データの占める比重が徐々に大きくなってきており、これらをどのように公共の福祉に利用するかということを考えたときに、オンサイト施設が利用できるということで、旧来は事業所系の政府統計の個票に利用されていたオンサ

イト施設の重心が、現在では行政データに移りつつあるというのが現状である。
・p9 に示したように、行政データには大きく分けて２種類ある。一つは統計に付随する情報で、法的には統計でなく行政データに入るが、統計を検証する上でかなり重要な情報になる。もう一つは雇用保険、医療保険等の本来の意味での行政データである。
・一つ目の、統計に付随するデータは、この統計委員会で扱える範囲なのだろうと思う（p10）。これらのデータの中には、保存期間が過ぎると廃棄されるものがあり、例えば、各府省の調査で独自に振られている番号とセンサスの番号の対応表や紙の調査票などが含まれる。それらは職業分類や産業分類の変更などの際に利用できると考えられるので、これらのデータについては、統計委員会において保存について何らかのアクションを起こしていただければと思う。
・もう一つの、行政に付随するデータについては、統計委員会で扱える範囲ではないと思う。これは過去に起きた事象を記録する役割を果たしていると思うが、行政のために作られているので、必ずしも使い勝手がよいわけではないという問題がある。アメリカの雇用保険データやフィンランドの医療保険のデータなどは、非常に長い時間をかけて、第三者が利用できるようにデータ構造を整えている。例えばアメリカの LEHD（Longitudinal Employer-Household Dynamics）というデータベースは 20 年ほどの経験を基に作られており、それぐらいの時間と労力がかかるということは研究者も認識すべきであり、コストを誰が負担するのかも議論すべきであると考えている（p11）。
・まとめるとデータアクセスについては、ほぼどの国も、オンサイト利用と匿名化データの２種類に収れんしているのが現状であり、研究者個人にデータそのものを提供するのであれば、匿名化データにするのが各国の基本的な動向であると考えている。我が国の現在の統計法第 33 条による調査票情報の利用のような方法は、各国と比較するとかなり例外的であると考えられる。オンサイト化は、事前審査のコストが軽減される一方で、事後の確認をしっかりと行う方式である。我が国の統計法第 33 条による調査票情報の利用はその逆で、事前の審査をしっかり行ということで、そのような考え方の違いがあると思う（p12）。オンサイト化を行った場合に、リモートアクセスを認めるかは、各国でも分かれる。デンマークなどでは、イタリアの大学などからデータを直接利用することが可能になってきており、またドイツの労働統計局は、アメリカのミシガン大学にオンサイト施設を開設している。このように、別の国のデータを利用できるということもオンサイト施設の一つのメリットである。行政データの保存と整理については、ここで言うべきことかはわからないが、オンサイト施設との親和性が高いと判断いただければと思う。

伊藤准教授から、資料３に基づき「イギリスにおける政府統計データの二次的利用の現状」の説明が行われた。
・イギリスには多様なミクロデータの提供形態があり、我が国の参考になると考えられる（p2、p3）。
・p4 に示したように、イギリスの政府統計のミクロデータは、（１）センサスミクロデータ、（２）サーベイミクロデータ、（３）LS データ（ONS Longitudinal Study of England and Wales）、（４）個体識別データ（identified data）の４つに大きく分けられる。これらのデータについては、インターネットによる利用やオンサイト施設におけるリモートアクセスによる利用など、様々な形で提供されており、その概略図を p5 に示している。
・まず１番目に、センサスミクロデータについて説明する。センサスミクロデータには、匿名化標本データ（Samples of Anonymised Records：SARs）、小地域ミクロデータ（Small Area Microdata：SAM）、CAMS（Controlled Access Microdata Sample）３つのタイプがある（p6）。

・匿名化標本データは、人口学的な属性や、就業属性、家族属性等を含むデータであり（p7）、そのニーズとしては、人口学・社会学系のものと、地理学的なものの２つの方向性があった。
・これらの異なるニーズにどう応えるかということで、最終的には 1991 年の人口センサスから、世帯 SAR と個人 SAR という２つのタイプの匿名化標本データが提供されている（p8）。個人 SAR は、抽出率が高く、地域区分が詳細であるものの、属性の区分は粗いものとなっている。一方で、世帯 SAR は、世帯属性が詳細であるものの地域区分が大まかであり、地域区分と属性がトレードオフの形で、データが作られている。また、1991 年の匿名化標本データには、センサスの項目からの導出変数の形でウエイトが作成されており、これはマンチェスター大学の CCSR というデータ・アーカイブが付与している。
・2001 年の人口センサスにおける世帯 SAR については、利用に特別なライセンスが必要な Special License 型で提供されており、地域区分も大まかなものとなっている。一方で、個人個人の匿名化標本データについては、Special License 型ではなく、抽出率が高いデータとなっている（p9）。
・レコードの抽出については、先に世帯のデータを抽出し、その後に残りのデータから個人のレコードを抽出するというように、個人と世帯のデータで重複しないような形で行われている（p10）。
・これらのデータの匿名化措置については、p11 に示しているように、データの削除や地域属性の制限など、いくつかの方法がある。1991 年の匿名化標本データには、データの削除などの非撹乱的な匿名化が行われている一方で、2001 年のものには、データの撹乱を行う手法が用いられているのが特徴的な点である（p12）。
・p13 では、イギリスでのデータの入手方法の事例について説明している。イギリス国内で研究者や学生が学術的に利用する場合、エセックス大学内の施設である UK データ・アーカイブに利用登録を行い、End User License Agreement に同意すると、ウェブ上で、無料でダウンロードすることができるようになっている。このようなライセンス型のミクロデータは、匿名化の程度が強いパブリックユースミクロデータではないという点に注意が必要である。これとは別の方法として、NESSTAR と呼ばれるオンデマンド型のリモート提供システムを通じたデータの提供もあり、これについてはモデルによる分析も可能となっている。
・2001 年の人口センサスにおける世帯 SAR のような Special License 型のデータの利用はハードルが高く、ライセンスを入手する際には、承認された研究者（Approved Researcher）の資格を得ることが条件となっている（p14）。
・次に、小地域ミクロデータ（SAM）について説明する（p14）。これは抽出率が高く、非常に細かい地域区分が利用可能なデータとなっており、先ほどの匿名化標本データ（SARs）とレコードが重複しないような形で抽出が行われている。
・一方で SAM は、p16 に示しているように、地域区分以外の他の属性については、区分が粗いか、あるいは利用可能でない場合があり、例えば、産業や職業等の属性は利用できない。このように、属性が細かく地域区分の粗い SARs とは、利用目的が異なるデータとなっている。
・p17 に示している導出変数とは、人口センサスで捕捉されている調査事項を基に付加された変数であり、これは個人 SAR や世帯 SAR には含まれるものの、SAM には含まれていないものである。導出変数は、イギリス国家統計局（ONS）が付加する場合と、マンチェスター大学のデータ・アーカイブである CCSR が新たに付加する場合がある。
・次に、CAMS（Controlled Access Microdata Samples）について説明する（p18）。CAMS は、ONS 内のオンサイト施設の内部のみで利用できるデータである。これが先ほどの SARs や SAM と

はタイプが異なり、法律上は個人情報に位置づけられている。個人 CAMS と世帯 CAMS が用意されており、属性は非常に細かく、例えば年齢は 95 歳までは各歳区分で利用可能であり、職業も非常に細かい区分が利用できるが、データを取得する際の審査が非常に厳しい。

・２番目に、サーベイミクロデータの提供について説明する（p19）。これについては、イギリス最大のデータ･アーカイブ施設である UK データ･アーカイブにより様々なデータが提供されており、ライセンスを取得すれば UK データ・アーカイブを通じて無料でダウンロードできるようになっている。

・３番目に、LS データについて説明する（p20）。これは、1971 年～2001 年のセンサス個票レコードをリンケージし、さらに政府保健中央レジスターの行政登録データをリンケージすることにより作成されている。

・LS データは、個人が特定される可能性が非常に高いデータであり、ミクロデータ提供審査委員会の審査が必要であるなど、申請の手続きが非常に煩雑なものとなっている。また、利用については、オンサイト利用か、あるいはオーダーメード集計の形に限定されている（p21）。

・４番目に、イギリス国家統計局（ONS）の個体識別データについて説明する（p22）。これは基本的に、ONS 内のオンサイト施設（Virtual Microdata Laboratory）のみで利用可能なデータである。個体識別データは、UK データ･アーカイブの中にあるセキュアアクセスと呼ばれる施設の中でも、リモートアクセスの形で利用することができる。個体識別データについては、承認された研究者のみが利用できるということがイギリスの統計法（統計登録サービス法）で規定されている。

・イギリスでは 2007 年に統計登録サービス法が定められており、ここに個人情報の秘密保護に関する条項が明記されている（p23）。

・統計登録サービス法では個人情報の利用対象について規定されており、その第 39 条で、承認された研究者によって研究が行われる場合には個人情報の利用ができると明記されている。このことが一つの大きなよりどころとなって、個体識別データが利用可能となっている（p24）。

・p25 に示しているように、個人情報に含まれるものには、個体識別データと Special License 型のデータがある。先ほど述べたように、個体識別データは、ONS 内のオンサイト施設内で利用できるが、利用の申請をして認められた者のみが利用できるようになっている。イギリスでは、こうした利用形態を safe setting かつ safe people と呼んでいる。

・こうした個人情報を利用するためには、研究者の資格を取得するために申請書類を提出し、p26 で示したような様々な形での承認がなければ利用できない形になっている。

・イギリスでは、匿名化データの提供とオンサイト施設での提供、さらにはその中間的な形で、非常に詳細な情報を含み匿名化の程度は低いが手続きが非常に厳しいものという様々なタイプのミクロデータが提供されており（p27）、こういった提供の状況が、今後の我が国のミクロデータ提供の方向性を考える上で、一つの参考事例になるのではないかと思う（p28）。

（２）意見交換

・リモートアクセスによりミクロデータを利用する際に使用するソフトウェアについては、データの提供側が用意する形になると思うが、その際に利用可能なソフトウェアの種類はどのようになっているのか。研究者が利用しやすいようなものが数多くそろえられているのか、あるいは種類が限定されているのか。

→ソフトウェアについては、各国で事情が違っているが、基本的に非常に制限されている。SAS や

Stata が代表的なソフトウェアであり、さらに通常のプログラム言語が装備されているが、最近ではＲなど、簡単な言語も使用できるようになってきている。リモートアクセスの場合には、計算機資源が中央に全て集中するため、ソフトウェアの種類を増やすと、計算機に負荷を与えることになるので、各国の統計部局はその点、かなり慎重に、制限をしている。また、オランダの例では、研究者が資金を集めて中央のサーバを増設するのでこのソフトウェアの導入をお願いする、といったことが行われている。

・先ほど、インプットではなく、アウトプットを審査する方法についての話があったが、これは有望な方法であると考えられる。その際に、使用するソフトウェアにもよるが、かなり多様なアウトプットが出てくるので、審査の方法が非常に難しいと思うが、そのような審査ができる人をどのようにリクルートして、どうトレーニングしていくのか。これは、このようなシステムを運営する際に、実務上大切なことではないかと思う。

→ソフトウェアを制限しているので、アウトプットはある程度想像がつくというのが、各国の統計部局の意見である。日本と異なり、欧米の統計局の担当者は、大体が経済学や社会学の博士号取得者であり、どのようなアウトプットが出てくるかはある程度想像がつくようである。オランダの場合は、匿名化が破られていないかを自動的に判断するプログラムを作成することも試みており、Stata のアウトプットを流すと自動的にそのリスクがわかるようになっている。出力の確認には専門家が必要となるので、オンサイトにそれらの者を配置すると、かなりの負担となると理解されているのだと思う。

・オンサイト施設の運営費用はどこが負担しているのか。アメリカでは、ホストとなる研究機関や教育機関が負担していると聞いたことがある。あるいは、利用者から料金を徴収しているのか。

→これも各国で異なるが、カナダ等では基本的に、オンサイト施設を設置する機関が全額を負担する。よって、アウトプットのチェックをする者を雇う場合には、その人件費も負担することになる。統計部局は一切補助を行わない。欧州の場合は様々な交渉が入ってくるのでよくわからない部分もあるが、フランスでは費用の負担を巡って対立が起こり、それがリモートアクセスを促進する方向につながったと聞いている。フランスのオンサイト利用では、統計部局が専用のソフトウェアを開発し、研究者が自分のコンピュータにインストールすることでネットワーク上独立した経路を確保するという形になっており、その場合にはランニングコストがほとんどかからないので、そのような形で、各大学にオンサイト、リモートアクセスが広まっているのが現状である。

・イギリスでは、一部のデータの利用に当たり、認証された研究者の資格が必要とのことであったが、その審査では何が要求されるのか。

→個票データを利用するための研究計画や学術研究プロジェクトの概要などについて提出し、申請手続きを行い、ミクロデータ提供審査委員会等で審査され、承認を得て初めて、承認された研究者の資格を得ることができる。我が国でいえば、科研費などの競争的資金を導入し、統計法第 33 条第２号により調査票情報を利用する形に近い。

・アウトプットのチェックはかなり個別的に行われるようであるが、どのように確認が行われるのか。例えばアメリカでは DRB（Disclosure Review Board）という組織を設けて、そこが確認を行う体制をとっているようであるが、ヨーロッパではそれがどのような形になっているのか。

→DRB はかなりフォーマルな組織であると聞いているが、欧州ではそこまでの形をとっているものはないと聞いている。ただし、最初にオンサイト施設を利用する際に、アウトプットについて、統計部局の承認がないと持ち出せないという約束になっているので、その過程でチェックが行われ

ていることになる。
・認証された研究者の資格については、いったん承認されれば、他のミクロデータの申請にも用いることができるのか。
→研究プロジェクト単位で申請するものであり、プロジェクトの終了とともに資格は失われる。
・イギリスの個体識別データについては、我が国でいえばどのような位置づけになるのか。
→我が国でいえば、統計法第 33 条による調査票情報の利用に相当するものと認識している。これには世帯・人口系や事業所・企業系のデータのほかに、ビジネスレジスターのようなものも入ってくるが、これらは秘密保護に留意する必要のあるデータであることから、オンサイト施設やリモートアクセスなどの非常に限られた方法で、かつ承認された研究者という資格を得ない限りは利用できないというように、非常にハードルが高いものとなっている。
・匿名化処理の程度の高いパブリックユースファイルが提供される場合、統計法第 33 条による利用を考えなくても、研究上支障は生じないか。パブリックユースファイルでも、研究のレベルは十分保たれるのか。
→パブリックユースファイルでできる研究の範囲は非常に狭いので、提供がパブリックユースファイルのような形に限定された場合には、一般的に研究水準は落ちると思われる。これについて、個人的な感想に過ぎないが、北米では、パブリックユースファイルをどう分析するかを競うべきであり、自分だけが利用できるデータを用いて論文を書くのは反則であるという雰囲気がある。一方で、欧州では、現実のデータから必要なことを適時に発信することが必要であるという考え方があり、パブリックユースファイルだけに拘泥してそのタイミングを逃したりすることなどは避けなければならないという雰囲気があると感じるので、大陸によってそこは考え方が分かれると思う。日本に関してはおそらく、パブリックユースファイルに限定すると研究水準は落ちると思う。
・リモートアクセスを行う場合にカメラや指紋認証の機器を導入するという話があったが、そのような機密保護の水準をクリアするような技術水準は既に確立されていると考えてよいか。
→確立されていると思う。指紋認証に関しても、安価にできるようになっている。
・それらに関する共通の技術的な合意、スタンダードのようなものはあるのか。
→徐々にできつつある。それを作成するために、各国で情報交換が行われている。WDA では、シンクライアント技術や、画面出力制御ソフトなどに関することが話し合われており、ヨーロッパではそれらの情報が流通している。
・アウトプットの確認では、どこに基準を置いて審査しているのか。
→匿名化が破られる可能性があるか、個体を識別できるかという点を基準に審査が行われている。欧州では統計部局の中で共通認識ができており、基本的には論文に掲載する表や図など、本当に外部に持ち出す必要があるものだけを審査する形となっている。それ以外の中間データ等は、オンサイト施設の中で確認することとされている。よって、推定結果や散布図などは問題になる場合がある。
・ドイツではミシガン大学にオンサイト施設を設置したとのことであるが、今後そのような動きは進む方向にあるのか。
→ドイツのような取組は、各国でも進められる方向にある。その背後には、データの国際競争が起こっているということがあり、いかに自国のデータを利用してもらうかということに、各国の統計部局がかなり力を割いてきているところがある。
・承認された研究者の資格はプロジェクト単位に付与されるとのことであったが、実際にデータを

利用するのは研究者個人であり、プロジェクトと個人というのは区別されているのか。
→研究プロジェクトのメンバー全員を含め、プロジェクト全体として、承認された研究者という資格を得るが、実際の手続きはその代表者が進める形になるので、プロジェクト単位という表現をした。
・リモートアクセスによる分析を行うことができる調査は限られているのか。あるいは全ての調査で利用できるのか。
→リモートアクセスによる利用を許可するデータについては、漏洩しても問題ないものである場合が多い。例外はスウェーデンで、全てのデータがアクセスできるようになっている。北欧では、健康に関わる情報にリモートアクセスを認めることは非常に難しいとのことで、現在デンマークで、リモートアクセスを認めるかということを検討しているようである。
・行政データの保存に関する話があったが、日本ではこの部分はどうなっているのか。統計として利用されていない行政データについては、どの程度の期間保存されているのか。行政記録の保存についてのガイドライン等はあるのか。
→保存は各府省に任されているようである。原票の保存は府省によって異なるかもしれないが、電子化された情報については永年で保存されることとなっている。行政データについては、各府省の文書管理規則の範疇であると思うが、個別の状況については、現時点での情報はない。逆にあまり長く保存してはいけないという考え方もあり、一定の期間後に責任を持って廃棄することになる。
・行政データの位置づけやその研究対象としての位置づけについて、日本と欧米との違いの背景には何があるのか。
→データの作成に関する研究者の意識が異なると思う。アメリカでは研究者自身が PSID（Panel Study of Income Dynamics）などの調査データを作成してきたという経緯があり、政府統計の設計、改善などに関しても研究者が研究目的で意見を言うといったことが多く行われている。それに対して欧州では、データは政府が作るものであるとの認識が一般的であり、研究者の側がデータの蓄積を行っておらず、利用できるデータがないという状況で、北欧中心に行政データの利用が進んできたというのが大まかな傾向である。日本はこの点ではどちらかというと欧州に近く、研究者がデータを作成することに無関心であった。
→北欧では人口が800万人程度であり、統計調査を行うだけのコストをかけることができないので、行政データを使わざるを得ないという理由も大きいと思う。

以上

＜文責内閣府大臣官房統計委員会担当室＞