資　料　２

保 存 期 間 ： ５ 年
（ 令 和 ８ 事 務 年 度 末 ）
令 和 ３ 年 １ ０ 月 ２ ９ 日

国税庁保有行政記録情報の整備に関する有識者検討会

国税庁　企画課

本日の資料内容

1. 匿名データの保有及び公表の目的

2. 当検討会の範囲

3. 検討の方向性

4. データ項目抜粋

5. 既存の統計作成作業フロー及び匿名化の状況

6. 今後のスケジュール

１．匿名データの保有及び公表の目的等

● 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。

（参考）政府保有データのオープン化に係る政府方針

● 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

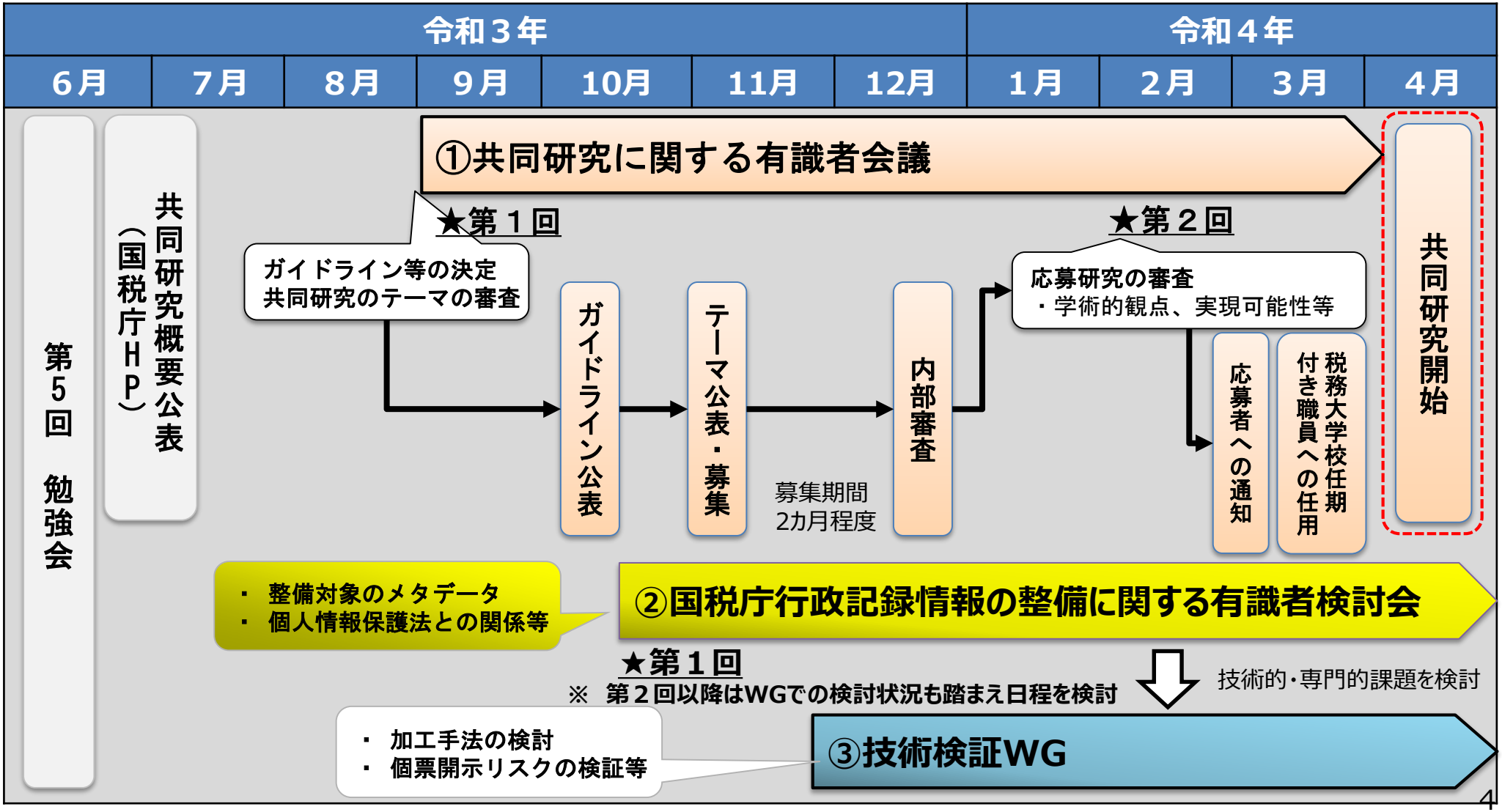
- オープンデータの取組については、「オープンデータ基本指針」に基づき、利活用者のニーズを的確に反映しながら進めることが重要。

● 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．当検討会の範囲

○　共同研究及び税務データのオープン化については、以下のスケジュールに沿って検討を進める。
○　国税庁行政記録情報の整備に関する有識者検討会（以下、当検討会）は、整備対象のメタデータ及び個人情報保護法との関係等が主な検討内容となる。状況に応じて、技術検証WGを設置する。



３．検討の方向性（１）匿名化ステップを進める前提としての整備（総論）

○　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。

○　今後、SUF や PUF（Public Use File、一般公開型ファイル）といった提供データの整備に先立って、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どのような論点があるか整理してはどうか。（整備第１ステップ）

○　その後、研究用行政記録情報の整備状況を勘案しつつ、どのような提供データ（SUF/PUF）を整備するか議論することとしてはどうか。（整備第２ステップ）

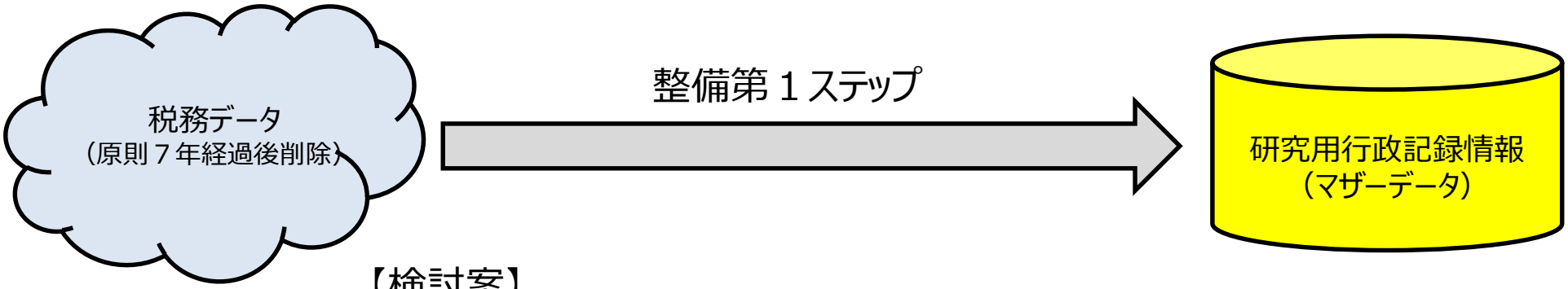
【想定される主な論点】
・　氏名、住所といった個体識別事項についてどの程度削除等を行う必要があるか。

【想定される主な論点】
・　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SUFやPUFとして整備していくべきか。

３．検討の方向性（２）整備第１ステップ：マザーデータの整備

○　研究用行政記録情報の整備に当たっては、税務データと遜色のないものを整備することを目指しつつ、
①　税務データが保有する個体識別情報項目（住所・氏名等）をどの程度削除する必要があるか
②　売上や所得項目が極端に高額であり識別が可能である可能性を踏まえ、削除ないし補正をする必要があるか
③　令和3年の個人情報保護法改正を含む法的観点や技術的観点等
から議論が必要。

○　なお、研究用行政記録情報の整備に当たって、まずは申告所得税関係書類（確定申告書１・３表、青色決算書、収支内訳書）から検討を始めることとしたい。



【検討案】
- 住　　所　→　市区町村レベルまでをデータで保持
- 氏　　名　→　削除
- 生年月日　→　すべてデータで保持
- マイナンバー→　削除

3．検討の方向性（３）整備第２ステップ：提供用データ（SUF/PUF）

○　研究用行政記録情報の整備の状況を前提としつつ、政府全体のオープンデータに係る方針等も踏まえ、国税庁においても提供データ（SUF/PUF）の整備を図っていきたいと考えているところ。

○　他省庁の取り組みや諸外国の状況も踏まえつつ、その第一ステップとして、SUFの整備を念頭に置き、議論を進める。

○　なお、将来的に、HP等での公開を前提としたPUFの整備についても検討を行う。

【検討のイメージ】

オープン化のステップ	共同研究 （ガイドライン公表済）	匿名化ステップ （SUFの整備）	匿名化ステップ （PUFの整備）
匿名化の度合い	匿名化なし	匿名化（低～中）	匿名化（高）
利用対象者	採択された研究の 研究者	当局に利用申請を 行った者	制限なし （HPで公開）
利用場所	税大和光	原則国税組織内施設利用/データ貸出	制限なし
データ持ち出し	不可	可	制限なし
利用可能年数	７年（拡充は検討）	７年（拡充は検討）	７年（拡充は検討）

	匿名化ステップ①	匿名化ステップ②	匿名化ステップ③
匿名化の度合い	匿名化（低）	匿名化（低～中）	匿名化（中）
利用対象者/制限	大学等の研究者のみ	民間シンクタンクを含む研究者	申請者すべて
利用場所	国税組織内施設利用のみ	原則国税組織内施設利用/データ貸出	原則データ貸出

3．検討の方向性（４）整備第２ステップ：提供用データ（SUF/PUF）

○ 「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。

３．検討の方向性（５）研究用行政記録情報の共同研究への活用

○　研究用行政記録情報の整備に当たっては、広く提供するデータ（SUF/PUF）の生成について今後議論すると同時に、共同研究における活用についても整理する必要。

○　現状、共同研究においては、国税庁保有の税務データを利用しているところ、研究用行政記録情報の整備後にはこちらの活用を前提としつつ、希望する場合には国税庁保有の税務データを利用しても差し支えないこととしてはどうか。

	税務データ	研究用行政記録情報	提供データ（SUF/PUF）
匿名化の度合い	匿名化なし	匿名化（極低）	匿名化（低～高）
守秘義務による制約	あり	あり	なし
他のデータとのリンケージ	可	不可	不可
パネル化	あり	あり	あり
利用可能期間	７年	7年以上	７年以上

共同研究においても活用

３．検討の方向性（６）今後議論いただきたいテーマ（案）

Ⅰ　研究用行政記録情報（マザーデータ）の整備について

現状、国税庁がシステム内で保有する税務データは、原則７年経過後に削除することとしている。
今後、SUFやPUFといった外部提供用データの整備に当たり、その基礎となる研究用行政記録情報（マザーデータ）を整備することが必要と考えられるところ、以下の項目について法的・技術的観点から検討を行いたい。

★　研究用行政記録情報の整備に当たり、税務データで保有する個人識別子（住所・氏名等）をどのように取り扱うべきか

★　個人識別子以外の項目について、例えば個人が特定される可能性がある項目（極端に所得が高い等）についてどのように取り扱うべきか

★　その他、検討を要するべき事項はあるか

Ⅱ　提供用データ（SUF/PUF）の整備について

「研究用行政記録情報」の整備を踏まえて、外部提供用データを整備していきたいと考えているところ、以下の項目について法的・技術的観点から検討を行いたい。

★　まずは、SUFの整備を目指すこととするが、データの利用対象者、利用範囲、利用場所等をどうするべきか

★　整備するSUFのデータ項目をどうするべきか

★　SUFの整備に当たって、どのような手法を用いて匿名化を実施すべきか

★　その他、検討を要するべき事項はあるか

４．データ項目抜粋

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

5．既存の統計作成作業フロー及び匿名化の状況

- 「調査票情報のオンサイト利用手引き」との相違点は特になし。

✓ 一次・二次秘匿等を実施し、安全性を担保。

統計表における秘匿措置 （「調査票情報のオンサイト利用手引き」と同等の規定）

1.集計区分の変更	各セルに集計される区分を変更して再度集計を行い、表1の内容を満たすようにすること。 （既存の区分の分割、他の区分と統合、新たな区分の設定等）	
2.集計対象の変更	集計対象の範囲を拡大又は縮小して再度集計を行い、表1の内容を満たすようにすること。 （調査客体グループの統合、外れ値の除外等）	
3.セルの値を秘匿	秘匿措置	申出者が提示する情報
	①一次秘匿 内容を満たさないセルの値を「X」などのマークに置き換え	秘匿前の統計表
	②二次秘匿 一次秘匿を行ったセルの値が他のセル等から算出できる場合、該当セルを全て「X」などのマークに置き換え	一次秘匿した各セルの位置を明示する情報
	秘匿インターバル 一次秘匿した各セルが取り得る値の最大と最小の差が、度数10以上ないし当該セル値の30%以上であること	一次秘匿した各セルが取り得る最大値、最小値及び両者の差ないし両者の差を当該セイルの値で除した割合

※参考：秘匿措置の例

外国法人の都道府県別法人数、所得金額

国税局・都道府県 ＼ 区分		外国法人 申告法人数	外国法人 所得金額
		社 Number	百万円 Million yen
仙台	青森	1	x
	岩手	3	x
	宮城	9	x
	秋田	－	－
	山形	1	x
	福島	7	x
	計	21	8
東京	千葉	97	2,269
	東京	3,976	425,544
	神奈川	227	9,550
	山梨	6	1,242
	計	4,306	438,605
福岡	福岡	31	33
	佐賀	1	x
	長崎	3	x
	計	35	36

（出所）国税庁統計年報報告書
「法人税」 平成30年度版一部抜粋

（参考）匿名加工の技法について

● 匿名加工の技法

- 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）

６．今後のスケジュール（案）

○　検討に当たっては、有識者検討会における行政記録情報の整備に係る全体的な方向性の議論を中心に進めることとし、令和3事務年度（令和3年7月～令和4年6月）内に中間とりまとめを行う。

○　なお、行政記録情報の整備については、中長期的な議論を要すると想定されるところ、全体的な方向性については令和４事務年度（令和4年7月～令和5年6月）中に整備の方向性について結論を得るとともに、令和5事務年度から具体的な整備を進め、令和6事務年度から対外的に行政記録情報の提供を始めることを目標とする。