資料１－２

平成 24 年 12 月 19 日

オンサイト利用に関する提案

統計研修所　小林良行*

1　用語の整理

・オンサイト利用

研究者がオンサイト施設に出向いて、施設内で秘匿性の低い統計データを利用し統計を作成すること。

・オフサイト利用

オンサイト利用以外の統計データ利用形態。

・オンサイト施設

政府統計機関[1]またはその委託を受けた独立行政法人(以下、「政府統計機関等」)により、高度な情報安全性(運用管理に必要な手続き、規定等の整備を含む)を備えていることを認定された施設及び／または設備。

・オンサイト施設設置機関

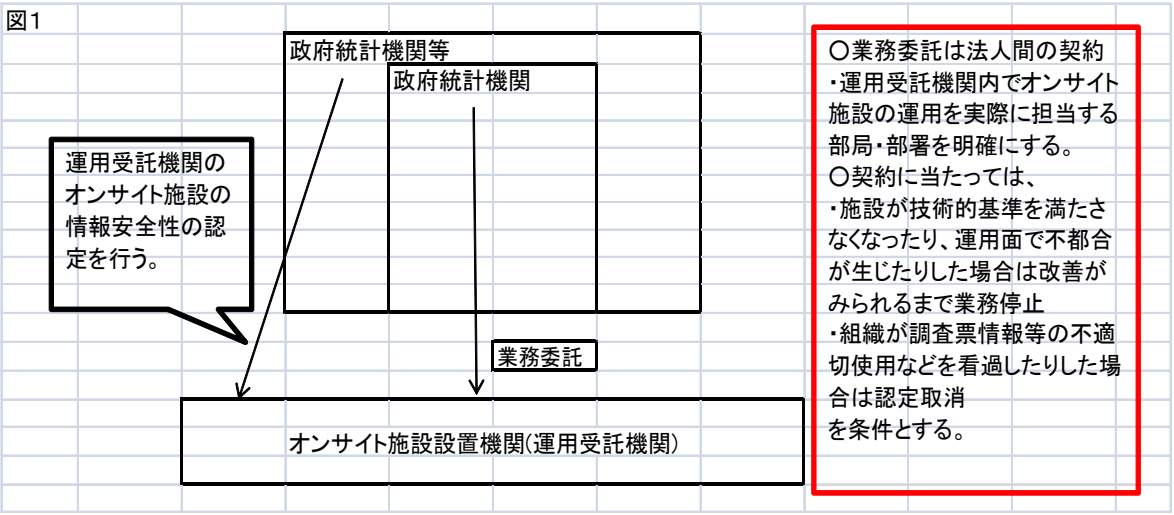
オンサイト施設を設置し、その運用を行う機関。

・オンサイト施設の運用形態

①直轄方式：政府統計機関等が、その建物内に設けたオンサイト施設の運用を行う方式。オンサイト施設設置機関は政府統計機関等である。

②委託方式：政府統計機関等からオンサイト利用の運用業務の委託を受けた大学法人または大学共同利用機関法人(以下、「運用受託機関」)が、その建物内に設けたオンサイト施設の運用を行う方式。オンサイト施設設置機関は運用受託機関である。

図1



---

* 筆者は総務省統計研修所に所属しているが、本稿中の意見・見解は個人のものである。本稿の記述内容に誤りがあればそれは筆者の理解不足や誤解によるものであり、その責はすべて筆者にある。

[1] 本稿では、公的統計の作成を行っている府省のこととする。

２　オンサイト利用に関連したいくつかの提案

○匿名化の程度から見た現行の我が国のミクロデータを整理すると表１のようになる。

○諸外国の政府統計機関が統計データを研究者に提供する形態として、近年では以下のものがある[2]。

(1)統計表及びデータキューブ[3]

(2)匿名化データ(Anonymized Microdata Files)

(2a)Public Use Files(PUFs)

(2b)licenced files(利用に所定の要件があり政府統計機関の承認が必要。独の SUF、日の匿名データが相当)

(3)Remote Access Facilities(RAFs)

(3a)Remote Execution(プログラムを送付し、実行結果の秘匿性の検査を受けた後、成果物をインターネットで受取るもの)

(3b)Remote Facilities(オンラインで分析を実行し、結果を得るもの。蘭以外では行っていない)

(4)Data Laboratories(DL)

我が国の二次的利用を考える上で、現在実現していないデータキューブ、PUF、リモートデータアクセス、オンサイト利用の提供を総合的に考えていくことが必要。また、情報安全性を保証する情報通信技術の動向についてもフォローしておく必要がある。

○オンサイト施設での調査票情報の利用は事後チェック型とし、オンサイト施設内で行う実証研究は検証型の研究に加えて探索型の研究も認めるようにする(小林(2011b))。

○オンサイト施設は、高度な公益性を持つ研究を行う公的統計共同利用拠点として利用される施設であるべき。利用者の便宜を考えた場合、オンサイト施設設置機関が政府統計機関等であれ運用受託機関であれ、一つのオンサイト施設内ではどの政府統計機関の調査票情報も利用できるようにすることが必要。

⇒・ある政府統計機関が情報安全性を認定したオンサイト施設を、他の政府統計機関も自機関で認定したのと同じとみなすようにすることによりオンサイト施設を各府省が個別に作るという無駄を省ける。

・利用者が選択できる場所の数が増える。

○運用受託機関のオンサイト利用では、継続的なサービス提供を行っていけるよう、研究用 PC やサーバーなどのハード・ソフトの将来の更新等のための推定費用、利用者が使用する消耗品代などに充てるための料金を設定し、利用者から徴収することができるようにする。

○オンサイト施設が満たすべき情報セキュリティ基準は、国際的な標準を踏まえて作成される

---

[2] これらの統計データ提供形態については、政府統計機関が学術研究を支援する際のガイドラインとして UNECE がまとめた UNECE(2007)の p.9-15 を参照のこと。また、各国の事例が UNECE(2007)p.26-104 に紹介されている。

[3] データキューブは情報通信技術の進歩を背景に 2000 年代に入って新たな提供形態として登場したもので、公表統計表(マクロデータ)より詳細な中間集計データを用意し、利用者の要求に応じて中間集計データの持つ属性を組み合わせた集計表を作成、提供するものである。小林(2011c)はデータキューブをプログラム送付型オーダーメード集計のテストデータとして利用することを提案。

ことが必要(たとえばISO)。ただし、政府統計機関職員が随時、利用者を監督できるような直轄方式のオンサイト施設運用では、パーティションで執務室内の一画を区切るなどし、情報安全性を確保した PC を設置した区画をオンサイト施設として運用してもよいこととする。

○以下の個別データは個体識別性、秘匿の困難性が特に高いと考えられることからオンサイト利用のみ認めることとする。

・フルセットのセンサスデータ

・リンケージデータ(※)

・ビジネスデータ(事業所・企業に関するデータ)のように、重要な変数の分布が非常にskewで匿名化が難しいデータ

・位置情報付きデータ(個体の時空間座標と属性情報が組になっているようなデータ)

(※)調査票情報をマッチングしてリンケージデータを作成することを許しているのはカナダ、スウェーデンのみで、作成を認める要件も限定的(UNECE(2007))。

しかし、オンサイト施設内で作成、利用したリンケージデータの項目のうち、リンケージ情報(複数の調査票情報をマッチングするのに使われるキー項目の組合せを示す項目のみを取り出したもの)は、学術研究の共有材として極めて有用であるので、①その存在をオンサイト施設設置機関間で情報共有し、②作成されたリンケージ情報は作成場所のオンサイト施設設置機関が管理し、③政府統計機関は、オンサイト施設設置機関間で相互にリンケージ情報の貸出し及びオンサイト利用をしようとする研究者からの申出に応じたリンケージ情報の提供を認める、といったことの実現が必要。

○諸外国では、オンサイト施設の設置(新たな建物建設が伴うなどのため)に費用がかかること、必ずしも利用者が便利なところにオンサイト施設があるとは限らないことなどの理由で、リモートデータアクセスによる調査票情報の利用手段を提供している。我が国でも調査票情報や秘匿性の低い匿名データのオフサイト利用はリモートデータアクセスのみを認めることとする。現在行われている研究室などでのオフサイト利用は、個別データの利活用に対する社会的成熟(国民の理解、個々の研究者のデータ利用に関する倫理、国際的な動向など)が進むまでひとまず凍結とすべき。

○ドイツでは新しい試みとしてRDCinRDC[4]が進められている。

○調査票情報や匿名化の程度が低い個別データの利用者には、法令、運用手続きに関する講習(統計法の趣旨、二次的利用の法的根拠、罰則、申出手続きなど)と講習終了後の理解度評価で一定のレベルに達することを義務付けし、また受講後一定年数経過後の再受講を義務付け(再受講は、法令や制度の運用に改正があり得るため)。講習内容等は総務省が定め、講習は総務省が行うようにすればよい。

○オンサイト利用を含む統計データ提供形態について表２に提案する。

---

[4] たとえばHeining,J(2009)、Brandt et al.(2009)を参照。

参考文献

Brandt,M.,Croessmann,A. and Gueke,C.(2009) . *Harmonization of Statistical Confidentiality in the Federal Republic of Germany*,WP.5,Joint UNECE/Eurostat work session on statistca data confidentiality,2009,Bilbao

Heining,J(2009) . *The Research Data Centre of the German Federal Employment Agency:Data Supply and Demand between 2004 and 2009*, Working Paper 129, German Council for Sacial and economic Data

小林良行 (2011a).「匿名データの教育目的利用に関する一考察」『統計学』第 100 号, 100-105

小林良行(2011b).「公的統計ミクロデータの現状と展望」『日本統計学会誌』,41(2),412-413

小林良行 (2011c).「プログラム送付型オーダーメード集計のテストデータ　－　メソデータの可能性」 , 2011 年度「官庁統計データの公開における諸問題の研究と他分野への応用」 研究集会での報告

UNECE(2007). *Managing Statistical Confidentiality & Microdata Access*, United Nations,UN

表1　匿名化の程度からみた我が国のミクロデータ　ー　ドイツとの比較

日本のミクロデータ		ドイツのミクロデータ				
ミクロデータの種類	利用目的	匿名化の程度	情報損失	情報の有用性	ミクロデータの種類とアクセス形態	利用目的
調査票情報	(注1)	非匿名化	低	高	アクセス不可	
		形式的な匿名化(注2)	↑	↑	RDC、CRDPによる利用(注3)	学術研究目的
匿名データ(国外での利用可)	・学術研究目的 ・高等教育目的 ・国際比較統計利活用事業目的	事実上の匿名化	↓	↓	SUF(注4) (国外からの購入不可)	学術研究目的
ー	ー	絶対的な匿名化	高	低	PUF (国外からの購入可)	一般I利用目的
					CAMPUS Files(注5)	教育目的

(*)小林(2011a)の表1に加筆。
(注1)調査票情報には個人の氏名や会社等の名称，住所を含んでいるものと氏名や名称，住所を含まないものの二種類がある。前者のタイプの調査票情報は，紙媒体や画像に記録された調査票及び事業所・企業名簿のもとになる統計調査の電磁的記録が該当し，表1では非匿名化データのカテゴリに分類されるものである。一方，電磁的記録媒体に記録されているほとんどの調査票情報は後者のタイプ、すなわち形式的な匿名化データである。
(注2)氏名，住所のような直接的識別子をなくしたもの
(注3)CRDP(Controlled Remote Data Processing)は，我が国のオーダーメード集計と対応。
(注4)Research Data Center内でのみ利用可能なオンサイトSUFとオフサイト利用に供する標準化SUFの2種類ある。
(注5)大学レベルの教育用。統計的な方法、分析の能力(Methodological skills)の学習と社会経済問題(sociological & economic issues)の分析能力の教育に利用することを想定して作成されている。1998年ミクロセンサス、1998所得税統計など複数のファイルが提供されている。大学以下の教育で利用できるような、データ量が小規模の教育用データの作成も検討されている。

表2 統計データの提供形態(想定)

データ粒度	提供形態		利用目的の範囲		直接利用		間接利用		
			学術研究	教育	オンサイト（注１）	オフサイト	オーダーメード集計		
							依頼集計型	プログラム送付型(注2)	オンライン集計型(注3)
細かい	調査票情報		○	×	○（事後チェック方式）	×（凍結）	○	○	×
	匿名データ	弱匿名化データ	○	×	○（事後チェック方式）	×	×	○	×
		標準匿名化データ（現行の匿名データ）	○	○（高等教育）	○	○	—	—	—
		一般利用匿名化データ	—	○（高等教育）	—	○	—	—	○
		教育利用匿名化データ	—	○（高等教育）（中等教育）	—	○	—	—	—
	データキューブ（注4）		—	○	—	○	—	—	—
粗い	公表結果表		—	○	—	○	—	—	—

○:可　×:不可　ー:対象外

(注1)オンサイト施設から成果物を持ち出すに当たり、政府統計機関に申請してから検査を受けて承認を得るまでの標準事務処理期間を設ける必要がある。
(注2)プログラム送付型のオーダーメード集計は、メールでプログラムを送付、政府統計機関の秘匿性の検査を受けた後、結果を返送する。秘匿性の検査から結果の返送までの標準事務処理期間を設ける必要がある。
(注3)組合わせ可能な分類項目に制限を設けるなど制約が必要。
(注4)提供の元となるデータの粒度が細かい場合は、直接利用ではなくオンライン集計型のオーダーメード集計とし、組合わせ可能な分類項目に制限を設けるなどの制約が必要。