

## 複数のデータ連携基盤にまたがるデータ利活用の課題に関する一検討 A Study on the Issues of Data Utilization Across Multiple Data Collaboration Platforms

石井 陽介<sup>†</sup>  
Yohsuke Ishii

伊藤 大生<sup>†</sup>  
Hiroki Ito

### 1. はじめに

データ連携基盤とは、複数のシステムやサービス間で蓄積されたデータを効率的に収集・変換・管理・利用する仕組みを提供するものである。このデータ連携基盤は、行政機関間連携、官民連携、および民間サービス同士の連携などの目的に応じて複数存在している。例えば、DSA(データ社会推進協議会)[1]が設立・運営する、分野横断的なデータ交換のためのプラットフォームである DATA-EX[2]や、内閣府の戦略的イノベーション創造プログラム(SIP)において分野間データ交換プラットフォームとして開発された CADDE(Connector Architecture for decentralized Data Exchange)[3]や、欧州における企業間データ連携基盤の枠組み、仕組みづくりを担っている Gaia-X[4]が提供するプラットフォームなどがある。

それぞれのデータ連携基盤でデータ形式や取り扱い方が異なることがある。一つのデータ連携基盤内でデータ活用する場合は問題ないケースであっても、図 1 に示すような複数の異なるデータ連携基盤からデータを取得して活用する場合は、当該データ形式や取り扱い方の違いへの対応が課題となる。本稿では、複数のデータ連携基盤に跨るデータ利活用の課題と、課題解決方針について報告する。

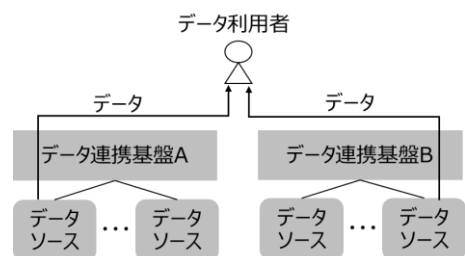


図 1 複数のデータ連携基盤を利用したデータ活用

## 2. 複数データ連携基盤に跨るデータ利活用の課題

### 2.1 データ交換プロセス

データ利用者がデータ活用を行うためには、データ提供者からデータを取得する必要がある。データ連携基盤では、このデータ取得を実現するために、データ提供者とデータ利用者との間のデータ交換を実現する仕組みを提供する。例えば、CADDE では、データ交換を実現するために表 1 に示す 5 つのプロセスを示している[5]。なお、ここで示すプロセスは、単一データ連携基盤内でのデータ交換ケースに限らず、複数データ連携基盤に跨るデータ交換ケースにもあてはまる。また、CADDE 以外のデータ連携基盤を利用するケースにおいても、同様のデータ交換プロセスになると考える。

表 1 CADDEにおけるデータ交換プロセス

#	名前	説明
1	データ準備とデータカタログ作成	データ提供者が、提供データを準備し、データカタログを作成する。
2	データ利用企画時データカタログ確認	データ利用者が、データ利用の企画を行い、必要なデータを、カタログ横断検索サービスで検索して特定して、データの取得判断を行う。
3	データ利用契約	データ利用者は、データ提供者と調整し、取得したいデータの利用契約を調整・締結する。データ提供者は、データ利用者が当該データの取得を可能にする認可設定を行う。
4	データ交換と利用	データ利用者は利用者コネクタを使用したデータ提供依頼を提供者コネクタに行い、提供者コネクタは認可設定を確認し、指定のデータ判断を行い、認可されていれば提供する。利用者は取得データを利用する。
5	来歴登録と確認	データの出自や利用状況を確認するため来歴管理サービスに履歴を登録する。来歴の確認要求時に来歴を検索して確認する。

### 2.2 データ取得判断

表 1 で示したデータ交換プロセスの 2 ステップ目「データ利用企画時データカタログ確認」において、データの取得判断を行う際に、当該データの「取得判断用情報」の確認が必要になる。取得判断用情報の例として、データ提供に関するデータ所有者の本人同意の有無や、データ利用条件、制限事項などがある。この取得判断用情報は、理想的には統一化・標準化されていることが望ましい。しかし、実際はデータソース/データ連携基盤毎に異なる。

本稿執筆時点で CADDE の API 仕様[5]や、Gaia-X 等のデータ連携基盤における API 設計で参照されている Dataspace Protocol[6]の API 仕様[7]においては、取得判断用情報に相当する情報について詳細な規定はない。それぞれの API 仕様では、データカタログサービスとその API 仕様が開示されており、例えば、データカタログ情報の一部として取得判断用情報を含めることは可能と考える。しかし、当該情報のフォーマット等はデータソース/データ連携基盤毎に異なる。

### 2.3 課題

取得判断用情報の統一化・標準化には時間がかかることが想定されるため、それまでのつなぎ対応が必要になる。この対応向けとして、2つのアプローチがある。

<sup>†</sup>株式会社 日立製作所 研究開発グループ  
Hitachi, Ltd., Research & Development Group

第 1 のアプローチは、データ利用者が各データの取得判断情報を個別に取得し、当該情報のフォーマットや情報量の違いを意識して手作業で確認するものである。第 2 のアプローチは、取得判断用情報の共通フォーマットを暫定的に策定し、当該フォーマットに合わせて取得判断用情報を取得・加工・突合する機能(データ取得判断支援機能)をデータ連携基盤で提供するものである。前者のアプローチは、データ利用者の作業負担が大きく、データ利活用の促進を阻害しかねない。後者のアプローチは、データ利用者の作業負担は軽減できるものの、取得判断用情報を取得・加工・突合する機能を実現する工数が大きく、データ連携基盤における多種多様なデータソースとの連携を阻害しかねない。また、後者のアプローチは、複数の異なるデータ連携基盤に跨るデータを対象とする場合、各データ連携基盤における取得判断用情報のフォーマットの違いにも対応する必要があり、さらなる工数増大要因になると考える。

データ利活用を促進するためには、後者のアプローチが望ましい。そこで、データ取得判断支援機能について、多種多様なデータソースの追加に柔軟に対応でき、かつ少ない工数で実現することが課題となる。

### 3. 課題解決方針

課題解決方針として、OpenAI 社 [8] が提供する ChatGPT [9] などの生成 AI を活用することを提案する。データ取得判断支援機能について、従来方式と提案方式を図 2 に示し、以下に説明する。

従来方式は、各データソースから取得判断用情報を取得する情報取得機能、取得情報を暫定共通フォーマットに対応付ける加工・突合機能、およびユーザインタフェースで構成される。データ利用者は、ユーザインタフェースを介して、利用データ候補群の取得判断用情報の提供を依頼する。データ取得判断支援機能は、構成機能群を利用して当該情報を提供する。ここで、加工・突合機能では、データソースならびにデータ連携基盤の違いにより多種多様な形式が存在する取得判断用情報を突合するために必要な各種定義情報等を持つ。データソースならびにデータ連携基盤の種類と扱うデータの種類の比例して、この定義情報ならびに定義情報を利用した処理プログラムの作成ならびに保守の作業工数が大きくなる。

提案方式は、加工・突合機能のかわりに生成 AI 問合せ機能を提供する。生成 AI 問合せ機能では、生成 AI のモデルに対して、情報取得機能が取得した各データソースの取得判断用情報群を入力に、取得判断用情報に所望の加工・突合をした結果を出力するよう問合せを行う。近年の生成 AI の飛躍的な精度向上が、この問合せ対応の実現に貢献している。利用する生成 AI のモデルについては、既存の公開サービスを利用する。また、生成 AI 問合せ機能では、生成 AI のモデルに対して、上記の問合せに必要なプロンプト定義情報等を持つ [10]。このプロンプト定義情報について、各々の取得判断用情報の具体的な違いを意識して定義する必要はなく、加工・突合時における各種条件や出力形式等を定義すればよい。そのかわりに、利用する生成 AI サービスが、入力情報と当該サービスのモデルを利用して、問合せ時で指定された内容を基に、各々の取得判断用情報の加工・突合の役割を担う。

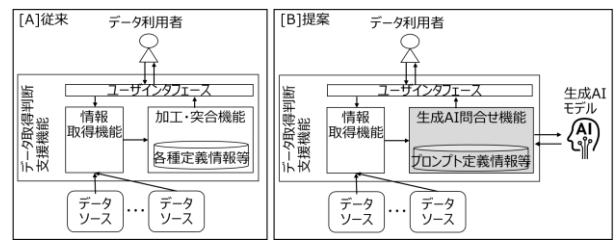


図 2 データ取得判断支援機能の構成

従来方式と提案方式の比較について述べる。従来方式は、加工・突合機能における定義情報ならびに処理プログラムの作成ならびに保守の作業工数が必要で、利用するデータソースならびにデータ連携基盤の種類が増えるたびに当該作業工数も増加する。提案方式は、プロンプト定義情報ならびに処理プログラムの作成ならびに保守の作業工数が必要であることは従来方式と同じである。しかし、当該プロンプト定義は、各々の取得判断用情報にあわせたカスタマイズが不要であり、利用するデータソースならびにデータ連携基盤の種類が増えても作成ならびに保守作業はほとんど発生しないと考える。したがって、提案方式は、多種多様なデータソースの追加に柔軟に対応でき、かつ少ない工数で実現することが可能で、データ取得判断支援機能向けに好適であると考えられる。

### 4. おわりに

複数のデータ連携基盤に跨るデータ利活用におけるデータ取得判断の必要性、データ取得判断支援機能の課題を示した。課題解決方針として生成 AI を活用する方針を示し、データ取得判断支援機能の提案方式を示した。今後の課題として、具体的なユースケースに基づく提案方式の試作ならびに検証、複数のデータ連携基盤を利用するデータ利用者へのデータ取得判断支援機能の提供方式の検討(データ連携基盤の提供機能にするか、外付けの機能にするか等)がある。

### 参考文献

- [1] 一般社団法人データ社会推進協議会。  
<https://data-society-alliance.org/>
- [2] 越塚 登, “社会環境変化への対応を支えるデータ基盤: 国内外のデータ基盤の現状・展望と DATA-EX,” 計測と制御, 62 巻, 4 号, pp.197-202 (2023).
- [3] SIP 分野間データ連携基盤. <https://sip-cyber-x.jp/>
- [4] Siska, Veronika, Vasileios Karagiannis, and Mario Drobnic. "Building a Dataspace: Technical Overview." (2023). <https://www.gaia-x.at/wp-content/uploads/2023/04/WhitepaperGaiaX.pdf>
- [5] CADDE V4 外部仕様書。  
<https://github.com/CADDE-sip/documents/tree/master/doc/4>
- [6] Dataspace Protocol – ensuring data space interoperability.  
<https://internationaldataspaces.org/dataspace-protocol-ensuring-data-space-interoperability/>
- [7] Dataspace Protocol 2024-1.  
<https://github.com/International-Data-Spaces-Association/ids-specification/tree/main/releases/2024-1>
- [8] OpenAI 社. <https://openai.com/>
- [9] ChatGPT. <https://chatgpt.com/>
- [10] OpenAI Platform, Prompt engineering.  
<https://platform.openai.com/docs/guides/prompt-engineering>