ライフサイエンスデータベース統合推進事業（統合化推進プログラム）
研究開発実施報告書　様式

# 2024 年度 研究開発実施報告

## 概要

| | |
|---|---|
| 研究開発課題名 | 創発的再解析のためのメタボローム統合データベース |
| 開発対象データベースの名称（URL） | integMET<br>(https://未定) |
| 研究代表者氏名 | 早川 英介（20739809） |
| 所属・役職 | 理化学研究所　環境資源科学研究センター　客員研究員<br>（2025 年 3 月時点） |

# □目次

# §1. 研究実施体制

| グループ名 | 研究代表者・研究分担者氏名 | 所属機関・役職名 | 研究題目 |
|---|---|---|---|
| 早川（理研）グループ | 早川 英介 | 理化学研究所・客員研究員 | integMETデータベースの解析機能とインターフェースの開発 |
| 河野グループ | 河野 信 | 北里大学・教授 | integMETデータベースのRDF化・知識グラフの作成 |
| 津川グループ | 津川 裕司 | 東京農工大学・教授 | integMETデータベースのグラフ埋め込み（エンベディング）・AI活用 |
| 早川（九工大)グループ | 早川 英介 | 九州工業大学・准教授 | integMETデータベースの搭載データ拡張およびそのためのシステム構築 |

# §2. 研究開発対象とするデータベース・ツール等

## (1) データベース一覧

【主なデータベース】

| No. | 名称 | 別称・略称 | URL |
|---|---|---|---|
| 1 | integMET | | 未定 |

【その他のデータベース】

| No. | 名称 | 別称・略称 | URL |
|---|---|---|---|
| 1 | | | |

## (2) ツール等一覧

| No. | 名称 | 別称・略称 | URL |
|---|---|---|---|
| 1 | | | |

# §3. 実施内容

## (1) 本年度に計画されていた研究開発項目・タスク

【項目1】解析機能とインターフェースの開発
　【項目 1-1-1】「メタデータによるフィルタリング機能の開発」
　【項目 1-1-2】「ネットワーク解析機能の開発」
　【項目 1-1-3】「エッジ属性情報の付与」
　【項目 1-1-4】「データ要約機能の開発」

【項目 2】データの RDF 化および他のデータベースとの連携
　【項目 2-1-1】「使用するオントロジー・RDF スキーマの検討・設計」
　【項目 2-2-1】「接続先の外部 DB の検討」

【項目 3】統合データのエンベディング生成と機械学習・AI との連携
　【項目 3-1-1】「ノード分類機械学習ワークフローの開発（上流工程)」
　【項目 3-1-2】「ノード分類機械学習ワークフローの開発（下流工程)」

【項目 4】搭載データ拡張およびそのためのシステム構築
　【項目 4-1-1】「Metabolomics Workbench のデータ前処理」
　【項目 4-1-2】「Metabolomics Workbench の代謝変動ネットワーク統合」
　【項目 4-1-3】「Metabolomics Workbench のメタデータネットワーク統合」

## (2) 進捗状況

### 概要

本年度は「創発的再解析のためのメタボローム統合データベース (integMET)」の基盤構築として、4つの主要項目に関する開発を実施した。解析機能とインターフェースの開発では、メタデータを活用したフィルタリング機能やネットワーク解析機能を実装し、ユーザーが直感的に大規模なメタボロームデータグラフデータを探索できる環境基盤を整備した。データの RDF 化および外部データベースとの連携に向けては、標準オントロジーの選定とスキーマ設計を行い、今後の拡張性を確保した。統合データのエンベディング生成と機械学習連携では、ノード分類のための機械学習ワークフローを構築し、データ駆動型の解析基盤の検討を行った。さらに、搭載データの拡張として、Metabolomics Workbench からのデータ統合により、すでに処理済みの Metabolights と合わせ現在の2大メタボロームレポジトリ両方を包含する大規模ネットワークを構築し、代謝物変動情報と研究メタデータを体系的に整理した。これらの取り組みにより、異なる研究間の代謝変動パターンの類似性を視覚化・探索できる統合プラットフォームの基盤が完成した。

以下に、各項目の実施内容の進捗について報告する。

### 【項目 1】「解析機能とインターフェースの開発」

#### 【項目 1-1-1】「メタデータによるフィルタリング機能の開発」

搭載したデータからユーザーの目的・興味に基づいて重要な研究（スタディ）のサブセットを抽出できるフィルター機能の開発を行った。具体的には分析手法（LC-MS、GC-MS）、試料の種類、代表的な研究

メタデータ（MeSH Term）を用いてグラフ上のスタディのノードを抽出する機能の実装を行った。この機能は integMET の Web ユーザーインターフェースに統合され（図 1）、ユーザーが容易に操作・可視化を行うことが可能となった。

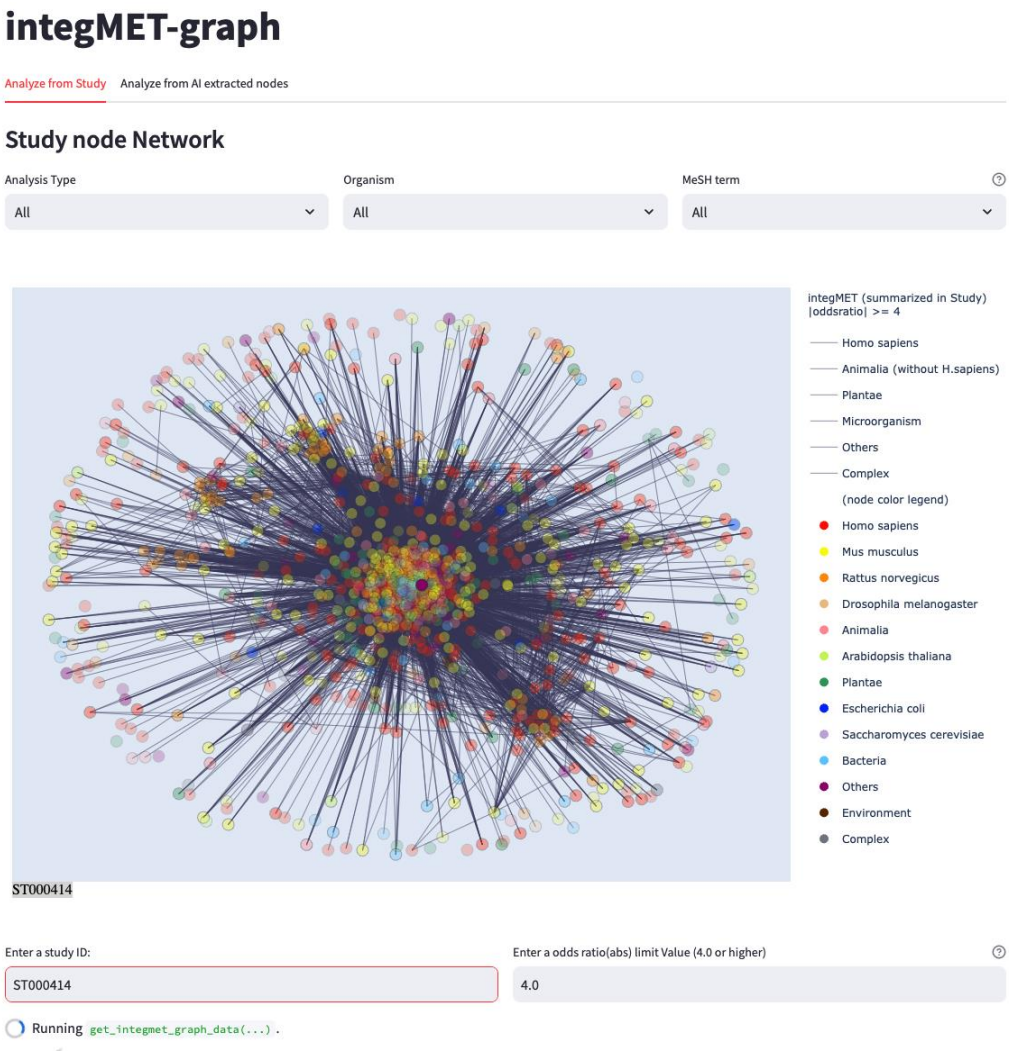


**図 1 integMET の開発中のユーザーインターフェース**

**【項目 1-1-2】「ネットワーク解析機能の開発」**

本データベースではレポジトリから収集した各スタディにおいて、比較定量テーブルと試料情報から、スタディにおける群間比較を組織的に生成する。群間比較は例えば「"コントロール群"対"薬剤投与群"」のような"群"対"群"での代謝物の変動の傾向を示すデータ構造となり、この群間比較をスタディを横断して類似度を計算することで研究横断的な代謝変動の類似性をネットワーク化している。このネットワーク構造を活用した機能として、ユーザーが指定したスタディのノード（エゴ）とその直接的な隣接ノード（アルター）、およびそれら周辺のノードから成るサブネットワークを抽出する機能を実装した。このエゴセントリックネットワークの抽出機能は、特定の興味のあるスタディと代謝変動が類似したスタディの集合体を抽出することで、複雑なネットワーク構造をもとにした解析を容易にする。加えて、代謝変動の類似度で繋がったノード間での研究メタデータの類似性をグラフ構造として可視化する機能の実装も行った（図 2）。この機能により、代謝変動の類似度で捉えたスタディのペア間の研究背景の類似性・違いをインタラクティブな可視化を通して直

感的に解析することが可能となった。

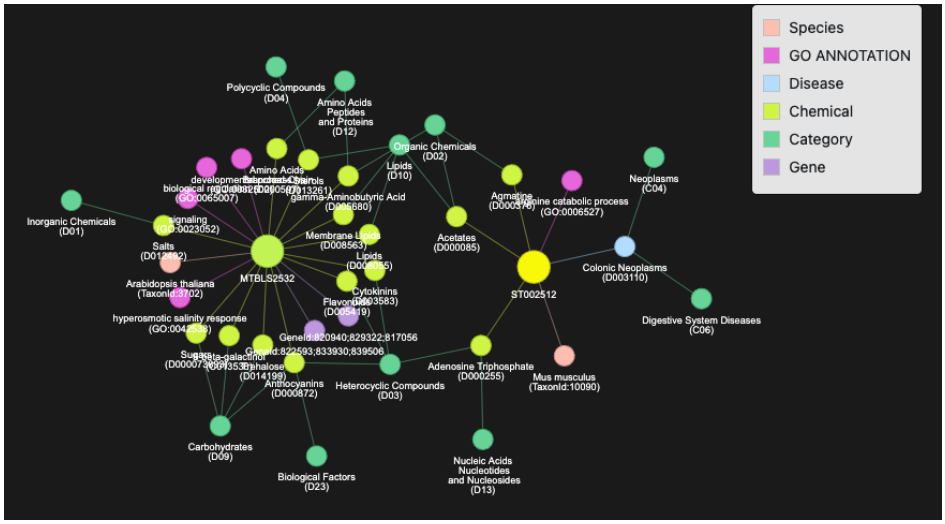


図 2 integMET インターフェース上でのスタディ間の研究類似性の可視化

### 【項目 1-1-3】「エッジ属性情報の付与 」

本グラフデータベースのデータ構造のエッジは、スタディにおける群間比較の代謝物の変動の類似性をオッズ比で定量化したものであり、エッジ自体に代謝物に関する情報は含まれていなかった。しかし、複数のスタディにまたがる統合解析をする際には代謝物に関する具体的な情報が必要となるケースが考えられる。そこで各エッジの代謝物変動の類似性に寄与している(スタディ間の群間比較ペアで代謝物シグナルの増減が一致する)代謝物に関して、元の比較定量データから収集しエッジの属性情報として付与されたデータベースの仕様を検討し、実装を行った。この開発により代謝物レベルでの研究間での共通性といったより詳細な解析が可能となった。

### 【項目 1-1-4】「データ要約機能の開発」

スタディおよび群間比較に関して、レポジトリの研究テキスト情報から研究情報を要約し、グラフデータベース上の属性情報として付与した。ユーザー解析のインターフェースで選択したノードの要約情報を表示させる機能を実装した。また、選択されたノードペアに関して、共通する研究メタデータをグラフで可視化する機能も実装した。これらの機能により、ユーザーが膨大なネットワークから解析・抽出した情報の解釈を視覚的に容易にすることができた。

## 【項目 2】データの RDF 化および他のデータベースとの連携

### 【項目 2-1-1】「使用するオントロジー・RDF スキーマの検討・設計」

本年度は、統合データベースを RDF 化し、外部リソースと相互運用可能な形式での公開基盤を整備するため、まず内部データ構造を忠実に表現するスキーマを設計した。このスキーマ設計において最優先すべきはオッズ比算出にいたるまでのデータ構造と考え、現段階では LinkML を用い、スキーマを https://github.com/integMET/integMET-LinkML/blob/main/integmet.yaml のように暫定設計した(図 3)。現状では RDF の強みである多 DB との「つながり」を重視した設計は保留し、我々が 群間比較に基づく統合グラフ を構築する際に利用するデータ構造(代謝物 ID としての InChIKey、群間比較およ

び代謝物の変動情報）に忠実なスキーマをまず設計した。

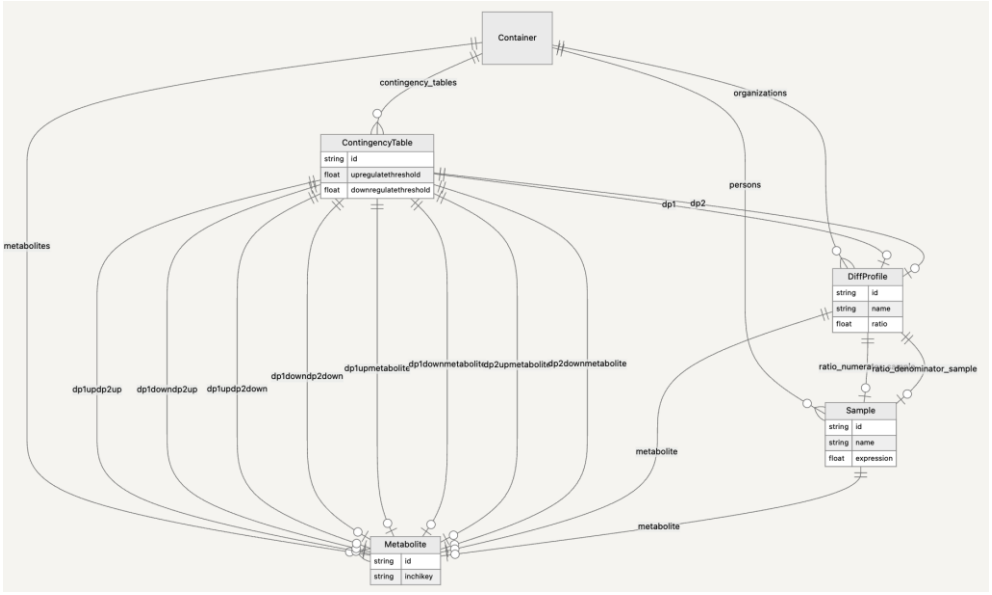


図 3 LinkML スキーマから生成した integMET スキーマの ER 図

### 【項目 2-2-1】「接続先の外部 DB の検討」

本 DB に搭載されるレポジトリの研究情報からより外部の情報を連携させるための接続先 DB の検討の結果、代謝パスウェイ情報源として Wikipathways、化合物情報 DB としては PubChem を連携先として決定した。各データベース運営機関と協議を進め、技術的連携方式の詳細設計を行う予定である。前項目の暫定スキーマ中の Metabolite クラス中の InChiKey を TogoID への入力とし、その InChIKey とのリンクを持つ他 DB (PubChem, LIPID MAPS 等) の ID を Metabolite クラス中のアトリビュートとして追加する予定である。なお、現状 TogoID には InChIKey からパスウェイデータベースのダイレクトリンクは無いため、 WikiPathways、TogoID と協議の上 togoid-config に対してそのデータ追加を行っていく予定である。

## 【項目 3】統合データのエンベディング生成と機械学習・AI との連携

### 【項目 3-1-1】「ノード分類機械学習ワークフローの開発（上流工程)」

本年度は、グラフデータベース上のノードに低次元ベクトル（埋め込み）を割り当てることで、生物学的要素の類似性探索を可能にする基盤を構築した。本 DB のネットワークノードの生物学的要素の類似性の検索をグラフ機械学習によって実現するために、グラフデータベースソフトウェア Neo4j 用 Graph Data Science (GDS) ライブラリ の Fast Random Projection を用いて 32 埋め込み次元のエンベディング生成を試験的に行った。なおグラフの規模が大きいほど大きい埋め込み次元を必要とするため、今年度はまずは試験的に植物に関連する部分ネットワーク（graph_group='Plantae'）に対して 32 次元の埋め込みを生成した。

この過程では以下のステップを実施した。

- グラフ投影：対象ノードと関係性を抽出し、GDS 用の論理グラフとしてプロジェクト化。
- パラメータ設定：埋め込み次元（`embeddingDimension`）や反復重み（`iterationWeights`）を試行し、安定したベクトル表現が得られる構成を探索。

- 埋め込み生成：`gds.fastRP.mutate` コマンドにより、各ノードに埋め込みを付与し、プロパティとして格納。

これにより、代謝物変動の類似性グラフデータ構造上の類似性を、数値ベクトルとして扱うことが可能となった。

**【項目 3-1-2】「ノード分類機械学習ワークフローの開発（下流工程)」**

上流工程で生成された埋め込みを基に、ノード間の類似性情報を活用した分類を行うパイプラインを構築した。現時点では最も基本的な手法である K 近傍法（kNN）を前項目のエンベディングに適用し、類似ノードの識別に関する評価を検討中である。この実装により、ある群間比較ノードが特定の属性(何らかのスタディメタデータ[ある InChIKey や ある MeSH term を持つか等) を予測することが可能になると期待される。今後は、上流工程と下流工程と統合した end-to-end の グラフニューラルネットワークパッケージの活用や、複数の埋め込み手法を比較することで、さらに高精度な表現を目指すことも検討する。

**【項目 4】搭載データ拡張およびそのためのシステム構築**

**【項目 4-1-1】「Metabolomics Workbench のデータ前処理」**

Metabolights データレポジトリ用に開発した情報抽出・処理スクリプト群を、メタボロミクス研究の主要公共レポジトリの一つである Metabolomics Workbench (MW)に対応するように改変を行った。MW には 2025 年時点で約 3,500 件のスタディが登録されており、そのうち代謝物の比較定量情報およびメタデータが適切に記述された約 2,700 件について処理を完了した。具体的には、レポジトリに格納された代謝物比較定量テーブルおよびメタデータから代謝物同定情報、サンプル群間の比較定量情報、実験条件・生物種等のメタデータを構造化抽出し、本データベース用の統一フォーマットに変換した。これらのデータは後続プロセスにおいて群間比較の生成、および Pubtator 等を用いたメタデータの抽出(図 4)および標準化処理へと供された。その結果、すでに処理済みの Metabolights のデータと合わせると、約 4,133 スタディを本データベースに搭載しており、測定された代謝物の種類は約 68,600 種類、メタデータは約 6,000 種類の規模となった。


図 4 MW のスタディから抽出された研究メタデータのワードクラウド

**【項目 4-1-2】「Metabolomics Workbench の代謝変動ネットワーク統合」**

【項目 4-1-1】で抽出・処理された MW の代謝物比較定量テーブルから、試料情報メタデータに基づく群間比較のデータセットを体系的に生成した。これらの群間比較をノードとして、iDMET アルゴリズム (Matsuta et al., BMC Bioinformatics 2022, 23(1):508)を適用することで、MW 由来のノード間、および MW と Metabolights 由来のノード間で代謝物変動の類似性を評価するためのオッズ比および p-value の算出を網羅的に行った。計算された統計値に基づき、有意な類似性を持つノード間をエッジで接続することで、Metabolights と MW 両レポジトリ由来の研究データを包含する大規模な統合代謝変動ネットワークを構築した。このネットワーク構造により、異なるデータソースから得られた代謝研究結果の横断的な比較解析が可能となった。(群間比較の総数:1,138,100、エッジの総数:19,021,508)

**【項目 4-1-3】「Metabolomics Workbench のメタデータネットワーク統合」**

【項目 4-1-1】において抽出した MW のスタディのテキスト情報から Pubtator を用いて構造化されたメタデータセットを基盤とした研究間の類似性評価を実施した。具体的には、MW 内の各研究間および既存の Metabolights 由来のスタディとの間で総当たりで MeSH Term の類似度算出を Jaccard 指数を採用し行った。算出された類似度スコアに統計的閾値を設定し、有意な類似性を示すスタディ間にエッジを形成することで、メタデータに基づく研究間類似性を表現するネットワーク構造を構築した(図 5)。

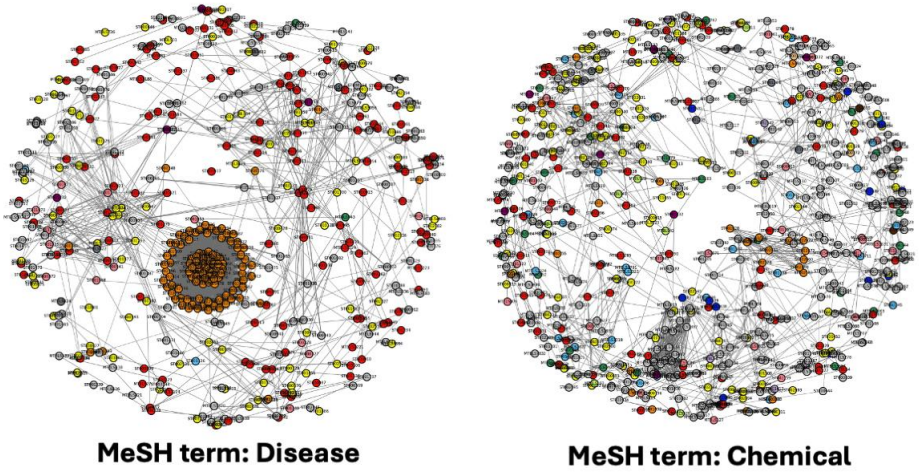


**図 5 スタディメタデータ(MeSH Term Disease および Chemical)に基づいて生成したメタデータネットワーク。各ノードはスタディを示している。**

# §4. 成果発表等

## (1) 原著論文発表

### ① 論文数概要

| 種別 | 国内外 | 件数 |
|---|---|---|
| 発行済論文 | 国内(和文) | 0 件 |
| | 国際(欧文) | 0 件 |
| 未発行論文 (accepted, in press 等) | 国内(和文) | 0 件 |
| | 国際(欧文) | 0 件 |

### ② 論文詳細情報

該当なし

## (2) その他の著作物(総説、書籍など)

該当なし

## (3) 国際学会および国内学会発表

### ① 概要

| 種別 | 国内外 | 件数 |
|---|---|---|
| 招待講演 | 国内 | 3 件 |
| | 国際 | 0 件 |
| 口頭発表 | 国内 | 0 件 |
| | 国際 | 0 件 |
| ポスター発表 | 国内 | 0 件 |
| | 国際 | 1 件 |

### ② 招待講演

〈国内〉

1. 早川英介、創発的再解析のためのメタボローム統合データベース、トーゴーの日シンポジウム 2024、品川ザ・グランドホール、2024 年 10 月 5 日

2. 早川英介・松田りら・高橋みき子 ・山本博之・有田 正規、創発的再解析に向けた研究横断的メタボロミクスデータの統合、JPrOS2024 & JSCP20th (22nd JHUPO & 20th JSCP)、リンクステーションホール、2024 年 6 月 26 日～28 日

3. 早川英介、メタボロミクスデータの研究横断的な統合: 包括的再解析に向けて、第 18 回メタボロームシンポジウム、鶴岡メタボロームキャンパス、2024 年 10 月 23 日～25 日

〈国際〉

該当なし

### ③ 口頭講演

〈国内〉
該当なし
〈国際〉
該当なし

### ④ ポスター発表

〈国内〉
該当なし
〈国際〉
1. E. Hayakawa, R. Matsuta, M. Takahashi, H. Yamamoto, M. Arita、Network-based Integration of Cross-study Metabolomics Data: Towards Comprehensive Reanalysis、Metabolomics 2024、大阪 ATC ホール アジア太平洋トレードセンター、2024 年 6 月 16 日～20 日

## (4) 知的財産権の出願 （国内の出願件数のみ公開）

### ① 出願件数

| 種別 | | 件数 |
|---|---|---|
| 特許出願 | 国内 | 0 件 |

### ② 一覧

#### 1) 国内出願

該当なし

## (5) 受賞・報道等

### ① 受賞

該当なし

### ② メディア報道

該当なし

### ③ その他の成果発表

該当なし

# §5. 主要なデータベースの利活用状況

### 1. アクセス数

① 実績

表 1 研究開発対象の主要なデータベースの利用状況

| 名称 | 種別 | 2024 年度(月間平均値) |
|---|---|---|
| integMET | 訪問者数 | 公開前 |
| | 訪問数 | 公開前 |
| | ページ数 | 公開前 |

② 分析

現在開発対象のデータベース integMET はプロジェクト参加者による試験的な利用と検証のために Amazon Web service 上で試験的に運用されている。ベータ版の公開は 2025 年度後半を予定している。

### 2. データベースの利用状況を示すアクセス数以外の指標

データベースは公開前であり、該当する項目はない。

### 3. データベースの利活用により得られた研究成果(生命科学研究への波及効果)

データベースは公開前であり、該当する項目はない。

### 4. データベースの利活用によりもたらされた産業への波及効果や科学技術のイノベーション(産業や科学技術への波及効果)

データベースは公開前であり、該当する項目はない。

# §6. 研究開発期間中に主催した活動(ワークショップ等)

## (1) 進捗ミーティング

| 年月日 | 名称 | 場所 | 参加人数 | 目的・概要 |
|---|---|---|---|---|
| 2024 年<br>5 月 15 日 | チームミーティング(非公開) | オンライン | 6 人 | 開発の進捗報告・情報共有のためのミーティング |
| 2024 年<br>7 月 12 日 | チームミーティング(非公開) | オンライン | 7 人 | 開発の進捗報告・情報共有のためのミーティング |
| 2024 年<br>8 月 9 日 | チームミーティング(非公開) | オンライン | 6 人 | 開発の進捗報告・情報共有のためのミーティング |
| 2024 年<br>9 月 30 日 | チームミーティング(非公開) | オンライン | 5 人 | 開発の進捗報告・情報共有のためのミーティング |
| 2024 年<br>10 月 18 日 | チームミーティング(非公開) | オンライン | 6 人 | 開発の進捗報告・情報共有のためのミーティング |
| 2024 年<br>11 月 8 日 | チームミーティング(非公開) | オンライン | 7 人 | 開発の進捗報告・情報共有のためのミーティング |
| 2024 年<br>12 月 13 日 | チームミーティング(非公開) | オンライン | 6 人 | 開発の進捗報告・情報共有のためのミーティング |
| 2025 年<br>1 月 24 日 | チームミーティング(非公開) | オンライン | 6 人 | 開発の進捗報告・情報共有のためのミーティング |
| 2025 年<br>2 月 21 日 | チームミーティング(非公開) | オンライン | 5 人 | 開発の進捗報告・情報共有のためのミーティング |
| 2025 年<br>3 月 21 日 | チームミーティング(非公開) | オンライン | 6 人 | 開発の進捗報告・情報共有のためのミーティング |
| 2025 年<br>1 月 21 日 | アドバイザリボードミーティング(非公開) | オンライン | 5 人 | アドバイザリボードと開発に関する意見交換 |

## (2) 主催したワークショップ、シンポジウム、アウトリーチ活動等

該当なし

以上

## 別紙１　既公開のデータベース・ウェブツール等

| No. | 正式名称 | 別称・略称 | 概要 | URL | 公開日 | 状態 | 分類 | 関連論文 |
|---|---|---|---|---|---|---|---|---|
| 1 | 該当なし | | | | | | | |