資　料

保 存 期 間 ： ５ 年
（令和９事務年度末）
令 和 ５ 年 １ 月 １ ０ 日

# 第３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３回となる本WGでは、各データ提供形態における課題の把握や、施すべき匿名加工技法の検討を行う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３回WGの構成員は、以下のとおり（敬称略）。

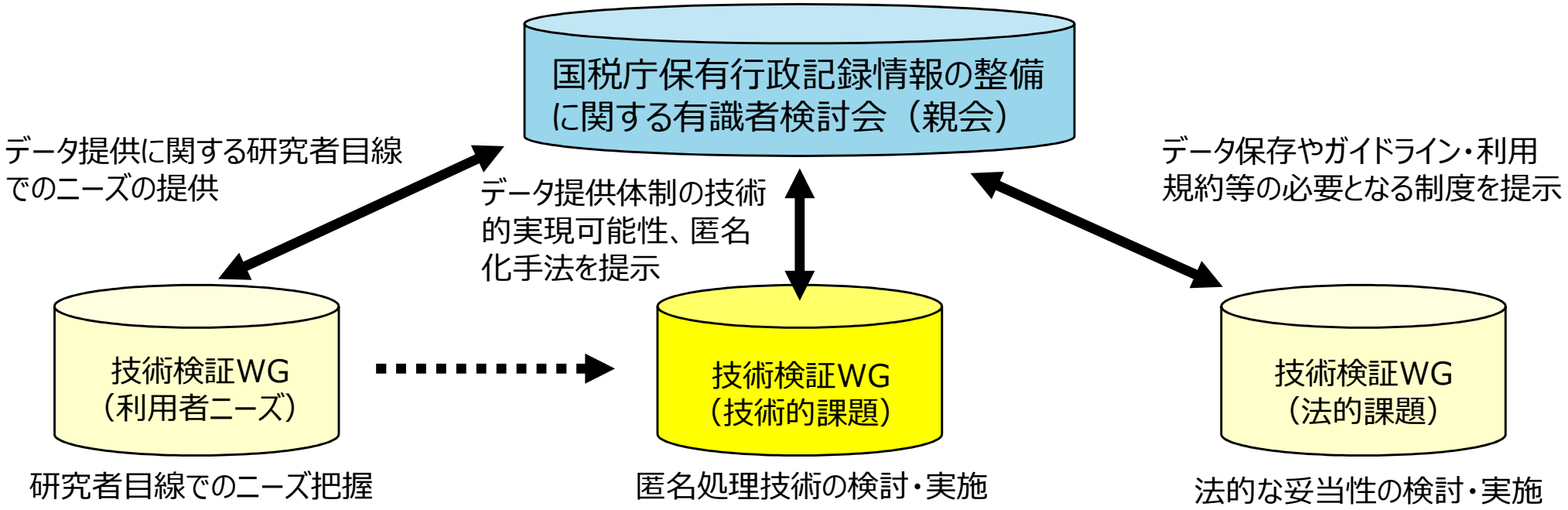
| | |
|---|---|
| 伊藤　伸介 | 中央大学　経済学部　教授 |
| 菅　幹雄 | 法政大学　経済学部　教授 |
| 星野　伸明 | 金沢大学　人間社会研究域　経済学経営学系　教授 |
| 南　和宏 | 統計数理研究所　データ科学研究系　教授 |

# １．本ワーキンググループの経緯・位置づけ

※令和４年10月14日開催
第２回技術検証WG資料の再掲

○　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。



# ２．これまでの議論（匿名データの保有及び公表の目的等）

※令和３年10月29日開催
第１回有識者検討会資料の再掲

## ● 匿名データの保有及び公表の背景

・ 国税庁保有行政記録情報を用いた税務大学校との共同研究（以下、共同研究）は、各府省庁が保有するデータは、公開することが適当でない情報であっても、限定的な関係者間での共有を図る「限定公開」とする「オープンデータ基本指針」を踏まえ、国税庁独自に有識者を交え検討を重ねた結果、まずは共同研究という形式から始めることが適切であるという結論が得られた。

・ 国税庁の税務データは、申告納税制度の下、納税者の信頼や協力によって集積しているものであることに留意し、適切に取り扱う必要がある。したがって、共同研究において個票データを利用する者は、守秘義務の観点から国家公務員の身分を有する者のみに限定する。

・ 一方で、国家公務員の身分を有することなく、かつ、より多くの研究者が税務データを分析することにもニーズがある。

・ 現状、共同研究において、分析結果等利用者は国家公務員の身分を有することなく、加工した税務データにアクセス可能であるが、このスキームを参考に、研究者等が加工したデータにアクセスできる仕組み（国税庁版SUF（Scientific Use Files、学術研究用ファイル）の可能性を検討する。

# ２．これまでの議論（政府保有データのオープン化に係る政府方針）

※令和３年10月29日開催
第１回有識者検討会資料の再掲

## ● オープンデータ基本指針（平成29年５月30日高度情報通信ネットワーク社会推進戦略本部・官民データ活用推進戦略会議決定　令和３年６月15日改正）抜粋

・　各府省庁が保有するデータはすべてオープンデータとして公開することを原則とする。

・　個人情報が含まれる、又は法人・個人の権利利益を害するおそれがある等の理由によりオープンデータとして公開することが適当でない情報であっても、支障のあるデータ項目を除いて公開すること、限定的な関係者間での共有を図る「限定公開」といった手法を積極的に活用する。

※ オープンデータとは、国・地方公共団体及び事業者が保有する官民データのうち、国民誰もがインターネット等を通じて容易に利用できるよう、①営利目的、非営利目的を問わず、②機械判読に適し、③無償で利用できるものとして公開されたデータをいう。

## ● 世界最先端デジタル国家創造宣言・官民データ活用推進基本計画（令和２年７月17日　閣議決定）抜粋

・ オープンデータの取組については、「オープンデータ基本指針」に基づき、利活用者のニーズを的確に反映しながら進めることが重要。

## ● 財務省デジタル・ガバメント中長期計画（平成30年６月25日　令和２年３月27日改定）抜粋

・ 保有データのオープン化については、データ連携・標準化等に関する政府の方針を踏まえ、個人情報保護、守秘義務等に関する法令を遵守しつつ、可能な限り、利用者ニーズを踏まえた行政保有データのオープン化を進める。

# ２．これまでの議論（整備ステップ）

※令和４年６月22日開催
第２回有識者検討会資料の再掲・一部修正

○　国税庁がシステム内で保有する税務データは、現状、原則７年経過後に削除することとしている。

○　令和３事務年度においては、提供データ（SUF/PUF※）の整備に先立って、長期間保存が可能となる、研究用行政記録情報（マザーデータ）を整備するに当たっての検討を進めてきたところ（整備第１ステップ）

（※）SUF：Scientific Use File、学術研究用ファイル、PUF：Public Use File、一般公開型ファイル

○　令和４事務年度においては、より具体的なデータ提供に向けて、どのような提供データを整備するか議論を進めているところ（整備第２ステップ）


【主な論点】
・　氏名、住所といった個体識別事項についてどの程度削除等を行う必要があるか。

【主な論点】
・　データ利用対象者の範囲や匿名化の程度、データの利用場所等をどのようにするか。
・　どのような形式のデータを整備していくべきか。

# 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４年10月14日開催
第２回技術検証WG資料の再掲

○　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<br>（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<br>（閲覧場所等の整備は不要、貸出作業は発生） | 高い<br>（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<br>（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<br>（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。<br>・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。<br>・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。<br>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<br>※ 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。<br>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）

# ２．これまでの議論　③サンプルデータについて

※令和４年10月14日開催
第２回技術検証WG資料の再掲

○　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<br>・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<br>(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）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 |

# （参考）第２回技術検証WG（令和４年10月14日開催）の議事要旨

## ○ データの提供形態について

・ 準備に時間がかかって、提供開始が遅れるよりは、まずは出来ることから取り掛かり、徐々に拡大していくべきである。
・ リモートエグゼキューションは、送付したプログラムの結果がエラーとなる可能性もあり、実効性が低いと考えられる。

## ○ 提供データの項目について

・ 他の統計情報から観察できない項目から優先して提供してはどうか。
・ 税制の研究においては、住所情報は市町村レベルまでは提供されることが研究の正確性を確保するためには望ましい。

## ○ サンプルデータについて

・ データ利用者がプログラムの正確性を確認する材料としては有用である。また、データ分析のニーズも高まっているため、自由な形で使えるデータがある方が望ましい。
・ 実際の税務データと関係のない架空のデータが想定されることから、その作成にあたっては、あまりコストや手間をかけるべきものではないように思われる。

## ○ データを利用できる者・利用目的の範囲について

・ リサーチアシスタントにも利用を認めるべきだが、利用者なのか研究協力者なのかによって、データへのアクセスについて整理する必要がある。
・ 利用目的の範囲は「税・財政施策に資すること」よりも広い範囲で認めてもよいのではないか。

# 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＞**

- ✓ 閲覧方式を念頭に置いた場合、分析結果持ち出しの安全性審査策定における検討課題について。
- ✓ 貸出方式を念頭に置いた場合、認識すべき具体的なコンプライアンスリスクや、それを下げるための方策（例：利用上の制約等）について。
- ✓ 上記のほか、閲覧・貸出方式それぞれにおける検討課題について。

**＜②施すべき匿名加工技法について＞**

- ✓ パーソナルデータ（個人）及びビジネスデータ（法人）に対する匿名加工を検討するに当たって、それぞれの特性に応じた課題について。
- ✓ サンプリング、識別情報の削除、トップコーディング、リコーディング、特異値の削除等、検討を要する匿名加工技法について（どのような匿名加工が有効と考えられるか）。
- ✓ 地理情報（住所情報）に係る匿名加工について。

**＜③サンプルデータについて＞**

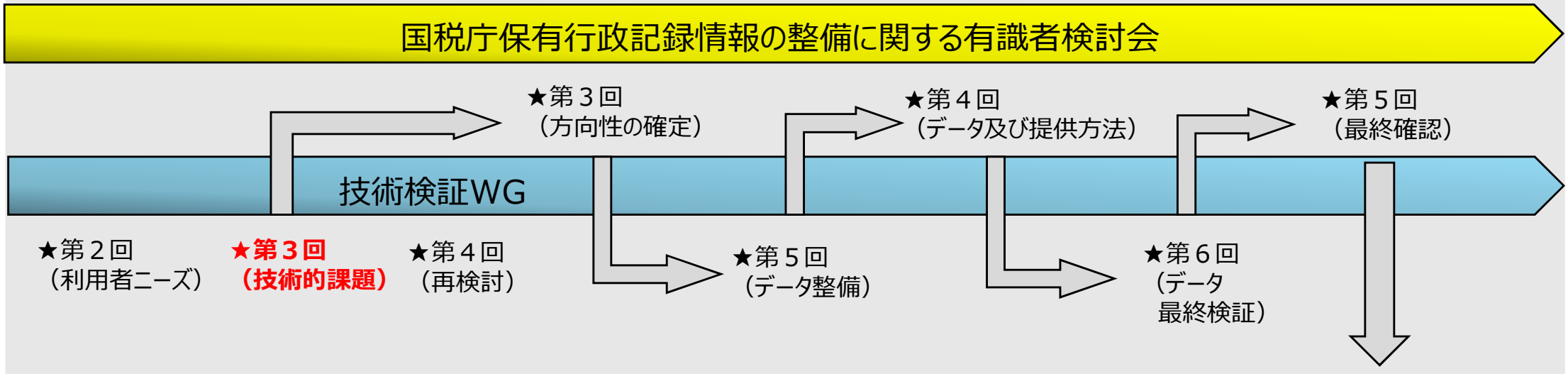
- ✓ 母集団と分布を近似させることを念頭に置いた場合のデータの生成可能性について。

# ４．今後のスケジュール（案）

○　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年度中に、準備が整い次第、対外的に行政記録情報の提供を開始することを目指す。

○　各WGにおける検証も踏まえつつ、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。

| 令和４年 | 令和５年 | | | | 令和６年 | |
|---|---|---|---|---|---|---|
| 10～12月 | １～３月 | ４～６月 | ７～９月 | 10～12月 | １～３月 | ４月～最終確認 |




**準備が整い次第、提供を開始する。**