マテリアル先端リサーチインフラ事業のデータインフラ設計

松波 成行 [1)]

1) 物質・材料研究機構 技術開発・共用部門

MATSUNAMI.Shigyuki@nims.go.jp

Data Infrastructure Design for Advanced Research Infrastructure for Materials and Nanotechnology in Japan (ARIM)

Shigeyuki Matsunami [1)]

1) Research Network and Facility Services Division, National Institute for Materials Science

概要

本発表では、文部科学省が主導するマテリアル先端リサーチインフラ（ARIM）において運用している大規模実験データインフラ基盤の設計・実装について報告する。ARIM は、全国 26 の研究機関に配置された約 1,200 台の共用研究機器から生成される多様な実験データを対象として、インフォマティクスや AI アプリケーションに使いやすいデータ構造化システムを構築・運用している。データ構造化プロセスにおいては、各測定装置固有のフォーマットで出力される実験データを、csv 形式をはじめとする標準化されたデータ交換フォーマットへと変換する自動化パイプラインを実装した。さらに、試料情報、測定パラメータ等の重要なメタデータについては、事前に定義したスキーマ型アプローチに基づき、json 形式のテーブル構造として統一的な管理を実現している。この統合的なデータマネジメント戦略により、FAIR データ原則に準拠した研究データ管理およびデータ共用の運用を進めており、事業を通じて収集された実験データの長期的な利活用促進と持続可能な研究データエコシステムの実現を目指している。

1 ARIM 事業の概要

我が国では、2021 年 4 月に内閣府によって決定された「マテリアル革新力強化戦略」の下、「マテリアル DX プラットフォーム」の構築が進められている [1]。その中で、文部科学省「マテリアル先端リサーチインフラ事業(ARIM ：エイリム)」では、全国的な最先端共用設備体制と高度な技術支援提供体制に加え、リモート化・自動化・ハイスループット化された先端設備を導入し、設備共用支援を継続する。加えて、設備共用に伴って創出されるマテリアルデータは、第三者が利活用しやすく、かつ機械可読性の高い構造化された形で、収集・蓄積を進めている。

2025 年 7 月からは、半導体分野の研究基盤を連携・強化させ、幅広い半導体研究開発のユーザーからのアクセスを可能とするためのネットワーク（ARIM 半導体基盤プラットフォーム）を構築し、我が国の半導体分野の研究開発・人材育成の裾野拡大を目指している。また、2025 年 9 月からは、一定のエンバーゴ期間（デフォルトでは、機器利用年度の翌年度から起算して 2 年間）を経たデータセットについて、ARIM データポータル[2]において、有償ライセンスによるデータ共用を開始した [3]。

2 データインフラ基盤の設計指針

2.1 データ登録からデータ共用まで

ARIM では、プロジェクトに参画する 26 機関が保有する約 1200 台の共用機器が利用可能である。共用機器からの実験データは、物質・材料研究機構がデータ基盤として整備するデータ構造化システム（サービス名：RDE）[4]をホストシステムとして利用し、各機関の保有する機器ごとに固有の AI Ready なデータ構造化様式で蓄積されている。これらのデータは、一定のエンバーゴ期間（デフォルトでは、機器利用

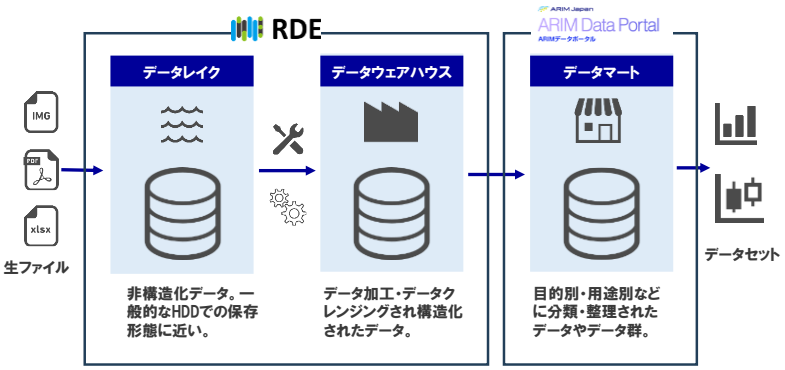

図 1：ARIM 事業におけるデータ登録から共用までのフロー： データ構造化システム（RDE）とデータ共用のための ARIM データポータルとの関係

年度の翌年度から起算して 2 年間）を経た後、ARIM データポータルにおいて、データにかかる利用申し込み（課題申請）を受けたのちに、申請者（申請グループ）に対してライセンスの形で共用される。

クラウドエンジニアリングの観点から見ると、RDE の前受サーバーはデータレイクとして機能し、構造化スクリプトが生成する構造化ファイル群を集積する部分はデータウェアハウス（DWH）に相当する。RDE はデータを管理する研究者間でのデータ共有機能も備えているが、ARIM ではさらに広範な第三者へのデータ共用を促進するために ARIM データポータル[2]の事業専用サイトを構築し、データマート形式でデータ共用を行っている（図 1）。

2.2 データ収集における課題

多くの組織がデジタルトランスフォーメーション（DX）の推進を目標に掲げているが、科学技術分野における DX 化では、一般的な電子商取引とは異なる特殊なデータ形式への対応が求められる。特に、データ駆動型の研究開発を進めるにあたり、データの収集や管理の段階で、しばしば乗り越えるのが困難な現実的な課題に直面する。以下に、代表的な DX 化の課題を示す。

- **組織単位でのデータ統合の困難性**: 各組織や部局で独自のデータ管理手法を採用しているため、組織横断的なデータ統合が困難となるケースがあること。
- **データベース構築の障壁**: データを一元管理するためのデータベースやシステムの構築に関するノウハウやリソース、またそのための専門的人材がいないこと。
- **データ抽出の複雑性**: 必要なデータを多様なシステムやファイルから探索し、抽出する作業が極めて困難であること。
- **ファイルフォーマットの多様性と非互換性**: 装置ごとにファイルフォーマットが異なり、統一された規定が存在しないため、Python 等のプログラムによる自動読み込みが極めて難しいこと。

これらの課題は相互に関連しており、その根本原因はインフォマティクスや生成 AI の利用しやすさとして必要条件となる「データ構造化」の難しさに帰結する。すなわち、データ収集の初期段階において、後続のデータ利活用を見据えた「設計」が不足していることが、課題の根源である。

2.3 データ構造化の考え方

多様なソースから収集されたバラバラの形式の「生データ」,「生ファイル」はその装置固有のドメイン知識を必要とするとともに、メーカ固有の表記や略語が使われていることから難解

図 2：科学技術分野における装置・計測にかかる実験データのデータ構造化スキーム

であり、そのままでは第三者が利用することはできない。共用されたデータを実験者ではない第三者が効率的に利用し、その能力を最大限に発揮するためにも、データを整え、意味を明確にし、統一された形式に揃える作業、すなわちデータ構造化が不可欠である。具体的に、データ構造化は以下の作業を含む [5]。

- **フォーマットの変換**: 前述のような多種多様なファイルフォーマットを、csv や json といった、機械可読性がありプログラムによる処理に適した標準的な形式に変換する。
- **メタデータの付与**: 「この数値は測定温度（℃）である」「この文字列は試料名である」といったように、データ個々に意味を説明する情報（実験メタデータ）を明確に紐付ける。
- **クリーニング**: 欠損値（データ欠如部分）の処理、単位の統一等を実施し、データ品質を向上させる
- **統合**: 複数のデータソースからの情報を統合し、一貫性のあるデータセットを作成する。

この作業によって初めて、データは AI による利用も可能な「構造化データ」となり、その真価を発揮する。

3 データ構造化の概要

3.1 データ構造化の方法と流れ

一般に装置からの出力ファイルは、その装置メーカ固有のフォーマットに従って装置出力パラメータ（計測パラメータ）と数値データから記載されている（図 2）。ARIM では共用装置ごとに、次のようなコンセプトで装置（計測）メタデータと数値部を分けて構造化を行う。

・数値部データ： 汎用性の高い csv フォーマットとし、その後の機械学習などのデータ処理に使いやすい形態にしている。

・装置（計測）メタデータ： 各機器固有のファイルフォーマットに基づいて、必要となる装置・計測メタデータの中から選定し、その選定項目をスキーマとして定義し、パーサースクリプト等から抽出しデータベースへ格納する。

スキーマ方式の考え方は、装置からの出力形式の多様性は受け入れるため、各形式を自動で翻訳・変換するためのスキーマ定義を、装置ごとに用意する、一種の「多言語翻訳」を目指すアプローチである（図 3）。既存の装置にも、新しい装置にも、スキーマを追加するだけで柔軟かつ迅速に対応できるメリットがあるが、代わりに装置ごとにスキーマを作成する手間（初期のエンジニアリングコスト）がかかる。現在、1200 台の装置の内、約 1,000 台についてスキーマの整備を終え構造化対応となっている。

図 3：スキーマ方式によるデータ構造化の概念図

ARIM では、このデータ構造化された一連のファイルを利用課題の機器ごとに「データセット」として管理する。一定のクローズ期間を経たあと、この構造化されたデータセットをフォルダーとしてデータ共用する仕組みをとる。

3.2 構造化のフォルダー構成

構造化されたデータセットは、図 4 に示されるようなフォルダー構成となっている。ここには、登録された生データ（生ファイル）のほか、構造化後の実験データ、実験メタデータなどが含まれ、各ファイルには詳細な情報が記載されている。次に各ディレクトリおよびファイルの役割を体系的に示す。

まず、ルートディレクトリである dataset_[データセット ID]は、個々のデータセットを一意に識別する ID が付与されており、その直下にはデータセット全体の概要や定義に関するファイル群が配置されている。catalog.json は「付帯データカタログ情報」を格納しており、データカタログ（登録情報）に関する詳細が記述される。このファイルは、catalog.schema.json で定義された項目に従って情報が記載されており、ARIM システムに登録された機器やデータカタログに関連する情報が含めることができる。

次に、invoice.schema.json は「装置実験情報スキーマ」であり、主としてデータセットで使用された装置のオペレーションに関する実験情報を定義するためのスキーマとして機能する。具体的には、当該情報が ARIM のデータ構造化システム（RDE）へアップロードされる際に表示される手入力の情報を記入するデータ登録テンプレート（RDE では”インボイス”と称する）に記載される情報が格納される。これは一種の簡易的な電子ラボノート（ELN）であり、計測装置やプロセス装置を使った時の詳細な背景情報（例えばサンプルの調整条件、機器のコンディション）も登録することができる。

metadata-def.json は「装置メタデータスキーマ」として機能し、当該装置から出力される測定条件やプロセス設定（レシピ）に関するメタデータの記述方法を定義するスキーマファイルである。ARIM 事業では、同じメーカの機種であれば、26 機関で共通したスキーマ定義とすることで、機関間でのデータ連携を可能とするメタデータの選定基準を設けている [6]。

```
/dataset_{データセットID}/
  ├── catalog.json
  ├── catalog.schema.json
  ├── invoice.schema.json
  ├── metadata-def.json
  ├── /data_{データ番号4桁ゼロ埋め}/
  |   ├── invoice.json
  |   ├── data.json
  |   ├── filemeta.json
  |   ├── /attachment/
  |   |   ├── 添付ファイル1
  |   |   ├── 添付ファイル2
  |   |   ├── ・・・
  |   ├── /main_image/
  |   |   ├── 代表画像ファイル
  |   ├── /meta/
  |   |   ├── メタデータファイル1
  |   |   ├── メタデータファイル2
  |   |   ├── ・・・
  |   ├── /other_image/
  |   |   ├── 画像ファイル1
  |   |   ├── 画像ファイル2
  |   |   ├── ・・・
  |   ├── /nonshared_raw/
  |   |   ├── nonshared_rawデータファイル1
  |   |   ├── nonshared_rawデータファイル2
  |   |   ├── ・・・
  |   ├── /raw/
  |   |   ├── rawデータファイル1
  |   |   ├── rawデータファイル2
  |   |   ├── ・・・
  |   └── /structured/
  |       ├── 構造化ファイル1
  |       ├── 構造化ファイル2
  |       ├── ・・・
```

図 4： 構造化されたデータセットのフォルダー構成

データセットの中核となるデータ本体は、data_[データ番号4桁ゼロ埋め]というディレクトリに集約されている。このディレクトリ名は、個々の測定データを識別するためのユニークな番号が付与され、ゼロ埋めによって統一された形式が採用されている。この data ディレクトリ内には、以下の重要なファイル群が含まれる。

- **invoice.json**: これは、RDE で登録されたデータを、登録単位で紐づけるファイルである。invoice.schema.json で指定された項目に従って、実験に関する情報が詳細に記載される。
- **data.json**: 測定された機器名やメーカー名などの装置登録情報、および測定試料に関する詳細な情報が記述されるファイルである。これにより、測定が行われた具体的な環境や対象物を特定できる。
- **filemeta.json**: データ構造化処理を経てデータベースに登録されているファイル情報を含む。データの由来や変換プロセスを追跡する上で重要である。
- **main_image ディレクトリ**: データセットの代表的な画像ファイルが格納される。データセットの内容を視覚的に理解するための補助情報となる。
- **other_image ディレクトリ**: 代表画像ファイル以外の画像ファイルが格納される。追加の画像データによって、より多角的な視点からデータセットを評価できる。
- **meta ディレクトリ**: データ構造化処理によって生成されたメタデータが格納される。
- **metadata.json:** meta ディレクトリ内に測定時の測定条件や装置設定に関する詳細なメタデータが出力される。これは、metadata-def.json で定義された項目に準拠している。

さらに、データセットには生のデータおよび構造化されたデータ、そしてその他の添付ファイルに関するディレクトリも含まれる。

- **nonshared_raw ディレクトリ**: 機器利用者からアップロードされたデータ構造化前の生のデータであり、第三者への共用が許可されていないものが格納される。
- **raw ディレクトリ**: 機器利用者からアップロードされたファイルであるものの、第三者への共用が許可されているファイルが格納される。主に微細加工機器のプロセスデータについて ARIM が定めるプロセスデータ記載の Excel ファイルを登録した場合に使われる。
- **structured ディレクトリ**: 構造化処理が施された計測やプロセスの数値データが格納される。特に、主として CSV フォーマットとなっており、機械学習などのデータ解析に直接利用しやすい形式で提供される。
- **attachment ディレクトリ**: その他の関連添付ファイルが格納される。

このように、ARIM のデータインフラ基盤の運用では、figshare [7]や ZENODO [8] のデータリポジトリであるような、非構造化の生ファイルや pdf ファイルのファイル共有型システムではなく、構造化された形式、そして関連するすべてのメタデータを含む包括的なパッケージとして登録・蓄積・共用するように設計されており、FAIR 原則（Findable, Accessible, Interoperable, Reusable） [9] に対しても、マテリアル分野の研究者やデータサイエンティストがデータを円滑に利用するため、より精緻な運用を行っていることが見て取れよう。

また、大きな特徴としては、機器利用者もしくはARIMスタッフが日々の実験で得られたデータを RDE にアップロードするだけで、その後の複雑なデータレイクへの格納、データウェアハウスでの構造化という一連の処理が、すべてシステムによって自動的に行われることある。すなわち、ユーザーが装置固有の複雑なデータ構造化のパイプラインを意識することなく、ファイルをアップロードするという単純な操作だけで、自動的に AI Ready なデータセットが生成されることで、2.2 節に示した科学技術分野での課題を克服している。

4 まとめ

本報告では、ARIM 事業におけるデータ構造化の取り組みを通じて、科学分野におけるデータ駆動型研究開発を支援するための包括的なアプローチを提示した。全国 26 機関・約 1200 台の共用機器から生成される多様な実験データを、インフォマティックスや生成 AI 等でのデータの利活用が行えやすい構造化データとして体系的に収集・蓄積するデータインフラ基盤を構築し、また運用している点が、本取り組みの大きな特徴である。特に、従来のマテリアル研究において障壁となっていた「ファイルフォーマットの多様性」「メタデータの欠如」「組織間でのデータ統合」の 3 点に対し、物質・材料研究機構が運用する RDE システムをホストシステムとしてデータ構造化手法を整備した。これにより、装置固有のフォーマットを csv 等の標準形式に変換し、適切なメタデータを付与することで、最終的に機械可読性が高く、直接処理可能なテーブル形式への変換を実現した。

本データインフラ基盤は、科学技術分野の装置データに対して FAIR 原則にも準拠しており、一定のエンバーゴ期間を経たデータセットは ARIM データポータルを通じて一般に共用される。これにより、実験データの持続的な価値創出と利活用の促進が可能となった。これらの取り組みは、科学技術分野における研究データの保全と共用を促進する仕組みの構築に貢献しており、今後のマテリアルイノベーションの創出と加速に大きく寄与することが期待される。

謝　辞

本取り組みは，文部科学省マテリアル先端リサーチインフラ事業（ARIM）の支援のもとで行われた．ARIM の 26 機関のデータ業務に従事するスタッフには貴重な助言や指摘をいただいたことに深い感謝を申し上げたい．

参考文献

[1] 内閣府 “マテリアル革新力強化戦略”, 令和 3 年 4 月 27 日

[2] ARIM データポータル. https://nanonet.go.jp/data_service/

[3] 文部科学省プレスリリース ” 大規模マテリアルデータ基盤を構築・共用開始 ～国内 26 機関連携により、科学と産業を支える知のインフラを整備～”,令和 7 年 8 月 26 日

[4] RDE (Research Data Express) https://dice.nims.go.jp/services/RDE/

[5] 松波 成行, 松田 朝彦, 知京 豊裕, 原田 善之, 吉川 英樹 “IoT データ収集システムのデータアーキテクチャ”, *トランザクションデジタルプラクティス*, **2**, p. 80-89 (2021)

[6] 遠堂 敬史, 松波 成行 “マテリアル先端リサーチインフラ事業(ARIM)における SEM のデータ構造化の取り組みについて” ,日本顕微鏡学会第 67 回シンポジウム, PI-6, (2024)

[7] figshare: https://figshare.com/

[8] ZENODO: https://zenodo.org/

[9] Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* “The FAIR Guiding Principles for scientific data management and stewardship”. *Sci Data* **3**, 160018 (2016). https://doi.org/10.1038/sdata.2016.18