

計測分析プラットフォーム実現のための共通データフォーマットの標準化

A Comprehensive File Format Standardization to Realize the Measurement/Characterization Platform

重藤 知夫^{a*}, 安永 卓生^b, 永富 隆清^c, 藤本 俊幸^a, 一村 信吾^{a, d}

Tomoo Sigehuzi, Takuo Yasunaga, Takaharu Nagatomi, Toshiyuki Fujimoto and Shingo Ichimura

^a産業技術総合研究所

^b九州工業大学

^c旭化成株式会社

^d早稲田大学

要 旨 測定条件や試料条件などのメタデータを十分に備え、追加情報がなくとも多様な観点から実験を再現可能であるとき、そのデータは「独立可用」であるという。計測分析機器が独立可用データを出力するようになれば、データベースによる収集・整理を待たずに、他者の測定データを統合した解析が可能になる。これは、計測分析ビッグデータの構築を通じた AI 応用に大きく寄与する。直接の測定データ以外の情報が多いことから、独立可用データは、計測分析機器によらない共通データフォーマットで実現するのが望ましい。我々がその目的で開発したフォーマット MaiML について、成立の経緯を紹介するとともに、計測分析のモデル化と汎用データコンテナの活用を中心にフォーマットの概要を示す。さらに、MaiML データのデータレイクへの蓄積から利活用に至るプロセスの具体像と、その実現のための課題を検討する。

キーワード：データフォーマット，ビッグデータ，独立可用性，標準化，マテリアルインフォマティクス

1. 共通データフォーマット構想

各種計測分析装置のデータフォーマットが共通なら一定の便利さがあるのは確かで、共通データフォーマットは以前から提案されてきた。一方で、計測分析 DX 実現のための条件を検討すると、ある特長を備えた共通データフォーマットの必要性が明らかになる。ここでは後者を論じる。

計測分析 DX と聞いて何が思い浮かぶだろうか。顕微鏡像を AI に見せると、注目箇所とマクロな物性の予測を指摘のうえ追加実験を提案してくる、というあたりではないか。この基礎となるのが、各種条件を与えると、確率的ではあっても実験結果を予測できるようなモデルである。

その構築のためには、測定条件・試料条件・環境条件・機器条件といった多数の条件を少しずつ変えながら計測した網羅的なデータ集合が教師データとして必要となる。これを計測分析ビッグデータと呼ぶ。条件種数次元の多次元空間（条件空間）の膨大な格子点での計測がその構築には必須で、測定点密度が十分でないと実用的な予測精度が得られない。サ

ンプルング密度が小さいフィッティングでは狭い線幅のピークを検出できないようなものだ。一方で、予測が難しい「線幅が狭い」現象こそ、AI を駆使して発見したいのである。

世間では生成 AI が話題だが、その基礎となるモデル (Large Language Models: LLM) が構築可能だったのは、人間の創造の幅が意外に狭く、しかもインターネットにテキストが満ち溢れていたからだ。計測という高コストな作業を、格子点を埋めるという、意味を見出しづらい目的のために行ってもらいにはどうするか。それには2つの方法がある。

正攻法が「データ工場方式」である。ロボット等を用いたハイスループット計測により網羅的な計測を行う。この方法の発展には目覚ましいものがあるが¹⁾、計測種ごとにデータ工場を準備するのは当面は現実的でない。

別解は、出所が異なるデータを集めて格子点を埋めていく「データ共有方式」だ。本稿で検討するのはこちらであり、「他人の測定データ」でのビッグデータ構成を目指す。

別解で問題になるのが、各種条件（メタデータ）へのアクセスだ。条件空間での座標が確定できないと、データはビッグデータ構築には使えない。（測定データ「横軸」以外の）メタデータがどこに記録されているかといえば、実験ノート、計測器制御計算機のハードディスク、論文、ときには計測者の脳内であったりする。メタデータが足りないデータは、メタデータを採る実験者本人以外にはゴミである。必要なメ

〒305-8560 茨城県つくば市梅園 1-1-1 中央第 1

産業技術総合研究所

TEL: 050-3521-3395

* E-mail: sigehuzi.tomoo@aist.go.jp

2023 年 12 月 13 日受付, 2023 年 12 月 21 日受理

doi: 10.11410/kenbikyo.59.1_20

タデータを伴って初めて、データは再利用可能となる。

「他人のデータを有効利用する」困難を克服する努力は、これまで主としてデータベースプロジェクト（以下、データベースとよぶ）が担ってきた。データベースでは、設置目的にかなうデータを収集する。その質をチェックし、目的とする解析に必要なメタデータをも提出させる。データはメタデータとソフトウェア的に関連付けられる。こうして、データベースの設置目的の範囲内であれば、他人のデータが利用可能になる。解析のためのソフトウェアも提供される。データベースは、「野蛮の中にある文明の要塞」であり、城壁の内部に理想環境を構築しようとするものだ。

計測分析ビッグデータ構築のためには、データベース内のデータのようにメタデータと結びついた再利用可能データが、膨大な数、必要となる。「データ共有方式」でこれを達成する正攻法は、データベースを拡大することだ。先端機器の共同利用とともに、「材料開発」という広範な目的を設定することで多種の計測機のデータを収集している物質・材料研究機構（NIMS）を中心とするマテリアル先端リサーチインフラ事業（ARIM）などが、その先端的な試みであろう（筆者らは、2016年3月にNIMSを訪問し、後にARIMに繋がるマテリアルインフォマティクス関連諸プロジェクトの関係者の方々と情報交換を行い、協力していくことで合意している）。

「データベースの拡大」はデータ共用によるビッグデータ構築のメインストリームではあるが、そこから漏れる計測が圧倒的多数である。そんな「はぐれデータ」をも共用できるようにする試みが、データベースの拡大と並行して必要なのではないか。そんな疑問が我々の出発点だった。データベースの「城壁」の外でも、データベース内同様に他者データを利用できる未来を目指す。データベースがトップダウンであるのに対して、ボトムアップの解決策を提示する。そのための手段がデータフォーマットである。

データを入手しただけでは必要なメタデータが手に入らないことが、他人のデータが利用できない大きな理由だった。データベースでは、解析に必要なメタデータの収集とソフトウェア的関連付けによってこの問題を解決している。とすれば、データファイルにメタデータを含めてしまえば同様の効果が得られるのではないか。どんな目的でデータが利用されるか事前には分からないので、メタデータは幅広く集める。ファイルサイズの巨大化を防止するため、含める情報をリンク（ファイルごとの一意なIDの参照）で替えることを許す。

十分なメタデータ（条件）をファイルに内包することにより、多様な観点から実験を再現できるだけの情報が確保されていることを、独立可用性（independent availability）と呼ぶことにする。独立可用なデータ（データファイルのことだが、以後は単にデータとよぶ）とは、追加情報がなくても実験の詳細が分かる「自立したデータ」である。

計測分析装置が独立可用なデータを出力するなら、単にデータをクラウドに溜めるだけで、再利用可能なデータの集

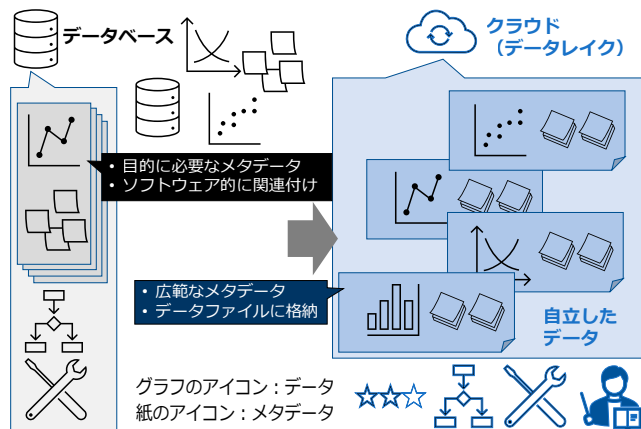


図1 「データの自立」による「データベース内環境」の実現

合「データレイク」となる。データベースやデータ工場が提供できる再利用可能データとデータレイク内のデータを合わせれば、網羅的計測の達成可能性が格段に大きくなる。

加えて、データ工場やデータベースの間でのデータ交換形式としても、データファイルだけの受け渡しで足る独立可用データは有用だ。データ工場・データベース・独立可用データフォーマットは、互いに補い合って、ともに計測分析ビッグデータの実現を目指す。

ここで考えてきたのは、データベースでのデータ相互運用性を、独立可用データフォーマットを用いてデータベース外で達成することであった。このとき、ストレージの他、データの質の評価・ソフトウェア開発・フォーマットの維持管理・普及教育など、データベースが内部向けに整備していた機能を、別途準備する必要があることに注意する。

また、データフォーマットが独立可用データを表現可能でも、機器やソフトウェアが十分なメタデータを提供できないならば、独立可用性は達成されない。機器が現時点で提供可能なメタデータに、解析上必須なものを加えた情報を格納するのが当初はせいぜいだろう。そんな「限定的独立可用性」を利用してAI応用等で成果を示し、計測分析機器メーカーが機器提供メタデータを増やす動機とすることが求められる。

「実験を再現できる情報がひとつのファイルにまとまっている」ことは、ビッグデータ構築以外にも非常に有用になり得る。複数種の計測機器からのデータを統合解析する場合などだ。この辺りも動機に加えたい。

独立可用データフォーマットで計測分析DXを目指すとき、具体的なフォーマットはどんなものであるべきか。

独立可用データは狭義のデータ以外に多くの情報を含み、項目の多くは計測法を超えて共通である。例えば同一試料を異なる装置で計測する場合、試料情報がごっそり共通化できるわけだ。とすれば、計測種ごとに別フォーマットを採用する意味は小さい。独立可用性以外の先進的な機能を採用するためにも、計測分析機器の種類によらない共通フォーマットとするのが望ましい。このような独立可用な共通データ

フォーマットが計測分析 DX には必要だったのであり、以下では単に共通データフォーマットとよぶ。

ボトムアップでのビッグデータ構築を目指すことから、共通データフォーマットの細部の構築は、トップダウン的に準備するより機器のメーカーやユーザーの調整に任せるべきだ。そのためにも、骨格はむしろ確固にする必要がある。現場に構造を知ってもらう必要上、人間が読めるテキストベースのフォーマットが望ましい。ソフトウェアの充実等を考えると XML ベースがよいだろう。

「先進的な機能」はどうか。クラウドに溜めるだけの管理を想定し、さらには再利用を図るというのだから、改竄防止機能は必須だ。メタデータやデータの公開相手を選びたいこともあるだろうから、何が入っているかは公開した上で、内容を暗号化できるとよい。実験の計画と実際の経緯を記録しよう。「実験の計画」は自動実験にも使えそうだ。

このような共通データフォーマットとして筆者らが提案しているのが MaiML (Measurement, Analysis, Instrument Markup Language) である。以下ではその成立の経緯をたどり、その後に MaiML における独立可用性の実現と、それだけに留まらない特長を、フォーマットの概要とともに紹介する。

2. これまでの経緯 (事業紹介)

デジタルトランスフォーメーション (DX) ということばが提唱されてはいたものの拡がっていたとはいえない 2014 年、日本学術振興会に「イノベーション創出に向けた計測分析プラットフォーム戦略の構築」に関する研究開発専門委員会 (以下、研究開発専門委員会とよぶ) が設立された。設置期間の 3 年間、計測分析機器産業の共創領域に構築すべきプラットフォームについて、ソフトウェア・ハードウェア・ソリューション^{注1)}・標準化の 4 つの WG を立ち上げ、各観点から産学出身の委員が検討を行った²⁾。そのソフトウェア WG が提案したのが共通データフォーマット MaiML (当時は XMAIL とよばれていた) である。

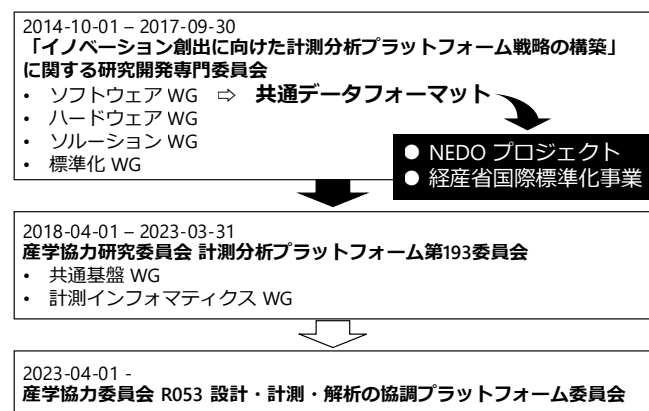


図2 日本学術振興会委員会における共通データフォーマットの検討の経緯

ソフトウェア WG 以外の提案の推進にも共通データフォーマットが役立つと研究開発専門委員会では考えた。デジタル技術による計測分析プラットフォーム創出の加速をイノベーションにつなげる試みであり、まさに DX である。

高性能の単機能計測分析装置を仮想的に統合して同一試料を計測する CPS (cyber-physical system) 型複合計測分析 (SEM で水平形状を、AFM で高さを計測する例を考えてみるとよい) の開発も筆者らが別途研究していたところであり、そのための技術要素としても、試料位置合わせ技術とともに共通データフォーマットは重要だった。

MaiML については、新エネルギー・産業技術総合開発機構 (NEDO) の技術先導プログラム³⁾ や研究開発事業⁴⁾ で実際の計測分析機器での計測を伴った実証研究が行われた。現在も例えば東京大学・東京工業大学で「データ・ロボット駆動科学を推進するデジタルラボラトリーの開発」⁵⁾ などへの応用の試みが続いている。実証を通して、データフォーマットの改善も進んだ。

共通データフォーマットの標準化についても、経済産業省の戦略的国際標準化加速事業⁶⁾ (以下、JIS 化委員会とよぶ) での検討を経て、JIS 規格化の最終段階に入っている (現在審議中の JIS 原案を、以下 JIS K 0200 と記載する)。国際標準の提案に向けての準備も始まっている。

この間、日本学術振興会には 2018 年、産学協力研究委員会として「計測分析プラットフォーム第 193 委員会」 (以下、193 委員会とよぶ) が設置された。同委員会では、研究開発専門委員会の幅広い提言を踏まえつつ、主として共通データフォーマットの普及と活用に関する検討を行った。共通基盤 WG では、共通データフォーマットの必要性と計測分析 DX 内での位置づけ、普及を図る上での障害と対策について、首尾一貫した説明と提案を試みた。

計測インフォマティクス WG では、研究開発専門委員会のハードウェア・ソリューション・標準化の各 WG からの提案を受ける形で議論した。その際、各 WG からの提案、即ち、計測分析 DX 時代を支えるハードウェアのあるべき姿、計測分析プラットフォームが実現したときに迅速かつ高効率でソリューション提供するための仕組み、計測分析プラットフォームを社会実装するにあたって求められる標準化を、ソフトウェア WG が提案した共通データフォーマット MaiML が横串として機能するという観点で進めた。さらに MaiML フォーマットの標準化・一般化が進まなければ各 WG からの提案を実現できないと考え、MaiML フォーマットのデファクト標準化と、マテリアルインフォマティクス (MI) の急速な進展に乗り遅れない共通データフォーマット採用戦略、フォーマットの特長を活かした MI 応用などが議論された。

193 委員会終了後の 2023 年 4 月には、日本学術振興会産学協力委員会として「R053 設計・計測・解析の協調プラットフォーム委員会」 (以下、R053 委員会とよぶ) が設立され、193 委員会までの議論を踏まえた幅広い検討を始めている。同年 9 月には、日本分析工業会主催の JASIS 2023 において

R053 委員会が第1回公開講演会⁷⁾を開催し、193 委員会の議論をどう引継ぎ発展させていくかについて報告を行った。

本稿は、193 委員会での議論を整理し、R053 委員会での検討により明らかになった課題を加えたものである。

3. データプラットフォームとしての MaiML

第1節では、独立可用性を中心に、計測分析共通データフォーマットのあるべき姿を議論した。ここでは他の視点を取り入れよう。

材料の物性・組成・構造の計測分析では、多種多数の機器やソフトウェアが同時に使われる。その出力フォーマットがバラバラなのは、以前から解析の妨げであった。AI などによるビッグデータ解析技術の発展により、MI への計測分析データの活用が重要となる昨今、統合解析のためにもデータフォーマット統一の重要度は増している。その共通認識があるから、共通データフォーマットの提案は絶えることがない。しかし、フォーマット策定のための用語の定義さえ、計測分析機器・ソフトウェアの変化の速さに追いつかないのが現状だ。共通データフォーマットの機器依存部分については、機器のメーカーやユーザーの調整に任せるべきだとの主張は前にも述べたが、その理由はここにもある。共通データフォーマットは、それを可能にするものでなければならない。

計測分析の前には、原料から材料を開発し、前処理して試料を作製する。そして計測が行われ、データ処理等の後処理に移る。この工程の全体の MI 解析が可能ならば、材料開発へのフィードバックさえ夢ではない。そのためには、材料開発に係る計測分析の過程全体を包括的データとして扱えるデータプラットフォームが必要となる。実験の計画と実際の経緯の記録ができるデータフォーマットが求められる所以だ。

以上の要求を満たす共通データフォーマットは、「半構造化データ」とであると考えられる。テキストベースの半構造化データである XML を MaiML が採用したのはこのためだ。

ここで、データ構造の形式としての、構造化データ、非構造化データ及び半構造化データについて整理しておく(図3)。構造化データは、Microsoft 社の Excel シート、CSV 形式などに例示される、行と列とからなるテーブル構造(表構造)を持つ。多くは、一行に「レコード」とよぶ1件のデータを記載し、列毎には、「フィールド」とよぶ1件のデータを表現するための属性(例えば、ID、計測データなど)を規定する。表構造の定義があるため、その後のデータの取扱が容易である。一般にデータベース(データウェアハウス)で使われる RDB (Relational Database) も同様の構造とみなせる。しかし、事前に、フィールドを決定しておく必要があることから、プラットフォームとしては柔軟性にかけることになる。

一方、非構造データとは、文章、画像、動画、音声など、そのままでは構造定義がされていないデータを指す。多くの場合は、メタデータも付随していない。計測分析、前処理、

後処理、材料のデータなどが、非構造化データとして記載されている場合も多く、それが故に多量で、多彩なデータが発生する。個々のデータは「ファイル」というまとまりで管理され、意味付けをされる。付随する「ファイル」としてメタデータが添付されている場合もあるが、その内部の記載方法

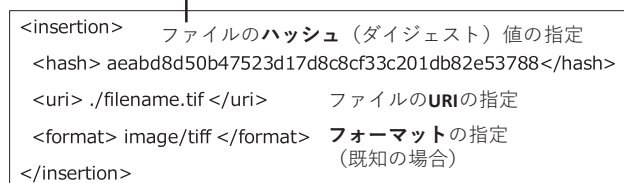
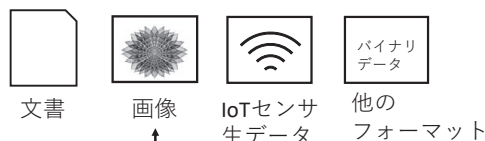
構造化データ

ID	date	Temperature	Humidity	ErrorOfHumidity	...
00012	2023-11-01	10.0	55.0	0.1	...
00015	2023-11-03	12.1	40.0	0.2	...
...	...	...	...	...	...

半構造化データ(MaiMLによる表現)



非構造化データ



MaiMLによる非構造化データなどの外部ファイルの挿入
一意性の保証: ハッシュ値

図3 構造化データ、半構造化データ、及び非構造化データの事例

半構造化データである MaiML での <result> 要素に記載された汎用データコンテナの事例、及び、非構造化データや他のフォーマットを外部ファイルとしてもつ場合に、MaiML において、そのファイルをデータとして挿入する場合の <insertion> 要素の記載事例を示す。

は公開されておらず、多様であるが故に、その後の処理で、データ及びメタデータを活用することが難しい。

今回、MaiML で採用した半構造化データは、XML、json、YAML に代表される、「フィールド」の意味と値を「ファイル」内にもつことで柔軟性を高めているデータ構造である。その中でも、XML は、「フィールド」の意味を「タグ」及びその「属性」に持ちながら、その値を要素の値として、また、付随するメタ情報などを子要素としてもつことができる。要素を階層化することにより「レコード」となるデータの階層をつくり、それを並列記載、または異なる「ファイル」として取り扱うことが可能である。

図 3 には、要素名・属性により、MaiML で記載した事例を示す。図右のブラケットで示す要素（例えば、<result> 要素）は、開始タグ（例えば、<result>）と終了タグ（例えば、</result>）で、要素の範囲を示し、その間に値を記載する。要素名（例えば、result）はデータの塊の名または意味を示す。MaiML の要素は、要素名と属性（例えば、xsi:type, key など）が規定され、階層化した子要素を記載する。

この事例では、構造化データのレコードに対応するものとして、<result> 要素が位置づけられている。例示した <result> 要素は、子要素として、<uuid> 要素、及び 4 つの <property> 要素が記載されている。また、図 3 で示すように、<result> 要素を並列して記載することで、構造化データのレコードに対応して、複数記載できる。

図 3 に示すように、MaiML は、タグ名、属性、および子要素として持つ事ができる要素（タグ名、個数）を規定する。したがって、階層化された要素名（タグ名）、属性、及び属性値を使って、必要なキーと値からなる構造化データに変換することができる。例えば、python 言語などでは、構造化データである CSV 及び XML で記載されたデータを取り扱うモジュールが用意されており、これらを用いることで容易に変換することができる。公開提供される MaiML ガイドライン⁶⁾では、簡単な変換プログラムも示している。

4. MaiML の特徴とデータマネジメント

MaiML は、XML の技術を用いて、① 計測分析のモデル化、② 一意性と改ざん検知、③ 計測分析の追跡可能性、④ 汎用データコンテナによる柔軟な記載、⑤ 外部ファイルとの連結による非構造化データなどとの連係を運用方針として、計測分析の包括的な表現が可能なデータフォーマットとして提案した。特に、サイバー空間であっても、そのデータだけで全ての情報を抽出できる「独立可用性」を担保することが目標である。本解説では、特に、①、②及び④の運用方針と実装を中心に記載する。詳細及びそれ以外の規定、運用方法及びそれらの事例紹介は、JIS K0200 の本体、及びガイドライン⁸⁾を参照にされたい。

まず、①に係り、MaiML では、計測分析の工程をモデル化するために、離散事象システムのモデル化のための数学的

表現であるペトリネット⁹⁾を利用した。ここで、ペトリネットでは、システムの状態をプレース（図 4：五角形、丸、四角形に対応、通常は丸印）、事象をトランジション（図 4：黒縦線）と呼び、トランジションへの入出力をアーク（図 4：矢印）を使って表し、離散事象システムのフローを表現する。MaiML ファイル内では、ペトリネットの XML 表現である PNML で記載している。ここでは、計測分析システムを離散事象システムとして捉え、計測分析の操作を離散事象（transition：図 4 黒縦線、計測分析の操作）とし、状態を示すプレースを、条件（<condition> 要素：五角形、計測分析のパラメータ及び情報）、材料（<material> 要素：丸印、計測分析の試料など物理的実体）及び結果（<result> 要素：四角形、計測分析の結果）の 3 つに分けて用いることとした。最小の計測分析の操作のモデルを、図 4 上段に記載している。これらのプレースは、アーク（図 4 矢印）をつかって、離散事象である計測分析の入力又は出力となるとした。

図 4 下段は、具体的な事例として、SEM の試料の前処理及びその後の SEM による撮影を行った際のフローを示したものである。

次に、②に係り、ビッグデータとして取り扱うことから、計測分析条件、試料及び結果などが、サイバー空間で一意性

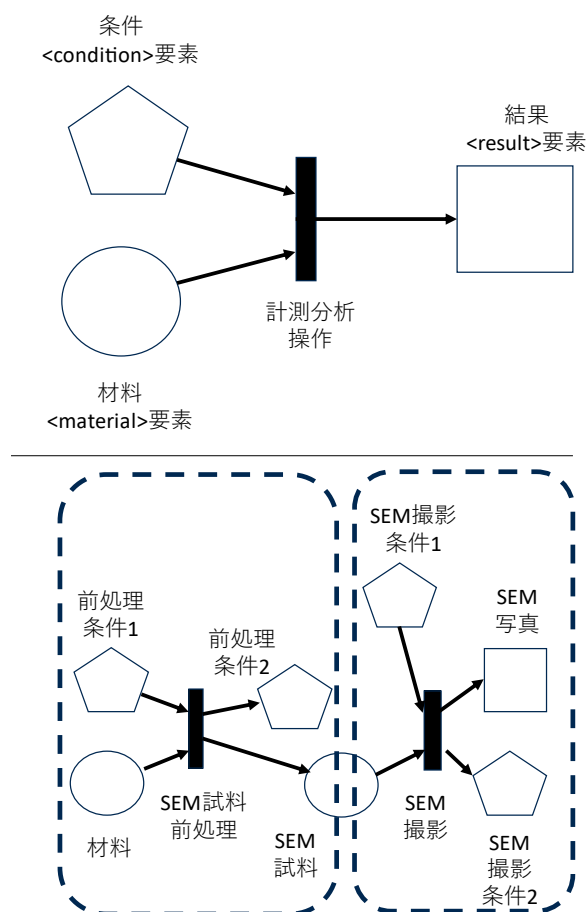


図 4 ペトリネットを使った計測分析の事例
SEM 試料の前処理と SEM 撮影の流れ。

をもつことを保証する必要がある。そこで、<uuid> 要素 (Universally Unique Identifier: UUID) を必須の子要素とした。UUID とは、128 bits からなる数値で、統制なしに重複や偶然の一致がほぼ起こらないことが保証出来ることから、対象を一意に識別するための識別子として用いられる。例えば、全く同一の試料を別の方法を用いて計測した場合には、<material> 要素の子要素の <uuid> 要素の値を同一とする。これによって、別ファイルに格納された <material> 要素でも同一であることが分かることになる。

④に係り、MaiML の規定の中では、データやメタデータの名前を定義しないこととした。その代わりに、<property> 要素 (図 3) 及び <content> 要素という二つの汎用データコンテナを用意した。前者は、単独の値を、後者は軸情報などを伴うベクトルデータを、子孫データの <value> 要素の中に格納する。<value> 要素の値のもつ「名及び意味」は、<property> 要素及び <content> 要素の属性ならびにその他の子孫要素を用いて表現した。xsi:type 属性は値の型を示し、key 属性は値の意味を表す。これ以外に、formatString 属性を用いて値の精度を、units 属性を用いて値の単位を、及び scaleFactor 属性を用いて定数倍であることを示すことが可能である。さらに、<value> 要素で示される値の不確かさを表すためには、図 4 に示す <uncertainty> 要素を並列して記載する。

ここで、汎用性の観点から MaiML における汎用データコンテナの考え方を述べる。AniML など多くのフォーマットでは、要素名をデータの意味 (キー) にしている場合が多い。この場合、フォーマットを規定する段階でデータの意味を表すキーを規定する必要がある。追加するには規定を変更する必要がある。前述のように、今回、柔軟なデータ表現ができること、また計測分析の進化や機器メーカー毎の違いによる方言や新しいパラメータを表現できることが目標であった。このため、key 属性の値として、<value> 要素がもつ値の意味を自由に記載できるとした。一方で、key 属性の値は、XML の名前空間技術を用いて、機器のメーカーやユーザーが設定することで、key 属性値の唯一性を保証できる。

ところで、データやメタデータの名前 (キー) を統一しないということは、多彩な方言を生み出すことになる。この問題点の解消は、前述の名前空間による名前の一意性の保証に加えて、人工知能の分野で用いられるオントロジー (ontology) という技術に期待している。JIS K 0200 では、すでに JIS や ISO で規定された語をもちいる際の名前空間の使い方を例示し、そのことを推奨した (JIS K 0200 附属書 A に“参考”として示しており、規定はしていない)。規定された言葉では表現できない (類似である、二つの語が混合したものであるなど) 場合は、データ作成者が一意性を担保した名前空間を利用し、JIS または ISO の規定語 (用語) との関係を、オントロジー技術を用いて記載する方法を (JIS K 0200 附属書 B に) 例示した。我々が期待しているのは、多彩なキーがサイバー空間上に表れることを許容することによって、そ

れがオントロジー技術により連結することで、継続的な計測分析に関わる用語の定義へと繋がることである。このとき、最低限必要であるとして規定されたキーの値しか格納されないことは、それ以外のデータ及びメタデータが提供されないことになるので、「独立可用性」のあるデータであることが担保できない。MaiML では、機器メーカーやユーザーが自由にキーを設定できることを通して、規定されていない、将来必要になるかもしれないデータやメタデータも記載されていることを期待している。

本解説で説明した以外に、JIS K 0200 では、データマネージメントの観点からデータのオープン・クローズ戦略が可能となるように、XML の暗号化技術を用いたデータの秘匿化のルールを規定している。また、ISO9001 と関わる計測分析のマネージメントの観点からは、③に係る追跡可能性の保証として、XES⁹⁾ と呼ばれるログの記載方法に倣った、計測分析の操作のログを記載することが規定されている。

さらに、⑤に係り、非構造化データ (画像など) を、図 3 に示した <insertion> 要素を用いて、MaiML ファイル内に挿入できる。この時、<hash> 要素を用いて、ファイルの一意性を保証し、<uri> 要素を用いてファイルの位置を示すことができる。外部ファイルは、zip 形式によりアーカイブして、ひとつのファイルにして取り扱う。

5. 共通フォーマットが拓くインフォーマティクスの世界

第 2 節で述べた通り、MaiML フォーマットのデファクト標準化のためにはまずはフォーマットのユーザーを増やすこと、さらにそのためには計測分析装置から出力される計測データを MaiML フォーマットへ変換するためのコンバータの開発・普及が必須である。一方で MaiML フォーマットの開発と、MaiML フォーマットへのコンバータの開発は、193 委員会活動とは別の活動として進められていた NEDO プロジェクトの中で推進された^{3,4)}。そのため MaiML フォーマットと試作コンバータの検証は NEDO プロジェクトへ参画している数機関のみで行われていたことから、より広い事業分野のユーザーからのフィードバックが必要であると計測インフォーマティクス WG では考えた。そこで 193 委員会の参画機関ならびに、著者の一人が参画している公益社団法人新科学技術推進協会の MI 推進 WG の参画機関から希望ユーザーを募り、試作コンバータとデータフォーマットに対する意見集約を行った。得られた意見は、今後のデータフォーマットとコンバータ開発に活用していただくべく、各コンバータの開発を進めた装置メーカーと MaiML フォーマット開発メンバーへフィードバックした。

さらに MaiML フォーマットの普及のためには、独立可用性という MaiML フォーマットの特徴をユーザーに理解していただくことが必須であることはもちろんであるが、その上でユーザーにとって MaiML フォーマットを利用することへのメリットの提示も求められると考えた。各ユーザー機関が機関内に閉じたデータベースを独自に構築している現状で

は、データベース構築に関わる部署や人数が少ないほどデータの類似性が高くなるためメタデータが欠落したデータでもMIが可能となる。しかしながら最近では、機関内で蓄積されたデータの有効活用が求められている。多くの部署の多くの人が関わる場合や、過去に蓄積されたデータを掘り返して活用したい場合には、メタデータを暗黙に前提することができない。必要なメタデータを保持することで信頼性が担保されたデータ、すなわち独立可用なデータがいつ、どこでも利用できる環境が今後のMIには必須となる。さらに将来的には、ある機関内で蓄積されたデータだけではなく、公的なデータプラットフォームに蓄積されたデータの利用や他機関が取得した過去データの共有や売買による研究開発の加速などに向けて、データ蓄積・流通に関する共創領域の創出も必須である。このようなデータの共創に基づくものづくりの時代には、独立可用性を満たすことができるMaiMLフォーマットによるデータの蓄積が最も有効であろうと考えた。この機関間でのデータ共創が193委員会の計測インフォマティクスWGでの議論を経て提案されたMaiMLフォーマット利用のメリットであり、これを社会実装するためのコンセプトがMaiMLフォーマットに基づく「データレイク」である。

図5はMaiMLフォーマットで記述された計測分析データをデータレイクへ蓄積することを想定した計測分析データプラットフォームの構成を示している。データプラットフォームは4つの機能、すなわちデータの収集、データの整形、データの加工およびデータの活用機能を持ち、データレイク、データウェアハウス（データの倉庫）、データマート（データの市場）からなる3層のデータレイヤーで構成される。

データの共創を視野に入れたデータ収集では、種々の材料に対して取得された様々な計測分析装置からMaiMLフォーマットで出力された大量の計測分析データがデータレイクに保管される。ここで注意点は、一般的なデータレイクの定義では蓄積されるデータの形式は問われず、あらゆる形式のデータが蓄積されると想定されるのに対して、計測分析データに関するデータレイクでは独立可用性を担保するためにデータ形式が半構造化データであるMaiMLフォーマットに限定されている点である。画像データやテキストデータ、バイナリデータなどのメタデータは全て、MaiMLフォーマットで出力されたデータに紐づいた附属ファイルとして保存されることになる。またデータレイクのデータ容量は無制限で

長期間保管されることになる。データレイクの段階ではデータの用途は特定されないためデータの使途も汎用的である。一方で、単に集まったデータであり構造化されているわけではないためデータ分析には向かないデータ群である。現状、データの機関内での共有が始まったところで機関を超えるデータ共有はまだ先の話ではあるが、データの共創によるものづくりの時代には機関外の利用も想定したデータレイクが必要となる。

次のステップがデータの整形であり、データ分析に向けた構造化データへと変換するために、データレイクから構造化できるデータを抽出し（E: extract）、分析しやすいデータに変換（T: transform）、システムへ格納（L: load）する。この工程にはETLツールと呼ばれるツールが必要となる。また、このステップでは表記ゆれなどの修正や、データクレンジング、データ結合、データ統合などの作業も必要となる。その結果、データウェアハウスへ蓄積されるデータはデータ分析の基礎となる構造化データとなり、構造化データであるためデータ分析のスピードが高速化できる。ただしこの段階では単にデータを構造化しただけで用途が特定されていないためデータの使途は汎用的である。

最後のデータレイヤーであるデータマートでは、用途や目的に応じて必要なデータのみをデータウェアハウスから抽出・加工して蓄積する。また、用途や目的などに応じて小規模単位でデータを管理する点も特徴である。各データの単位は必要なデータのみで構成されており高速かつ高効率なデータ分析に適した形になっている。ただしこの段階までくると用途が特定されているため使途が限定的となる。また、各単位のデータの中身を利用者が簡単に把握できるよう、含まれるデータの素性を示すデータカタログが必要であり、機関を超えたデータ共有を実現するためにはデータカタログの標準化が求められる。

最後のステップがデータの利活用である。データマートから必要なデータを抽出し、既存の構造化データへ取り込むためのデータクレンジングやデータの結合、統合などが改めて必要となることも想定される。最終的に得られた構造化データに対してBI（business intelligence）ツールによるデータの可視化を行ったり、MIやプロセスインフォマティクス（PI）のためのAIや深層学習によるデータ解析やモデル構築が行われる。

このような計測分析プラットフォームを構築する上でどのような課題があるのか、データの流れの上流から順に考えてみる。まず計測分析を行う段階では装置の自動化やMaiMLフォーマットでのデータ出力・変換ツール、ユーザーによる多数の測定とメタデータの提供などが求められ、装置メーカーと計測分析機器ユーザーの協力が必須である。これらデータを収集するためには自動的に、かつ一定の規則をもって蓄積するためのツールの開発が必要である。その結果、計測分析データのデータレイクに蓄積されるデータはMaiMLフォーマットであるがゆえに材料開発の観点で独立可用性が

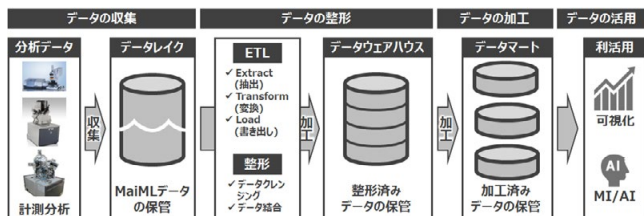


図5 MaiMLフォーマットで記述された計測分析データのデータレイクへの蓄積とその利活用を想定した計測分析データプラットフォームの構成

担保されたデータとなる。

データ整形・加工の段階では、計測分析データに適した ETL ツールが求められる。データウェアハウスからデータを抽出・加工する段階では当初は 1 機関内、将来的には業界内やサプライチェーン内でのデータカタログの整備が必要である。最終的にデータを活用するためには、データマートのデータを各機関が取り込む場合は 1 機関内、業界内あるいはサプライチェーン内でのデータ共有の仕組み、外部にあるデータを秘匿性を保って利用する場合は秘密計算技術が求められる。またデータの共創によるモノづくりのためにはデータ共有、特にデータ提供側に対するインセンティブ設計も大変重要となる。

このように課題を整理すると、データの流れの上流側は主にデータプラットフォームとしての課題であり、後半は主にデータ共有・流通の観点での課題と考えることもできる。現段階ではまだまだ課題は多いものの、当初の 1 機関内での計測分析データプラットフォーム構築から業界内、サプライチェーン内と対象が広がることで、独立可用性を満たす計測分析データの共創が起こり、MI による材料開発の加速が実現されると期待している。そのためには計測分析データの独立可用性を担保できる MaiML フォーマットが不可欠である。

193 委員会での議論を発展させる形で設立された R053 委員会では、計測分析プラットフォームのあるべき姿についても議論・提案しようと考えている。興味のある方には是非参画いただきたい。

謝 辞

「共通データフォーマットの標準化」は、参考文献に記載した日本学術振興会の委員会活動、NEDO プロジェクト活動、及び国際標準化事業活動を推進してきた成果である。それら

の活動にご参加・ご協力頂いた皆様にこの場をお借りして御礼申し上げたい。

注

注 1) ソリューション WG では、どんな機器で何を測ればよいかわからない計測問題を吸い上げ、適切な計測につなげて課題解決を図る社会的な「仕組み」などについて議論した。

文 献

- 1) Shimizu, R., Kobayashi, S., Watanabe, Y., Ando, Y. and Hitosugi, T.: *APL Mater.*, **8**, 111110 (2020)
- 2) 「イノベーション創出に向けた計測分析プラットフォーム戦略の構築」に関する研究開発専門委員会報告書（日本学術振興会、2017 年 9 月）
- 3) NEDO 技術先導プログラム「ビッグデータ適応型の革新的検査評価技術の研究開発」最終報告書（NEDO No. 2019000000536）
- 4) NEDO プロジェクト「省エネ製品開発の加速化に向けた複合計測分析システム研究開発事業」最終報告書（NEDO No. 20200000000670）
- 5) https://solid-state-chemistry.jp/pdf/JASIS_2023_pressrelease.pdf
- 6) 戦略的国際標準化加速事業「計測分析装置の計測分析データ共通フォーマット及び共通位置合わせ技術に関する JIS 開発」（2020～22 年度）。成果物の 1 つとして JIS K 0199「異なる顕微測定装置間における同一箇所分析のための位置合わせ」が 2023 年 1 月に発行されており、JIS K 0200「計測分析装置の分析データ共通フォーマット」も発行見込みである。
- 7) 日本学術振興会 R053 設計・計測・解析の協調プラットフォーム委員会、第 1 回公開講演会「計測分析データ共通フォーマット開発の現状と今後の展開『計測分析プラットフォーム第 193 委員会の活動成果と活用に向けて』、JASIS 2023 トピックセミナー（2023）
- 8) MaiML ガイドライン：<http://maiml.org/>
- 9) Petri, C.A. and Reisig, W.: *Scholarpedia*, **3**(4): 6477, http://www.scholarpedia.org/article/Petri_net