資料

保 存 期 間 ： 5 年
（令和９事務年度末）
令 和 ４ 年１０月１４日

第２回 国税庁保有行政記録情報の整備に関する技術検証WG

国税庁　企画課

資料内容

1. 本ワーキンググループの経緯・位置づけ

2. これまでの議論

3. 本日ご検討いただきたい内容

4. 今後のスケジュール

１．本ワーキンググループの経緯・位置づけ

○　国税庁が保有する行政記録情報のオープン化に向けた検討を効率的に行うため、法的な課題及び技術的な課題に対する具体的な対応方法について検討・確認を行うことを目的として、国税庁保有行政記録情報の整備に関する有識者検討会の下で、本ワーキンググループ（以下、WG）を開催する。

> 「国税庁保有行政記録情報の整備に関する有識者検討会」開催要綱（抜粋）
> ３ 運営
> （２）座長は必要があると認めるときは、検討会にワーキンググループを置くことができる。
> なお、ワーキンググループにおける検討結果は、有識者検討会に報告するものとする。

○　第２回となる本WGでは、匿名化データが効果的に活用されるために、研究者目線からみた利用者のニーズを把握することを目的として開催。

○　WGにおける検討結果については、事務局（国税庁企画課）において整理の上、「国税庁保有行政記録情報の整備に関する有識者検討会」に対して検討状況を適宜報告することとする。

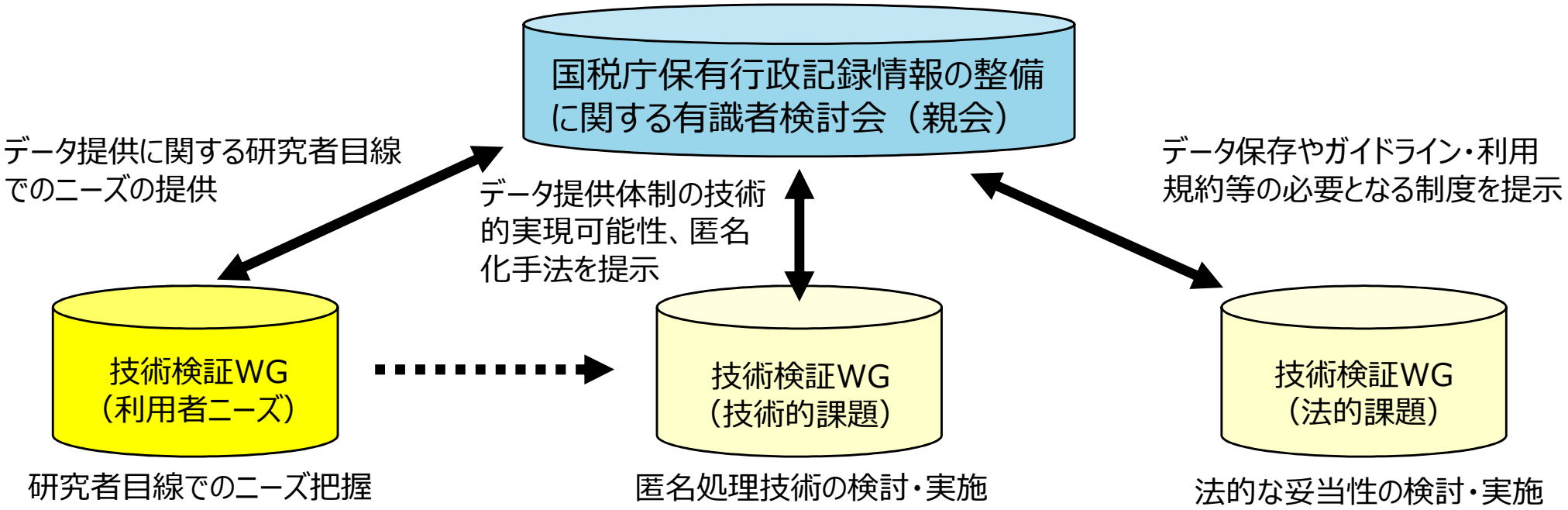
○　第２回WGの構成員は、以下のとおり（敬称略）。

伊藤　伸介	中央大学　経済学部　教授
宇南山　卓	京都大学　経済研究所　教授
土居　丈朗	慶應義塾大学　経済学部　教授

１．本ワーキンググループの経緯・位置づけ

○　国税庁保有行政記録情報の整備に関する有識者検討会は、統計学、経済学、法律の各専門家から構成され、全体の方向性を検討することを主な役割とする。

○　技術検証WGは、データ提供に関する研究者目線でのニーズを把握するための**利用者ニーズの把握**を目的としたもの、そのうえで匿名化を施すうえでの**技術的課題の検証**を目的としたもの、さらに、議論の進展に応じて、データ利用に際しての法的規律を検討する**法的課題の検証**を目的としたものの開催を検討する。なお、WGの検討内容は有識者検討会へ報告する。



資料内容

1. 本ワーキンググループの経緯・位置づけ

2. これまでの議論

3. 本日ご検討いただきたい内容

4. 今後のスケジュール

２．これまでの議論　①データの提供形態について

※令和４年６月22日開催
第２回有識者検討会資料の再掲

○　まずは、サンプルデータ及びメタデータを公開し、研究者に広く触れていただける環境を整備することとしてはどうか。

　併せて、サンプルデータ及びメタデータを入り口として、①リモートエグゼキューション、②データ貸出／閲覧※の２種類を用意し、それぞれの利点と手続き上の負担を周知することで、ニーズに応じた税務データの学術研究利用を促進させることが可能となるのではないか。

○　上記の整備・検討と並行して、パターン②のデータ提供を実現すべくデータを完全に匿名化する技術の検討を行うこととしてはどうか。

※　データ貸出：CD-R等の媒体にデータを格納して貸出し、使用後に返却する。
※　閲覧：国税当局の施設に来庁し、閲覧・利用する。

○　パターン①

ステップ１　サンプルデータ及びメタデータの公開

・実際のデータの分布に類似した、分析に耐えうる程度（データ量）のデータセットを作成
・特段手続を要することなく、審査不要で自由にダウンロードできるようにし、データ説明書である「メタデータ」も整備

ステップ２－１　データ提供（リモートエグゼキューション）

・研究者の方でプログラム等を送付し、結果のみを提供
・手続きは全てメールでのやりとりで完結、国家公務員の身分委嘱は不要、要審査

ステップ２－２　データ提供（データ貸出／閲覧）

・匿名化が施されたデータ（SUF）を貸出／閲覧
・貸し出しの場合、手続きは全てメールでのやりとりで完結、国家公務員の身分委嘱は不要、要審査
・リスク管理の観点から閲覧とする場合は、国税庁の施設に来訪する必要あり、要審査

○　パターン②

データ提供（匿名化されたデータの公開及びメタデータの公開）

・完全に匿名化が施されたデータ（PUF）を公表し、特段手続を要することなく、審査不要で自由にダウンロード
・併せてデータ説明書である「メタデータ」も整備

２．これまでの議論　①データの提供形態について

※令和３年10月29日開催
第１回有識者検討会資料を一部修正

○ 「データセット固定方式」にする場合、提供データ（SUF/PUF）の整備に当たっては、どのようなデータ項目で、どの程度の匿名化を施したデータを整備する必要があるか、守秘義務を遵守しつつ、研究ニーズ等を踏まえた上で検討していく必要がある。

○ 「オーダーメイド方式」にする場合、それに応えるための研究用行政記録情報について、どのような非識別加工をすれば長期保存可能かを検討することが前提となる。

	データセット固定方式	オーダーメイド方式
概要	整備するデータセットを予め決定し、そのデータのみを提供する。	ニーズに基づき、必要なデータ項目を指定させた上で、都度データを払い出す。
メリット	・データセットを予め固定するため、データの整備が比較的容易。	・細かいニーズに応えやすい。 ・指定するデータ項目が少なければ、粒度の細かいデータ提供が可能。
デメリット	・細かいニーズを汲み取りにくい。 ・整備するデータ項目が多岐にわたる場合、守秘義務の観点から粒度の粗いデータを整備せざるを得ず、研究目的に堪えない可能性。	・提供の都度、匿名化が施されているか確認する必要があり確認のためにリソースを割く必要がある。 ・提供した粒度の細かいデータ同士のマッチングにより、意図せず個人が特定される恐れ。
国税当局側の負担	比較的低い （一度データを整備すれば継続的に対応が可能）	高い （都度、オーダーに応じたデータを作成。 匿名化がなされたかの確認・審査が必要）
受入可能件数	広く受け入れることが比較的可能	広く受け入れることは困難 （公募等により年間数件程度か）

2．これまでの議論　①データの提供形態について

○　提供形態として、①データ貸出（CD-R等の媒体にデータを格納して貸出し、使用後返却）、②データ閲覧（国税当局の施設に来訪し閲覧・利用）のいずれかが考えられる。

○　コンプライアンスリスクに応じて、例えば、ガイドライン・利用規約における制限や、税目によっては提供形態を限定する等の対応も考えられる。

	データ貸出方式	データ閲覧方式
利用者の利便性	高い	低い （利用者は国税当局の施設に往訪する必要）
国税当局側の負担	比較的低い （閲覧場所等の整備は不要、貸出作業は発生）	高い （閲覧場所を整備するなどの対応が必要）
受入可能件数	広く受け入れることが比較的可能 （データの貸出事務のみが発生）	広く受け入れることは困難 （閲覧場所のスケジュール管理等が必要）
コンプライアンスリスク	高い	低い

２．これまでの議論　②データ項目抜粋

※令和３年10月29日開催
第１回有識者検討会資料の再掲

個人課税関連	法人課税関連
確定申告書	確定申告書
青色申告決算書・収支内訳書	法人税申告書別表ファイル
各種届出書	財務諸表（貸借対照表）
個人事業者の消費税申告書	財務諸表（損益計算書）
資産課税関連	連結グループ情報
相続税申告情報	各種届出書
贈与申告情報	法人事業者の消費税申告書

2．これまでの議論　③サンプルデータについて

○　サンプルデータの役割については、①データ提供に際して、研究者等が事前にデータ構造を理解することにより、広く利用されるきっかけを提供すること、②大学等におけるデータ分析等の教育用途としても利用可能であり、③将来的には匿名化データの匿名化のノウハウを蓄積する観点から、サンプルデータを提供することとしてはどうか。

○　サンプルデータの整備に当たっては、サンプルデータの役割（特に上記①）や実現可能性を考慮し、まずは、乱数を発生させるなどして作成する方法による疑似データを整備する方向性としてはどうか。

○　サンプルデータの公表タイミングについては、データ提供の実現時期を考慮する必要がある。

論点	疑似データの特徴
想定される利用目的	・データ分析等の教育目的 ・共同研究、匿名化データ利用への準備
税務データとの関連	なし（※）
研究分析における利用可能性	疑似データであり、論文への引用は不可
保持できる変数 (収入・所得・控除項目等)	制限なし
実現可能性	比較的容易

（※）疑似データの作成にあっても、一定程度税務データと所得分布等が一致していることが求められるか。

（参考）匿名加工の技法について

※令和３年10月29日開催
第１回有識者検討会資料の再掲

○ 非識別化の手法は、以下の表のとおり、様々な知見の蓄積がある一方、対象データや、求めるレベルに応じて、適用すべき技法は様々。
○ どの水準まで加工が必要か、技術視点、ユーザー視点、法的視点等から検討する必要。

No	代表的な技法例	技法例	概要
1	属性情報の削除	属性（列）削除	直接個人を特定可能な属性（氏名等）を削除すること。
2		仮名化	直接個人を特定可能な属性またはその組み合わせ（氏名・生年月日）を符号や番号等に置き換えること。例えば、ハッシュ関数。
3	属性情報の一般化	一般化	・属性の値を上位の値や概念に置き換えること。例えば、10歳刻み、キュウリ→野菜。 ・データ全体に行うものをGlobal Recoding、局所的に行うものをLocal Recodingと呼ぶ。 ・四捨五入や二捨三入などを丸め法（Rounding）と呼ぶ。
4		あいまい化	数値属性に対して、特に大きい、もしくは小さい属性値をまとめる。 例えば、100歳以上の人は「100歳以上」とする。
5	属性情報の可能技法 ※ 原文ママ	ミクロアグリゲーション	元データをグループ化した後、同じグループのレコードの各属性値を、グループの代表値に置き換えること。
6		ノイズ（誤差）の付加	数値属性に対して、一定の分布に従った乱数的なノイズを加えること。
7		データ交換	カテゴリー属性に対して、レコード間で属性値を（確率的に）入れ替えること。
8		疑似データ作成	元のデータと統計的に疑似させる人工的な合成データを作成すること。
9	その他技法	レコード（行）削除	特に大きい等、特殊な属性（値）を持つレコードを削除する。 例えば、120歳以上のレコードは削除する。
10		セル削除	センシティブな属性値等、分析に用いるべきでない属性値を削除する。
11		サンプリング	元データ全体から一定の割合・個数でランダムに抽出すること。

（出所） 高度情報通信ネットワーク社会推進戦略本部（ＩＴ総合戦略本部） パーソナルデータに関する検討会 技術検討ワーキンググループ報告書（2013年）

資料内容

1. 本ワーキンググループの経緯・位置づけ

2. これまでの議論

3. 本日ご検討いただきたい内容

4. 今後のスケジュール

３．本日ご検討いただきたい内容

＜①データの提供形態について＞

○ コンプライアンスリスク等も加味したうえで、データ貸出方式又はデータ閲覧方式のニーズについて。

○ リモートエグゼキューションに対するニーズについて。

＜②提供データの項目について＞

○ 対象税目や、対象個票の優先順位について。

○ データ項目の中でも、学術研究用途に際して必ず保持されておくべき変数及び許容される匿名化水準について。
例：住所情報について、どこまで残すべきか。
高額所得者における合計所得等の特定リスクが高い変数について、どの金額まで残すべきか。

＜③サンプルデータについて＞

○ サンプルデータの活用可能性をどのように評価できるか。

＜その他の論点＞

○ データを利用できる者の範囲
例：①指導教授等の監督の下であれば博士研究員や大学院生の使用も認める、②国・地方公共団体等の行政機関、③民間シンクタンク等の研究者等

○ データの利用目的の範囲
例：現状の共同研究では、ガイドライン上、税・財政施策に資することが要件となっている。

資料内容

1. 本ワーキンググループの経緯・位置づけ

2. これまでの議論

3. 本日ご検討いただきたい内容

4. 今後のスケジュール

４．今後のスケジュール

※令和４年６月22日開催
第２回有識者検討会資料を一部修正

○ 原則として有識者検討会と技術検証WGを交互に開催し、両方合わせて１年で４回程度の検討を行う。

○ 令和５年６月までに整備の方向性についての議論を終え、令和５年７月から令和６年６月までに具体的なデータの整備・検証を行い、令和６年７月から対外的に行政記録情報の提供を開始することを目標とする。

○ 各WGでの検証も踏まえつつ、提供するデータ、方式及び場所に関しては、有識者検討会において決定する。