令和６年度　最終評価様式

政府等保有データのAI学習用データへの変換に関する調査研究

令和７年５月
デジタル庁

● 実施する重点課題（特に該当するものには◎、そのほかで該当するものには○（複数可）を記載）

業務プロセス転換・政策転換に向けた取組	次期SIP/FSより抽出された取組	SIP成果の社会実装に向けた取組	スタートアップの事業創出に向けた取組	若手人材の育成に向けた取組	研究者や研究活動が不足解消の取組	国際標準戦略の促進に向けた取組
○						

● 関連するSIP課題（該当するものには○を記載）

持続可能なフードチェーン	ヘルスケア	包摂的コミュニティ	学び方・働き方	海洋安全保障	スマートエネルギー	サーキュラーエコノミー	防災ネットワーク	インフラマネジメント	モビリティプラットフォーム	人協調型ロボティクス	バーチャルエコノミー	先進的量子技術基盤	マテリアル事業化・育成エコ

1．社会実装に向けた施策・取組等の全体俯瞰の中での成果（進捗の説明）

① 全体概要

＜事業名変更に至った経緯＞
事業として計画の検討を開始したした時点では、クラウドでの機密性２情報の取扱いが困難であったためオンプレミスの使用を前提としていたが、その後機密性２情報の取扱い等、クラウドサービスでも十分対応できるようになった。それにより本事業においてオンプレミスの必要性は無くなり、事業者のサービス形態として主流であるクラウドサービスを利用して事業を実施する方が次々にリリースされる性能のより高いAIモデルを使用でき、将来的に継続性のある事業を展開することが見込まれるため、オンプレミスからクラウドサービスを使用することとした。

＜① 解決すべき社会課題＞
日本政府は課題先進国として抱える社会課題を解決する切り札として、データとそこから創出される付加価値・競争力に注目しており、データ利活用の更なる促進に向けて、2021年に「包括的データ戦略」、2023年には「デジタル社会の実現に向けた重点計画」を策定した。生成AIは、そうした社会課題の解決に資するデータ利活用のアプリケーションのひとつとして、研究機関やAI事業者による研究開発のもと、近年著しい発展を遂げている。生成AIの開発にあたっては、正確で新しい情報を含み、かつ不適切な情報が含まれない学習データが大量に必要とされており、特に日本語に対応する生成AIの研究開発には良質な日本語の学習データが必要となる。政府等が保有するデータの多くは、その正確性や権利、匿名加工処理が実施されている等を満たしており、学習データとして有用であることから、それらのデータ提供に対する期待がある。一方で、政府等が保有するデータは、そのデータ形式がpdf、JPEG形式など、直ちにAI学習に用いることが難しい場合も多く、またデータのアクセス権限などにより活用が難しいのものが散見され、その対応が必要となっている。

1．社会実装に向けた施策・取組等の全体俯瞰の中での成果（進捗の説明）

① 全体概要

＜② 取組施策の内容＞

生成AIの学習データとして政府等（中央省庁、地方公共団体、地方自治体及びその関係機関等）が保有するデータを活用することで、日本語に対応した生成AIの研究開発のさらなる促進が期待されるものの、現在政府が公開するデータは必ずしも直ちにAI学習への活用が可能な形式・環境で公開されていないため、研究機関やAI事業者が生成AIの学習に活用しやすいデータを公開するプロセスの構築等について調査研究を行った。

しかし、計画書を作成した時点から、マルチモーダルLLMの性能が飛躍的に向上したことから、一般的なWord・Excel、PDFファイルの機械判読は基本的に可能となった。そのため、事業開始時点でも特に機械判読困難であり、官報データのような日本独特であり政府等の資料でよく用いられる複合図表文形式のデータにフォーカスを絞って研究を行うこととした。

また、事業者や有識者へのヒアリングにより、日本の法令に関するデータセットのニーズが特に高いことが判明したため、研究対象とした。

加えて、データ公開方法については、独自のプラットフォームを整備するよりも、一般的に広く利用されている民間のプラットフォーム（GitHub、Hugging Face等）に掲載する方が、利用者のアクセシビリティが高いことが分かったため、民間のプラットフォームを活用する方針とした。

調査研究の結果、法務向けデータや図表と文の複合データなど、特定の目的に絞ったデータは有効性が高く、国内AI事業の成長には業務や技術に特化したデータ公開が重要であることが分かった。また、政府のデータ準備において、評価データは仮説検証サイクルを早め、開発の敷居を下げるため最優先されるとの認識に至った。加えて、AIによる業務改善部分を開発者と共有し、業務の専門家の助言を仰ぎながら学習データと評価モデルの業務適合性を共同で検証することが重要である。

＜③ 成果の社会実装＞

調査結果を踏まえ、まずは令和8年度から生成AIの評価用データセットの作成・提供に関する事業の開始に向けて検討を進めている。

作成した評価用データセットを民間のプラットフォーム上に公表することで、政府等が生成AIを調達する際の選定基準に、当該データセットにおける正答率を条件に付すことで、調達目的により則した性能の高い生成AIを調達することが可能となる。また、事業者においては、当該データセットにおける正答率が政府等で使用される生成AIの基準となることから、自社の生成AIの性能向上の基準にもなり、性能検証もしやすくなることから、開発が促進されることが期待できる。

このように、生成AIの性能を政府から品質が高い日本語によるデータセットが公開されることで、事業者や研究者による利用が進むことにより、国外との競争力強化が促進されることを見込む。

1．社会実装に向けた施策・取組等の全体俯瞰の中での成果（進捗の説明）

② 全体俯瞰図

解決すべき社会課題

現状は海外に比べてデータ量が不足

公開されている政府等保有データへのアクセス方法がわかりづらい

公開されている政府等保有データの機械可読性が低い

日本語に対応する生成AIの研究開発のために大量に必要な正確で新しく不適切な情報が含まれない良質な日本語の学習データの不足

取組施策の内容

①政府等保有データの利活用促進のための基盤技術※開発
データ公開の準備作業を一通り流していく中で、実際にどのようなワークや課題が発生するかを検証し、データ管理、公開のプロセス化に向けたノウハウを整理

②持続可能なデータ提供スキームの構築
データ公開に至るプロセスをロール別に細分化し、技術的な作業だけでなく、ライセンス等の権利関係の確認等もプロセスに含め、信頼性の高い評価用データセットの利活用が促進されるスキームを構築

期待効果・To-be像

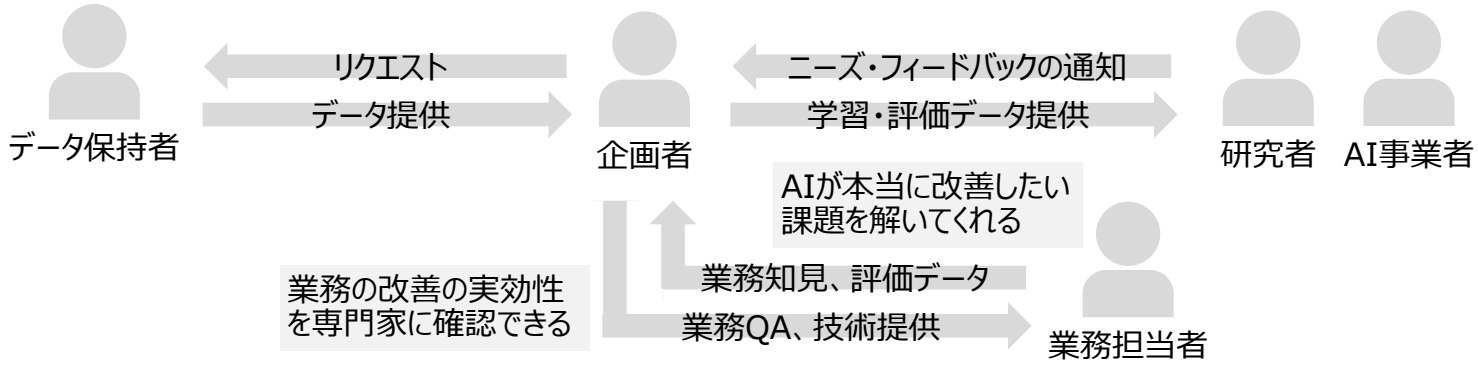
データ公開を企画する省庁・自治体、データ保持者である関連省庁、研究者・AI事業者、業務担当者が以下のとおり対応

どんなデータをどんな形式で持つと役に立つかが分かる

データソースの改善により効率的、効果的なデータ公開が可能になる

不足しているデータや推進すべき研究領域を把握できる

AI提供による業務効果、データ公開の意図が分かる



※社会実装に必要な基盤の基礎技術を「基盤技術」と呼ぶ

2．研究成果及び出口戦略、達成状況（取組全期間）

テーマ ① 政府等保有データの利活用促進のための基盤技術開発

① 研究成果及び達成状況

取扱うデータについては、本事業の企画段階の前提にあった「良質な日本語テキストの公開」という目標に対して、現時点の生成AIのトレンドから、日本固有の背景や情報に答えられるデータを持続的に公開することが重要であると考えた。本事業を、限られたリソースの中で効率的かつ継続的に実行すべく、データに優先度をつける概念整理、優先度の高いものを持続的に公開するプロセスの仮定と、それらの実践を行った。
題材として、「法解釈、条項特定の多肢選択問題」「契約書 要約結果の品質判定問題」「複合図表文問題」「日本語特有レイアウト認識問題」のニーズを発見し、これらの評価用データセット作成に必要な技術的な課題を整理し、その解決手法を検証した。例えば、多肢選択問題では、複数の生成AIで一貫して不正解だった場合、人手で作成した正答の方に誤りが混入される可能性が高い、といった誤り検出方法の有用性を示した。

② 出口戦略・研究成果の波及

研究結果を基に令和８年度からの評価用データセットの作成及び公開を目指し、令和7年度において省庁等のステークホルダーとの連携体制構築のため協議を進めていく。
データ変換に関する調査研究結果をデジタル庁ウェブサイトに公開し、そのノウハウを共有することにより、政府等だけでなく産学の各組織においてもデータセット作成が促進されることを見込む。

③ 目標達成状況等の特記事項

法令解釈等に関しては、一般的な解説書等の解説や多肢選択問題集の情報が非常に有効であるが、著作権は出版社に帰属することからデータセットに盛り込むことができない。

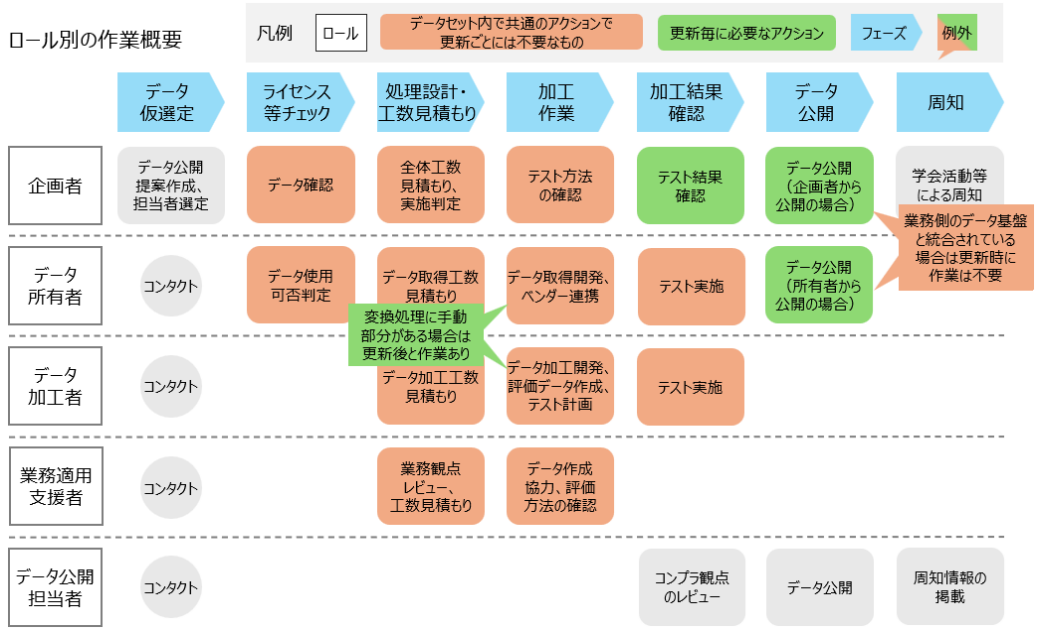
2．研究成果及び出口戦略、達成状況（取組全期間）

テーマ ② 持続可能なデータ提供スキームの構築

① 研究成果及び達成状況

データ公開における手法として、特定の組織による集中管理・公開とデータ保持組織による個別管理・公開の２パターンを想定し、データ提供スキームはどちらのパターンでも適用できる手法を検討した。データ公開に至るプロセスをロール別に細分化し、各作業の概要をまとめた。技術的な作業だけでなく、ライセンス等の権利関係の確認や、データ公開に対する周知活動についてもプロセスに含め、信頼性の高い評価用データセットの利活用が促進されるスキームとした。

ロール別の作業概要



② 出口戦略・研究成果の波及

当該データ提供スキームをベースとし、テーマ①における各ステークホルダーとの評価用データセットの作成・公開の実現に向けて、関係府省庁等の各ステークホルダーとの協議を進めていく。
政府等においてデータ提供の手法の共通化や民間のプラットフォーム利用の統一化等、データ利用者が求めているデータへのアクセス性、データの利便性を向上させる。
また、当該データ提供スキームは、本事業に限らず、データ提供の手法として汎用性があるため、調査研究結果の報告書を公開することにより、政府等以外の組織においても、データ公開が促進されることを見込む。

③ 目標達成状況等の特記事項

必要なプロセスやスキル要件を明確にした状態であり、これを運用するための体制構築は2025年度以降の課題である。

3．到達目標（KPI）に対する実績

テーマ名	実施内容の概要と 到達目標（KPI）	到達目標（KPI）に 対する実施内容と実績
① 政府等保有データの利活用促進のための基盤技術開発	政府等が保有するデータのうち、直ちにAI学習に用いることのできない、いわゆるマシンリーダブルでないデータをマシンリーダブルに変換するとともに、それらのデータを、よりAIモデル学習で使いやすいデータセットとして成形、管理するためのAIデータ整備基盤技術を開発する。	前述のとおり、特に機械可読困難な複合図表文をメインスコープに据えて変換方法の研究を行った。 作成した「法解釈、条項特定の多肢選択問題」「契約書 要約結果の品質判定問題」「複合図表文問題」「日本語特有レイアウト認識問題」のデータセットを使用し、複数の生成AIに対して性能評価を行った。作成したデータセットの形式はJSONである。 一般的な生成AIとしての性能評価に加え、法令検索・契約書要約というタスク特化の性能につき次の評価を実施した。また、各検証方法に対して異なる評価方法を採用した。 ● **汎用的な性能評価** 　➢ **Jaster** ● **法令検索タスクに対する性能評価** 　➢ **exact match** 　➢ **llm_jp_evalのchar_f1** ● **契約書要約タスクに対する性能評価** 　➢ **LLM as a judge** 　➢ **専門家による手動採点（部分点方式）** また、本事業において作成した「法解釈、条項特定の多肢選択問題」の評価用データセットの公開を予定している。 なお、＜事業名変更に至った経緯＞に記載のとおり、本事業においてはオンプレミスで実施する必要性が無く、データセットを作成する上で生成AIを利用することはあるが、特定の生成AIである必要はない。そのため、どの生成AIであっても事業を継続展開できるよう、AIデータ整備基盤として個別のシステムを構築するのではなく、クラウドサービスを利用し、データ公開の準備作業を一通り流していく中で、実際にどのようなワークや課題が発生するかを把握し、データ管理、公開のプロセス化に向けたノウハウの蓄積を行うこととした。

3．到達目標（KPI）に対する実績

テーマ名	実施内容の概要と 到達目標（KPI）	到達目標（KPI）に 対する実施内容と実績
② 持続可能なデータ提供スキームの構築	持続的に政府等保有データを提供・管理するためのスキームを構築するとともに、データセットを教師データとする生成AIの開発とその活用推進を支援する仕組みを構築する。	生成AIの進化は速く、事業計画時と比べ、生成AIの性能が非常に高くなり、また新たなサービスが次々にリリースされているため、特定の生成AIの開発を行うのではなく、汎用的なデータセットの作成方法の調査研究に変更した。 評価データは仮説検証サイクルを早め、実用性の担保にも直接的につながるため、最優先で検討すべき一方、評価データ作成はミスの発生しやすい作業であるため、基盤モデルを用いた答え合わせ等、品質担保のための自動化、共通化された仕組みの検討が重要であることがわかった。 加えて、データ公開方法については、独自のプラットフォームを整備するよりも、一般的に広く利用されている民間のプラットフォーム（GitHub、Hugging Face等）に掲載する方が、利用者のアクセシビリティが高いことが調査により判明したため、民間のプラットフォームを活用する方針とした。 その作成方法を主軸に沿えた、データの選定から、データの周知までの一連の業務プロセスを構築した。

4．実施体制及び実施者の役割分担

実施体制

AI戦略チーム

データPT

内閣府

デジタル庁

PD

デジタル庁
PD：山田政幸
データユニット長

テーマ１

政府等保有データの利活用促進のための基盤技術開発

テーマ２

持続可能なデータ提供スキームの構築

デジタル庁　森参事官

①生成AI 利用促進のためのデータ利活用に関する技術動向等調査

② 政府等に限らず民間企業等においても業務効率化・国民サービスの向上に資するようなオープンデータに関する整理

受託者：EYストラテジー・アンド・コンサルティング株式会社

- 技術的検証とその関連する技術などについての議論への参加
- 技術的検証およびその周辺領域を対象とする資料に対する情報提供
- 追加学習用データ確認・修正・試行的追加学習

再委託先：株式会社PreferredNetworks

③ 政府等保有データ等のAI 学習データへの変換に関する調査研究

④ 課題・示唆の抽出及び分析

⑤ 持続的に政府等保有データを提供・管理するためのスキームの検討

受託者：EYストラテジー・アンド・コンサルティング株式会社

データ構造化、画像認識、OCR等の技術動向や活用事例等、再委託先が知見をもつ領域を中心に、特定のテーマ（複数）に対する調査等

再委託先：株式会社シナモン

⑥ 生成AI に関連する規格・基準、法令・ガイドライン等の調査・仕様検討等

受託者：EYストラテジー・アンド・コンサルティング株式会社

専門的見地からの示唆・アドバイス、対象地域における法令の調査及びそれに基づく示唆・見解の提示

再委託先：EY弁護士法人

⑦ 政府に設置しているAI 学習用データに関するコミュニケーション窓口に係る業務

受託者：EYストラテジー・アンド・コンサルティング株式会社