情報処理学会論文誌　デジタルプラクティス　Vol.2　No.2　57–63 (Apr. 2021)

材料データプラットフォームシステム DICE における研究データフローの構築―実践と課題

谷藤 幹子[1,a)]　吉川 英樹[1,b)]

受付日 2020年8月25日, 再受付日 2020年11月9日/2020年12月25日,
採録日 2021年1月26日

概要：材料分野でのデータ駆動型の材料研究の進展を受けて，物質・材料研究機構（National Institute for Materials Science, NIMS）は，材料データプラットフォームシステムの開発に 2017 年に着手し，2020 年からサービス名 DICE として所内試験を開始した．DICE はオープンサイエンスに応えるデータ基盤であるとともに，データ駆動型研究の一つの手法であるマテリアル・インフォマティクスに利用する想定でデータを「つくる」「あつめる」「つかう」の三つを基本コンセプトとしている．そのため FAIR（Findable，Accessible，Interoperable，Reusable）なデータ流通基盤であることが必要であり，材料データベースや材料データリポジトリをオープンデータ基盤として再構築している．特に材料分野ではデータの質・量・安心を担保する必要があるため，FAIR につながる研究ワークフローの設計，構築も並行して実施した．本稿では実現可能な FAIR 原則に沿うデータプラットフォームとはどのようなものか，オープンサイエンスのフレームワークで実践する取り組みと課題として考察する．

キーワード：研究データ，データフロー，相互利用 FAIR，材料データプラットフォーム DICE，データリポジトリ，メタデータスキーマ，マテリアル・インフォマティクス

Research Data Flow in the Materials Data Platform System DICE — Practice and Future visions

Mikiko Tanifuji[1,a)]　Hideki Yoshikawa[1,b)]

Received: August 25, 2020, Revised: November 9, 2020/December 25, 2020,
Accepted: January 26, 2021

Abstract: In response to a progress of data-driven science in the materials science field, NIMS (National Institute for Materials Science) started development of the materials data platform system DICE in 2017, with in-house trials starting in 2020. As a platform system for accelerating data-driven materials research and development, a data management and analysis infrastructure was constructed on the NIMS intranet based on the concept of the three elements of "creating," "collecting," and "using" data. We provide two types of open platform; materials database and materials data repositories. We introduce our FAIR data system in response to an open science practice from the design of a research workflow for using data, while ensuring the quality, quantity and security of circulating data.

Keywords: research data, data flow, FAIR, Materials Data Platform DICE, data repository, metadata schema, materials informatics

1. はじめに

マテリアル・インフォマティクスの進展に伴って，物質探索や材料の性能予測を可能とするデータ駆動型プラットフォームや，予測アプリケーションの需要が高まっている．この背景には，急伸する国家間のデータ覇権の動き，ビックデータと AI 技術の急伸，インターネット通信網の

[1] 国立研究開発法人 物質・材料研究機構 MaDIS 材料データプラットフォームセンター
Materials Data Platform Center, MaDIS, National Institute for Materials Science, Tsukuba, Ibaraki 305–0044, Japan
[a)] tanifuji.mikiko@nims.go.jp
[b)] yoshikawa.hideki@nims.go.jp

大容量・高速化という主に三つの社会現象がある．ここに実験・計測装置の高度化やIoT化に伴うデータの質と転送容量，解析や機械学習のアプリケーションの汎用化に伴うファイル形式など，研究環境の変化というよりもデータ中心の研究スタイル，まさにdata-driven researchに移っている[1]．さらにオープンサイエンスというデータ公開を求める世界的な政策も相まって，「データ」と，それを駆動する「プラットフォーム（場）」そして「データ利活用」への期待が世界に広がった．IT業界やデータ企業が提供するデータプラットフォームは，2020年にはデジタルトランスフォーメーション（Digital Transformation：DX）へ進み，デジタル技術を駆使したデータ駆動の世界へ，研究開発の手法が大きく変わろうとしている[2], [3], [4].

2.　データプラットフォーム機能

(1)　グランドデザイン

「データ」と，それを駆動する研究の場「プラットフォーム」とは何か．プラットフォームの必須要素は「データ」であるが，単にデータファイルがプラットフォームにあるだけでは研究に使えない．データ駆動を可能とするためには，データの質（均一性，共通化，標準形式）が担保され，機械学習等の手法にかけられるデータの量（データセットの数）を両立するバランスが重要になる．一定の条件に合うデータを集め，質を確認し，機械学習にかけ，実験データと公知データをマッピングする，あるいはシミュレーションや評価などでのデータ駆動な手法では，データについての説明情報（「メタデータ」と言う）が不可欠となる．またデータの来歴情報や，機械可読なデータ形式と構造で表現されていることも必要になる．さらに材料分野の特徴として，再現性担保のため，実験や計測の条件や試料の生成レシピがデジタル情報として伴っていることが望ましい．現実には，それらを共通性のある形式で記録する習慣がない場合に，電子ラボノートのようなデジタルに記録し，整形（情報の階層化）ができる環境がまずは必要にある．本データプラットフォームDICEは，データについて，そのデータ創出の段階（データをつくる）と利用（データをつかう）の両端からみて，最低限に必要な情報をデータ構造として設計し，定義した．その定義に基づいて記述される情報を，実際の実験装置や計算過程からデータセットとともにエキスポートする研究データフローとして設計している．**図1**は，プラットフォームをデータと研究に必要な機能として整理したものである．電子ラボノートシステム，IoTデータ転送・収集システム，データ蓄積システム，および蓄積データをその場で視認する可視化システム（ビューアー）で構成している．これをシステム全体図として**図2**に示す．

DICEが対象とするデータは，人間と機械が理解でき，利用に必要な情報（以下，「メタデータ」）として出力する変換システム（後述するM-DaC），データのファイル形式の変換やクレンジングといった前処理を行うメタデータ編集システムを提供することによって，データ情報を補足する機能を持つ．材料データについてのメタデータを機械可読に提供することは，研究データワークフローとしても，NIMSにとっても初めての取り組みであり，研究の現場にビルトインすることが難しいところである．特に，データをインフォマティクスに使うまでの前処理にかかるキュレーターや，データ設計の専門家も，DICE稼働に必要な人材である．

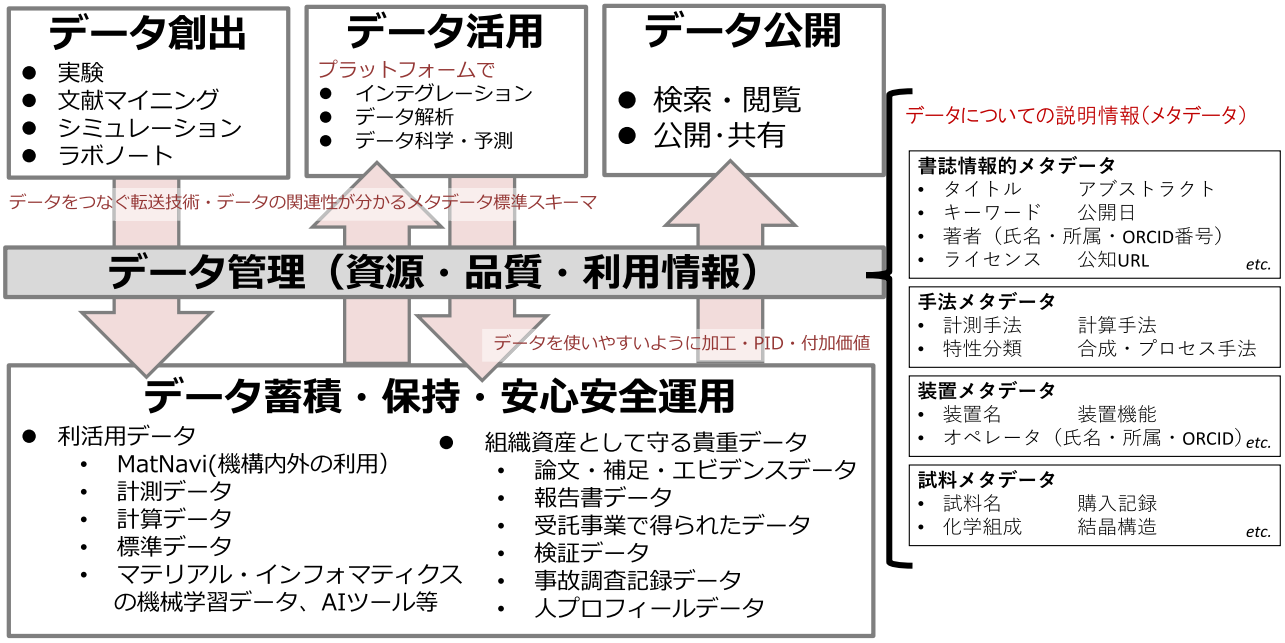

図1　材料データプラットフォームDICEの主要機能

Fig. 1　Main functions of the Materials Data Platform, DICE.

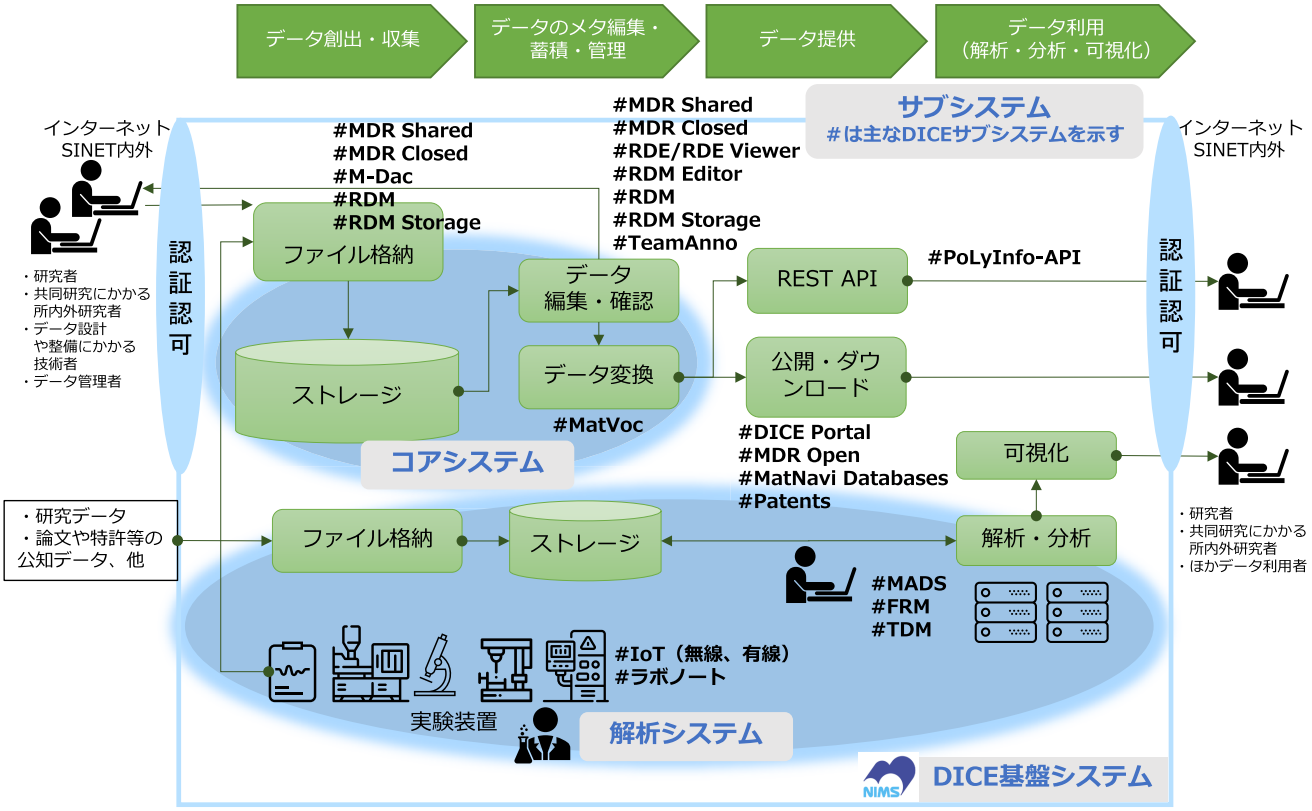

図 2　研究サイクルに伴うデータ創出・蓄積・管理・利用を実践する DICE システム全体図
Fig. 2　DICE system for research data cycle from data creation to data storage and to be used.

(2)　データプラットフォームシステム設計での工夫

DICE はサブシステム，解析システム，基盤システムを合わせて 80 ほどのシステムが連携し，高速解析クラスタと大容量のハードウェア基盤が支えている（図 2）．所内イントラネット，所外インターネットを介した各データ通信，安全安心のための内部監視・外部監視の体制を持つ．

最も重要なデータ資源となるサブシステムは，実験・計測装置からのデータ収集システム，NIMS が保有する既存の材料データベースシステム，公知情報として論文からデータを収集するシステムで構成している．外部機関との共同研究によるデータも，各々の研究契約によって同じデータ流通に乗る場合と，個別のクローズシステムで扱う場合がある．これら異なるデータ資源が，個々にサブシステムとして中央管理（コアシステム）と交信しながらプラットフォーム上で流通する [5]．このとき，サブシステムごとの目的や独自性を許容しつつ，プラットフォーム全体として流通できるよう PID システム（永続的識別子 Persistent Identfier：PID を使ったオブジェクト同定システム [6]）や専門用語や単位を統制する語彙統制システム（材料辞書 MatVoc），人や組織，装置情報などを共通管理するマスタ管理システムなどを，コアシステムを介して連動するサブシステムとし，サブシステム間のデータ流通において可能な限りに統制，変換，情報の場所と関連性を同定することを目指している．特に材料分野での挑戦課題として，材料辞書を作ることにより語彙の統制を目指し，データリポジトリシステムへの実装を通してデータ駆動に導く試みに取り組んでいる [7]．

(3)　サブシステムとコアシステム，API フレームワーク

DICE は，データ源として重要な①実験・計測データとメタデータをセットにして生成し，機械可読化，可視化する機能，②①を公知データとともに集積するストーレッジ機能，③集めるデータを中継して識別子 PID を付与し，来歴情報とともに管理して，解析・予測を実行する解析クラスタ等のサブシステムにデータセットとして渡すコア機能，④データ公開・共用のリポジトリ機能の四つからなる．多種データ源を横断した検索と抽出，解析を可能とするため，プラットフォームで流通するデータのメタデータの設計とその柔軟性が，極めて重要になる [8], [9]．DICE が扱うデータは，米国の標準技術研究所 NIST（National Institute of Standards and Technology）が提案するメタデータスキーマ [10], [11] を参考に三つの階層に仕分けて体系化した．第 1 階層は材料データの基本情報ついて材料分野全体に共通する情報を記述する．第 2 階層は計測装置，試料情報，分析条件，計算ソフトウェア名とバージョンなど，データ生成条件について記述する．これら以外の情報を第 3 階層で扱うこととしている．また FAIR 原則に沿うデータシステムであるために，これらの記述に用いる

図 3　Wikibase を基盤として構築する材料辞書 MatVoc で扱う語彙の例

Fig. 3　Example of terms in MatVoc structured with Wikibase framework.

語彙（用語や物理単位）を統制可能とするため，サブシステムとして材料辞書を開発した．**図 3** は，材料辞書が扱う語彙の階層表現の例を示している．コアシステムは，標準スキーマおよびデータモデル，解釈可能とするため，材料辞書と連携して語彙をセットとして各サブシステムに配信する中継機能を持つ．コアシステムとサブシステムの間をつなぐ API フレームワーク（API-FWK）を使い，新しくサブシステムを開発する場合や運用後に変更を行うときには，API-FWK が指定する共通メッセージ形式に対応することを前提に，コアシステムとサブシステムの接続工数を抑え，プラットフォーム上でのデータ流通の統制を可能にする予定である [12].

(4)　材料辞書 MatVoc

データプラットフォーム上で流通するデータが相互判読・機械学習にかけられるための語彙統制を可能とする材料辞書 MatVoc は，サブシステムの中でもオープンデータの要素を持つコミュニティー指向型システムである．メタデータに使うラベル名や値，データファイル内に出現する装置や計算機固有の表現を共通表現に統制するため，データ収集フローにおいては装置由来の名称や物理単位を変換するためのマスターファイル（対照表の元となる情報）として，データリポジトリではデータ登録や検索語彙として，MatVoc の利用を想定している．MatVoc システムはオープンソースの Wikibase を使い，Wikipedia で知られるインタフェースで入力する．プラットフォームのサブシステムに対して語彙を API でデリバリする設計である．

MatVoc は，材料データベース MatNavi の一つである高分子データベース PoLyInfo で扱う高分子名（図 3）や，理化学事典を出典とする物理単位，ほかオープンデータとして公開されている化合物名 PubChem，化学物質名 IUPAC などからも語彙として収集し，MatVoc の語彙源としている．材料分野は語彙統制という観点では広範囲なため，時間をかけて材料研究・開発に関わる研究者参加型のコミュニティーに育てていくことを目指したい．

3.　材料計測データの自動収集とワークフロー

DICE 上で連携したサブシステムとして，**図 4** に示す材料計測データの自動収集，可読化，メタデータ整理，データ公開のワークフローを開発した．実験や計測の現場で使うサービスとして，計測の専門家自ら設計したサブシステムである．

(1)　計測機器メーカとの協同による計測データを機械学習にみちびくツールの開発と公開

前章で述べた実験・計測データは，装置から出力される段階では，そのままでは FAIR に使えないという問題がある．ましてや機械学習によって統計処理し，新材料の開発を目指す場合，計測条件や試料情報等のメタ情報があることが望ましく，また機械可読性の高い XML 形式に出力ファイルが揃っていると使いやすくなる．しかし計測データの多くは，同一のメーカの装置であっても装置が異なるとデータ形式も異なることがあり，相互比較が難しいという課題がある．また計測条件などのメタ情報が可読性の高い独立したファイルとして記録されないために，対象データの検索も難しい．そのため，計測データを機械学習に使いやすいデータ形式へ変換するツールを開発した．

具体的には，計測機器メーカの協力を得て，生データより計測条件や試料情報等のメタ情報を抽出し，機械可読性の高い XML ファイルへと変換するツールを開発した．最

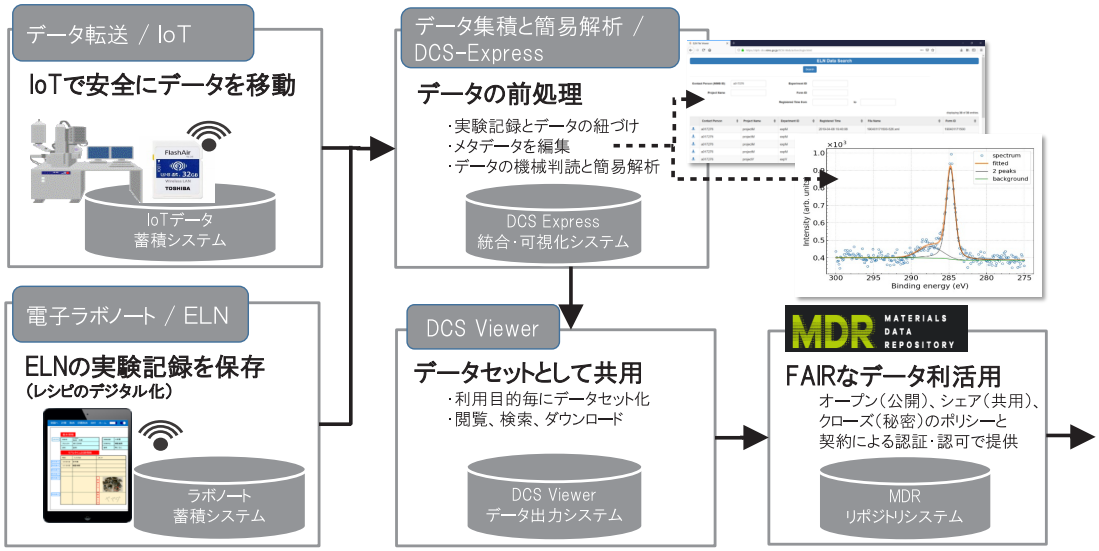

図 4　材料データプラットフォームにおける実験データの移送（IoT），蓄積（Data Curation System: DCS）から機械可読化，メタデータ整理（DCS Viewer），データセットとして公開する（MDR）までのサブシステム連携

Fig. 4　Sub-systems of materials data platform; IoT to deliver experimental data sets, Data Curation System to store and convert to machine-readable formats, DCS viewers for metadata management, MDR as data publishing via data repository.

初の取り組みとして，材料評価で広く用いられる X 線光電子分光法（XPS）と X 線回折法（XRD）の 2 種の計測データについて，メタ情報を付与するための用語変換を定義し，機械学習で主要となるパラメータを抽出するツールを 2 種，開発した．アルバック・ファイ社 Quantera SXM 等の装置で生成された XPS スペクトルと，リガク社 SmartLab の装置で生成された粉末 XRD パターンの計測データに対応する変換ツールである．また，バイナリデータのテキスト変換ツールや数値データ行列の構文解析プログラム（パーサ）を含むスペクトル等の可読化および視覚化の変換ツールも用意し，“M-DaC（Materials Data Conversion Tools）”というパッケージとして公開している [13]．（ソースコードの一部は MIT ライセンスのもと，利用者が改良することも想定し，出力サンプルデータとともにクリエイティブ・コモンズ・ライセンス CC BY-NC 4.0 で利用可能としている．）

(2)　収集・判読・解析を実施する共用基盤とワークフロー

実験データが効率的に集まる仕組みは，データプラットフォームでの重要な基盤の一つである．効率性が高くなるほど実験者のデータ登録の努力のインセンティブは高くなり，解析者の参加もしやすくなる．このため，図 4 に示すデータフローを設計し，単なるデータストーレッジではない機能として，前述の M-DaC も含む機械判読な形式で，自動解析が可能な情報が紐付く「データセット」としてデータが流通する基盤を構築した．このとき，実験装置が組織のネットワークにつながらない（オペレーションするパソコンの世代が古いなどの理由で，ネットワークに接続できない場合も含む）装置からのデータを，安全安心にデータ集積サーバに転送するデータ転送の仕組み（以下，「IoT システム」）を併せて構築した [14]．各システムは，プラットフォームシステムのサブシステムとしてつながり，サブシステムの仕様や対応データ種が増えるたびに，プラットフォームシステムの改修をしなくてよい設計方式をとっている．XPS の例では，解読困難な生データ（一次データ）を，データ判読化し，その際にメタデータの用語辞書を用いた自動翻訳，測定者・装置情報・ソフトウェア・試料情報等を機械出力と人間による補完で整理し（二次データ），データの数理統計処理やスパースモデリングを準リアルタイムで実行するデータ高付加価値化（三次データ）のフローを踏む．（図 4 の Data Curation System: DCS Express システムと連携して動作）

(3)　材料計測データの効率的な解析

前節の準リアルタイムで実行する計測データの高付加価値化として，文献データを含む種々の機関で計測した過去のスペクトルデータと自らのスペクトルデータを統合したデータ駆動型の新しいデータ解析技術の開発を進めている [15]．ここで材料計測のスペクトルデータを対象としているのは，（スカラーである）材料特性値よりも遥かに情報量の多い（ベクトルやマトリックスとしての）スペクトルデータを対象とすることで，新規に合成される多様な物質材料の本質的な物性の違いを説明する特徴量を見つけだすことを目的としているためである．通常，このスペクトルデータからの特徴量の探索は，ベテランのデータ解析者による日単位での人的作業を必要とするものであり，本研究ではその作業の自動化やハイスループット化を目指して

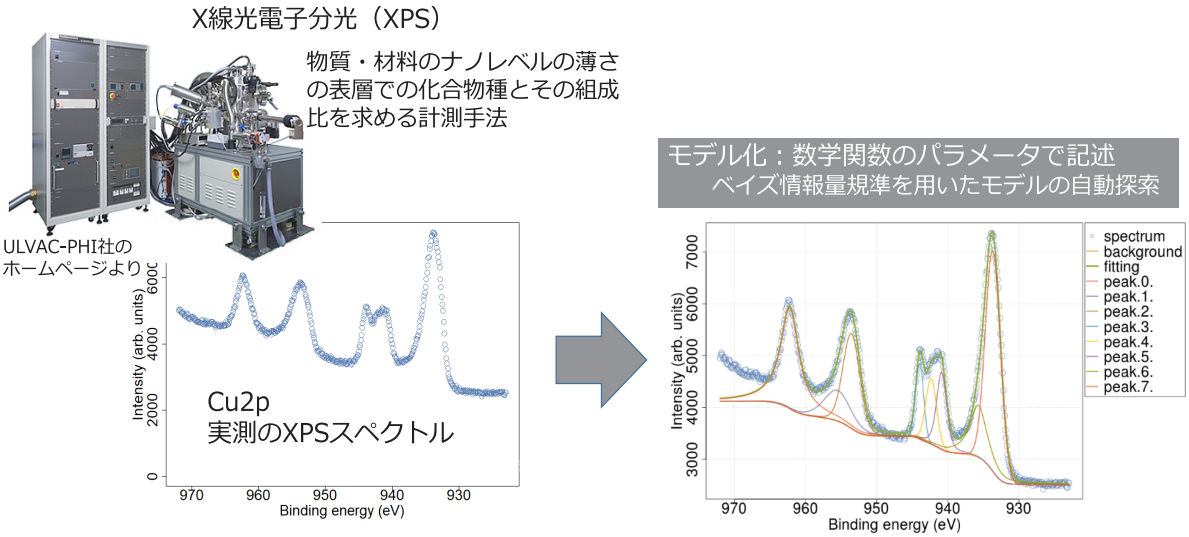

図 5　XPS スペクトルを例としたデータのモデル化
Fig. 5　Example of modeling XPS spectrum data.

いる．

計測装置のエネルギー分解能や透過関数などの装置の違いに由来するスペクトル形状の個体差が，物質材料に由来する特徴量の探索を困難にすることから，装置由来の個体差を推定し補正することで，物質材料由来の情報を得るデータ解析技術が求められる．その技術の基礎となるのが，スペクトル形状の特徴を過不足なく数学関数を使ってモデル化し，計測装置由来と物質材料由来のスペクトルの細部の違いを識別するスパースモデリングである．このスパースモデリングを自動化することで，属人性を排除しハイスループット化した物質材料の特徴量探索を可能にする．**図 5** は，銅酸化物の X 線光電子分光（XPS）のスペクトルを例としてスパースモデリングを行った結果である．モデル化されたバックグラウンド成分と複数個のピーク成分を全自動で推定したもので，多数のスパースモデリングの解の候補の中からベイズ情報量規準を使って最適な解を自動選別したものである [16]．このスパースモデリングのデータ解析のプロセスを前節の計測データの収集と自動判読化のフローに組み込むことで，スペクトル計測の実験中に数分レベルの計算時間内にスパースモデリングの解を見ることができ，過去のスペクトルデータのデータベースと統合したデータ駆動型の自動解析をすることが可能になる [17]．このため，公知データである論文からも該当するデータをマイニングし，実験値とマッシュアップを可能とするサブシステムも併せて開発した [18]．研究ワークフローのバックエンドで接続し，専門家による目利きを経たデータを作るプロセスを組み込むことにより，オープンデータの価値最大化につなげる予定である．

4.　今後の展望

物質・材料研究機構が 2020 年から所内試験を開始した材料データプラットフォームシステムと，そのうえで研究データが流れるデータフローについて概説した．本データプラットフォームシステムは，多数のステークホルダー間で材料研究データを流通させ，データ駆動の材料研究・開発を加速することを目的として開発したもので，「データ管理基盤」「解析基盤」「公知データ源」等の所内基盤と「データベース」「データリポジトリ」の外部公開基盤から構成されている．

「材料研究データはデータを取得した人やグループだけに属する」という意識が研究現場で依然として強いため，論文や報告書の形になったごく一部のデータのみが流通する現状がある．そのため，ステークホルダーの組織や人数が増えるほど，対象とするデータの範囲が（未公開データを含んで）広いほど，データ流通のプラットフォーム開発の難度は著しく増大する．したがって，その開発難度の増大を抑制するには，データプラットフォームの開発におけるデータ登録者を含めた多くのステークホルダーの賛同と協力が不可欠であり，そのためにはデータ流通に対するステークホルダーのインセンティブを常に意識する必要がある．その認識に基づいて，データ流通のワークフローの自動化を指向し，データのキュレーションや解析や管理を分業する仕組みをシステムに取り入れている．今後は，データ駆動型材料研究そのものが DX の時代に入り，FAIR なデータの記述，形式，つなぎ方，流通方法が多様化する．材料分野のデータプラットフォームもスケーラブルに，汎用のクラウドサービスとともに使いこなすことが可能な

DX を指向したい．

謝辞　材料科学の研究に資するプラットフォーム機能の概念設計には，所内多くの研究者とエンジニア専門家からアイデアや具体的なデータ設計まで，幅広く協力をいただいた．特に，出村雅彦氏（物質・材料研究機構），石井真史氏（同機構），門平卓也氏（同機構）からの多くの示唆と協力なくして今日の DICE 実装に至ることはなかった．ここに深い感謝の念を表します．また本システムの開発に携わった諸氏，データキュレーションの協力をいただいた皆様にも，謹んで感謝の意を表します．

参考文献

[1] 谷藤幹子：材料データプラットフォームシステムの設計と構築，機能材料，Vol.40, No.10, pp.4–16 (2020).
[2] 知京豊裕：マテリアル・インフォマティクスの現状と課題～海外の動向と日本の挑戦，情報知識学会誌，Vol.27, p.297, DOI: 10.2964/jsik_2017_032 (2017).
[3] 伊藤　聡：MI2I 拠点における物質科学データの整備と普及，〈https://elements-strategy.jp/pdf/03-OP-9.pdf〉.
[4] 出村雅彦，小関敏彦：SIP-MI プロジェクト，これまでとこれから，まてりあ，Vol.58, p.489, DOI: 10.2320/materia.58.489 (2019).
[5] 菊地伸治，門平卓也，鈴木峰晴，内藤裕幸：高付加価値科学データ創出を指向した研究データ管理プラットフォームのアーキテクチャ，信学技報，Vol.119, No.66, SC2019-2,〈http://pubman.nims.go.jp/pubman/item/escidoc:1890239:6〉.
[6] 門平卓也，菊地伸治，田辺浩介，谷藤幹子：分散システムに対する固有識別子管理―材料研究分野のデータプラットフォームにおける FAIR の実現，Submitted to Journal of Digital Practices (2021).
[7] Tanabe, K., Matsuda, A. and Tanifuji, M.: Designing a vocabulary service for a 'data-driven' materials data repository, 14th Int Conf on Open Repositories, DOI: 10.5281/zenodo.3554054 (2019).
[8] Meguro, S., Lippmaa, M., Ohnishi, T., Chikyowac, T. and Koinuma, H.: XML-based data management system for combinatorial solid-state materials science, Applied Surface Science, Vol.252, DOI: 10.1016/j.apsusc.2005.05.084 (2006).
[9] Tanifuji, M., Matsuda, A. and Yoshikawa, H.: Materials Data Platform - a FAIR System for Data-Driven Materials Science, 2019 8th International Congress on Advanced Applied Informatics, p.1021, DOI: 10.1109/iiai-aai.2019.00206 (2019).
[10] NIST Materials Data Curation System:〈http://mgi.nist.go.gov/materials-data-curation-system〉.
[11] NIMS 材料メタデータスキーマ：〈https://dice.nims.go.jp/services/MDR/〉.
[12] 谷藤幹子，吉川英樹：データ相互利用を可能とするデータプラットフォーム技術，応用物理学会，2019 年 3 月 11 日.
[13] NIMS Materials Data Conversion Tools (M-DaC):〈https://dice.nims.go.jp/services/M-DaC/〉.
[14] 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，Submitted to Journal of Digital Practices (2021).
[15] Suzuki, M., Nagao, H., Harada, Y., Shinotsuka, H., Watanabe, K., Sasaki, A., Matsuda, A., Kimoto, K. and Yoshikawa, H.: Raw-to-repository characterization data conversion for repeatable, replicable, and reproducible measurements, Journal of Vacuum Science & Technology A38, p.023204, DOI: 10.1116/1.5128408 (2020).
[16] Shinotsuka, H., Yoshikawa, H., Murakami, R., Nakamura, K., Tanaka, H. and Yoshihara, K.: Automated information compression of XPS spectrum using information criteria, J. Electron Spectrosc. Relat. Phenom. 239, DOI: 10.1016/j.elspec.2019.146903 (2020).
[17] Matsumura, T., Nagamura, N., Akaho, S., Nagata, K. and Ando, Y.: Spectrum adapted the expectation-maximization algorithm for high-throughput peak shift analysis, Science and Technology of Advanced Materials, Vol.20, DOI: 10.1080/14686996.2019.1620123 (2019).
[18] 天野　晃，高橋政臣，高田安裕，小野寺千栄，門平卓也，谷藤幹子：材料科学のためのテキスト・データマイニングプラットフォームの構築と運用，Submitted to Journal of Digital Practices (2021).

谷藤 幹子（非会員）

日本大学文理学部物理学科卒業，University of Leeds 国際学修了（修士）．2005 年物質・材料研究機構に入所，2017 年より統合型材料開発・情報基盤部門 材料データプラットフォームセンター長，材料データプラットフォーム DICE の構築に携わる．応用物理学会会員．内閣府オープンサイエンスの推進に関する検討会委員．

吉川 英樹（非会員）

1992 年大阪大学大学院工学研究科博士課程修了．博士（工学）．1995 年物質・材料研究機構の前身の無機材質研究所に入所．2017 年より同機構の統合型材料開発・情報基盤部門 材料データプラットフォーム副センター長．