特集　計測標準フォーラム第21回講演会【基調講演】

計測分析を未来につなぐためのデータの在り方と標準化

九州工業大学・大学院情報工学研究院・
物理情報工学研究系
教授　安　永　卓　生

1　はじめに

データ駆動型時代における計測分析を考えるために、はじめに技術と社会変革についておおむね総括しておこう。我々が学生時代であった1980年代のアルビン・トフラーの「第三の波」という本の中では、第1の波としての農業革命、第2の波としての産業革命、そして情報通信革命という第3の波が今押し寄せていることが記載されていた。この段階での情報通信革命とは、情報処理の高速化、大規模化、また、通信による情報の同時性といった、従来、ヒトが行ってきた「情報処理」の一部をコンピュータが担うということであった。私自身、こういったコンピュータをアナブレ（Another Brain）と呼び、ヒトの脳の一部の機能を補完するものとして考えてきた。その上で、明確なアルゴリズムに従って、高速に、大量のデータを生み出すことに成功した。

さらに、1990年代に商用化されたインターネットは、コンピュータ同士通信し合うという技術的革新であったが、ウェブサービスというキーテクノロジーが生まれていく中で、ヒトというデータ発生装置が大量にデータを創出し、公開することとなった。その結果として、人工知能（AI）は、それらのデータを学習し、各種の予測、検出などに使われるようになった。加えて、全ての機器がネットワークに繋がるIoT（Internet of Things）というテクノロジーが実装されるにつれて、大量なデータを生み出しつつある。これらの多量で、多様なデータを一般に「ビッグデータ」と呼ぶ。

科学技術振興機構（JST）における戦略的創造研究推進事業（CREST）における計測分析関連研究・開発でもみられるように、大量のデータを用いたAIによる高速の学習は、上述の情報通信革命と同様に、計測分析の分野では、「高度で精度の高い検出」のキーテクノロジーとなってきた。さらに、生成系AIに至っては、ヒトとの関わりの中で続々と新しい情報（画像、音声、文章など）を生み出すようになっている。日本では、次の時代をSociety 5.0とよび、現在は、まさに新しい社会へ向かう、第4の波とでもいうべき革命の真っ只中にいると言われている。ここでのキーテクノロジーは、人工知能（AI）技術を含む「データ駆動型テクノロジー」である。

いうまでもなく、計測分析というテクノロジーは、材料製造分野を始めとして、多様な分野でデータを生み出すマザーテクノロジーである。一方で、多様な計測分析器から生み出されるデータは、その出力データのフォーマットが、各々に多様であり、かつ、メタデータが不足したデータである場合が多い。例えば、単独の画像であり、その画像データにはその画像の試料情報と接続されていない場合が多い。また、試料の前処理、データの後工程による分析のワークフローの情報などがなく、論文などの手法を読むしか無い場合も多い。そこで、計測分析がこのデータ駆動型社会に向けて貢献していくためには、未来に繋がる「計測分析」が生み出すデータの在り方を考えておく必要がある。

ここに至り、我々は、「MaiML（Measurement Analysis Instrument Markup Languageの略）」と呼ぶ、計測分析に関わるプロセスを包括的に表現を試みた共通データフォーマットを提案している。これは、日本学術振興会における研究開発専門委員会（「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」、2014.10–2017.9）、産学協力研究委員会（計測分析プラットフォーム第193委員会）、経済産業省のNEDO、国際標準化事業などを通して、日本産業規格（JIS）（2014年予定）としてMaiMLの詳細を策定した。

現在、国際標準化機構（ISO）による国際標準化の事業を、ISOとリエゾン関係にあるVAMASを通して行っている。

本稿では、まず、MaiMLのもつ、ビッグデータとしての条件を満たすための概略を述べる。その後、特に、データの汎用的表現及び計測分析のプロセスの包括的モデル化について、その背景となる情報学的背景について記載することとする。

2 ビッグデータとしての計測分析データとMaiML

まず、計測分析とはなにかについて改めて考えておこう。「計測」とは、JIS Z8103：2019のなかで、「特定の目的をもって、測定の方法及び手段を考究し、実施し、その結果を用いて所期の目的を達成させること」と規定されている。この規定の中で現れる「計測」は、同一JISの中で「ある量をそれと同じ種類の量の測定単位と比較して、その量の値を実験的に得るプロセス」とされている。その際、「測定は、測定結果の利用目的にかなう量の記述、測定手順、その手順に従って動作する校正された測定システムの存在が前提」となると注記されている。一方で、分析とは、JIS K 0211：2013のなかで、「物質又は現象を成分、要素に分解し、定量、確認する操作」と規定されている。MaiMLが対象とする「計測分析」は、「計測」、「分析」、または「計測又は分析の一連の過程」を表すものとした。ここで、我々が注目したのは、多種多様といえる計測分析機器及びソフトウェアがもつ、「測定単位と比較した量の値の記載」及び「計測又は分析の一連の過程（プロセス）」という共通性を、計測分析はもつことであった。

つぎに、計測分析データが、将来に渡って利用されるために必要なことを検討した。ここで、「未来に繋がるデータ」とは、たとえ、データが計測分析機器・ソフトウェア及びその計測分析者・所有者を離れても利活用することができる状態にあること、性質をもつこととした。この性質を、我々は「独立可用性」と呼んだ。このために、多種多様な計測分析機器やソフトウェアから生み出される、多様な計測分析結果及びそれに付随するメタデータを、電子データとして包括的に取り扱う方法として、XML（Extensieble Markup Language）を選択した。このフォーマットは、1998年にW3CからXML1.0として勧告されたものである。現在の「ビッグデータ」を支える「拡張性」をもつデータフォーマットである。これは、計測分析データが、サイバー空間上で、「ビッグデータ」となることが「未来に繋がるデータ」といえ、そのために、計測分析の共通性を構造として担保しつつも、今後の計測分析技術の発展を包括できる「拡張性」をもつことが必要であると考えたからである。

このXMLテクノロジーは、1960年代末にIBMの研究者であったゴールドファーブ博士が考案し1979年にIBMが発表したGML（Generalized Markup Language）、1986年にISOにより規定されたSGML（Standard Generalized Markup Language）に端を発する。GMLやSGMLは、文書の電子化を行うことを目的としたものであり、同様の考えを使い、1980年代にクヌース博士が開発したTeXは、現在でも数式を多用する分野の標準として利用されている。これらの言語の基礎にある「マークアップ言語（Markup Language）」という考え方は、「<」と「>」に囲まれたタグとよばれる識別子（**図1**）を使い、文書の構造や体裁を表現するために用いている。

要　素		
<要素型>	マークアップの対象となる文字	</要素型>
開始タグ	要素の内容	終了タグ

図1：マークアップ言語における要素

このマークアップ言語の考え方を利用したのが、HTML（Hyper Text Markup Language）である。1980年代末に欧州原子核研究機構（CERN）のティム博士により開発されたWWW（World Wide Web）を支えるテクノロジーである。WWWは、日本を含めた世界中に分散し異なる大学・研究所で働いている研究者が情報を共有することを目標としたもので、1991年に世界に公開された。HTMLは、ハイパーテキストと呼ばれる仕組みをマークアップにより実装し、URI（Universal Resource Identifier）を使って、ローカル及びインターネット上にある多数の文書を参照し、接続し、紐付けするすることができた（**図2**<a>タグ）。ここで、開始タグ<a>は「アンカー（anchor）」として、文字列のリンク先を示す。さらに、開始タグ<a>は、「属性」が付記され、href属性がリンク先のURIを示す。このことが、インターネットを普及し、インター

ネット上の全てのデータをひとつの大きなデータ、「ビッグデータ」として生み出すことになる。

要　素		
＜a href＝“リンク先のURI”＞	リンクしたい文字列	＜/a＞
＜img src＝“画像のURI”＞	リンクしたい文字列	＜/img＞
＜audio src＝“音声のURI”＞	リンクしたい文字列	＜/audio＞

図2：HTMLによるハイパーテキストの実現例

著者自身、当時、理学部物理学科に所属しており、WWW/HTMLの存在をいち早く知る場にいた。続いて1993年にイリノイ大のNCSA（The National Center for Supercomputing Applications）から公開されたブラウザ（Mosaic）は、文書情報だけでなく画像を含めたマルチメディア情報をブラウザ上で表現できた（図2＜img＞タグ及び＜audio＞タグ）。このことは、現在、電子メールに文字以外の情報をおくる際に利用するMIME型（Multipurpose Internet Mail Extensions）と相俟って、多様な異なるデータ表現（の包括的な表現の可能性を示した。これらは、著者にとって、MaiMLの発想に至る原体験である。ここに至り、多様なデータを含めた「ビッグデータ」の原型ができ上がった。

さて、ビッグデータの特徴として、3V（Volume、Velocity、Variety ）、6V（ 3V＋Veracity、Value、Variability）、あるいは7V（6V＋Visualization）としてあげられている（**図3**）。現在の計測分析データ及びデータフォーマットはこの特徴を持ちうるだろうか。最初の3Vはビッグデータとなり得るデータの特徴を示している。かつてはチャンピョンデータのみが論文に掲載されるなど、計測分析そのものにコストがかかり、ビッグデータの特徴を示してはいなかったが、自動計測・分析が進む中で、データの量は増大し、生成速度が上昇している。また、多様なデータが、次々と開発が進む多種多様な計測分析装置・ソフトウェアから生成されている。また、以前であれば、研究者、材料開発者、計測分析者などのノートにしか無かったメタデータやデータも次第に電子化されつつある。それぞれのデータ（**図4**）は、電子化（デジタル化）はなされたとしても、構造化データもあれば、AniMLなどのXML型データで示される半構造化データ、画像などのような非構造化データとして存在している。もちろん、これらのデータを散在させず、サイバー空間に、あるいはイントラネットであったとしても、データレイク（多様なデータが存在するデータのストック）へと収集されることは、これからのビッグデータ解析を

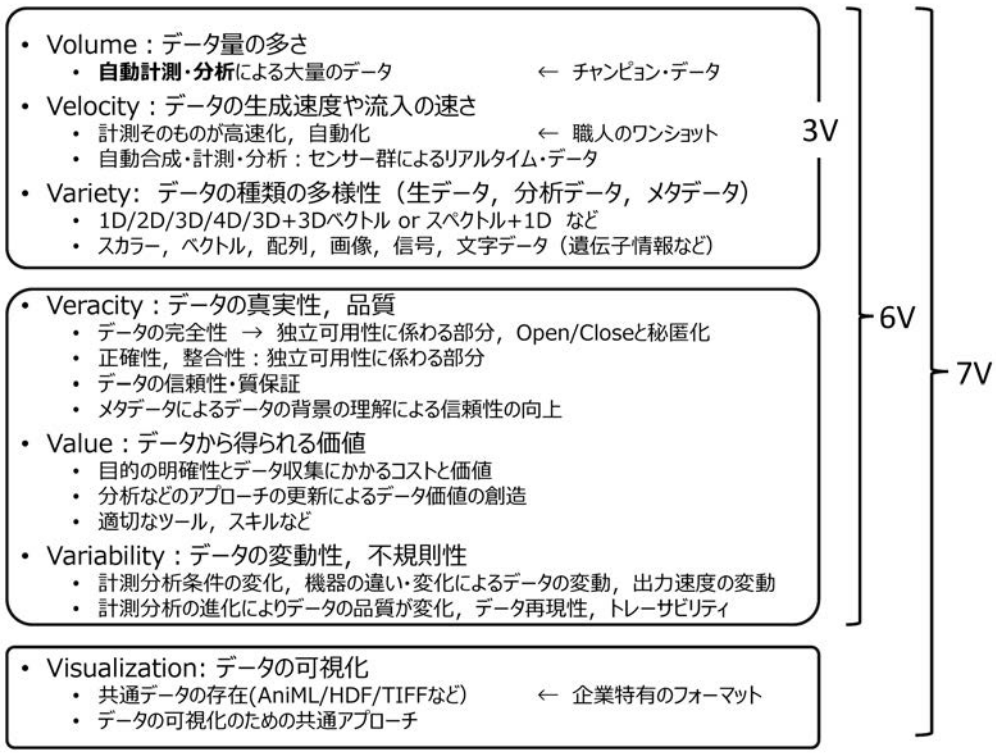

図3：計測分析データのもつビッグデータ性

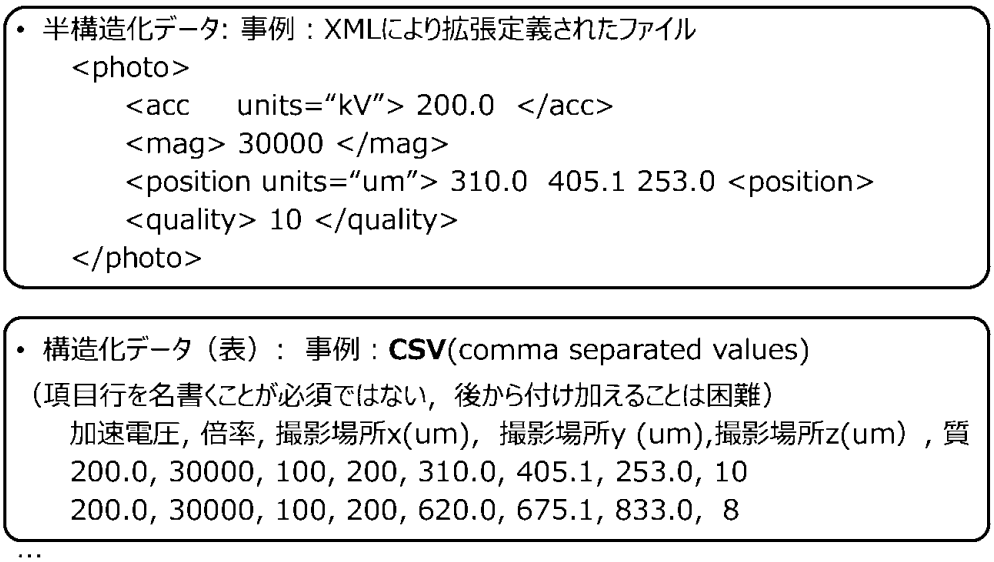

図 4 ：半構造化データと構造化データ

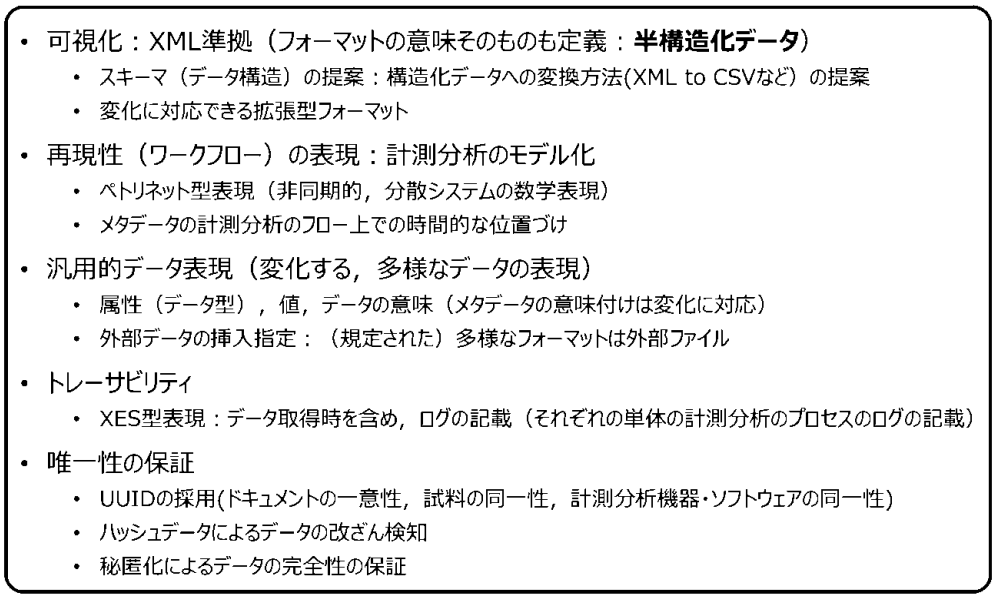

図 5 ：MaiMLの目指す包括的共通データフォーマットの目的

実現するには必須である。しかし、それだけでは、データ間の繋がりが分からないためにデータ解析を行うには不十分である。ここで、6Ｖ又は7Ｖは、ビッグデータへのアプローチ方法の留意点を示している。今回のMaiMLの目的は、上述のデータレイクとして散在しがちな多様なデータを、7Ｖで示されたデータへのアプローチ法を実現するための半構造化データフォーマットの提案である（**図 5**）。

まず、XMLを用いることで、計測分析データの半構造化を通して、データの「可視化（Visualization）」のためのアプローチ法を提案（**図 6**）した。これにより、データへの同一のアプローチが可能となった。データの変動性を加味したデータの汎用的表現（<property>タグ、＜content＞タグ、及びそのネスト構造）により、スカラー量、ベクトル量、表など、構造化された数値データの表現を可能とした。また、画像などの他のデータへの連結（＜insertio＞タグ）によりデータの多様性を担保し、データレイクのデータを接続可能とした。

次に、データの変動性を考慮して、計測分析の再現性（＜protocol＞層）及びトレーサビリティ（<eventLog>層）を設けた。さらに、データ自身の一意性、及び試料・材料、分析に用いるデータ、計測分析機器・ソフトウェアなどの同一性を、UUID（＜uuid＞タグ）によって表現した。また、データの不確かさ（＜uncertainty＞

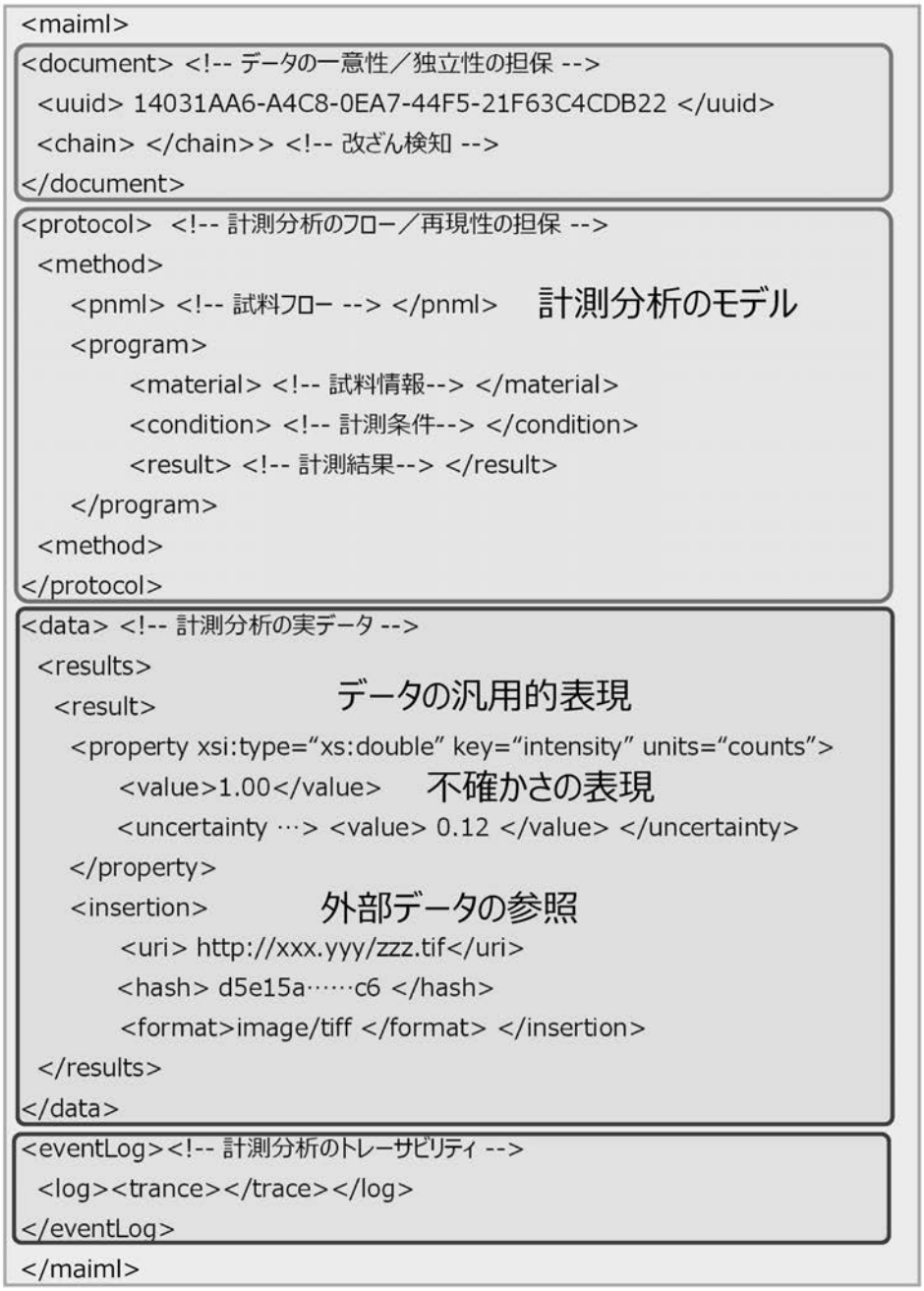

図 6 ：MaiML の全体構造。大きく 4 つの構造からなる

タグ）も記載できることとした。これらを通して、計測分析のデータの品質・信頼性（Veracity）の保証を目指した。本稿で、それらの全て述べることはできないが、これらによって、ビッグデータに向けたデータの「価値（Value）」が生み出せる仕組みづくりの一端を担ったと考えている。

3 データの汎用的表現

計測分瀬において、メタデータのキーを定義して利用することは大変重要なことであり、それぞれの計測分析の手法の中で、専門用語（Technical Term）の標準化が進められたきた。また、現在でもその作業は進められている。一方で、新規の計測分析技術が開発される中で、新しい専門用語が生み出されている事実もある。そこで、MaiML では、これらの専門用語について、すでに JIS や ISO で定義されている用語を利用しつつも、個別に定義可能とし、データ収集を通して、標準化にも寄与できる仕組みを検討した。

図 7 には、データの汎用的表現の事例を示している。＜property＞要素はスカラーデータ、＜content＞要素はベクトルデータを表現でき、子要素の＜value＞要素に量を記載している。ネスト構造（入れ子構造）での

図 7 ：汎用データ表現

表現を使えば、構造体型や表型（2 次元以上）のデータも記載できる。計測分析データは、「測定単位と比較した量の値の記載」であることから、いずれも「単位」をもつタグとして定義され、units 属性をもつ。データ解析を行う際には、データ型が必須であることから、データ型を表現するための xsi:type 属性をもち、精度に当たる記載方法を示した formatString 属性をもつ。また、遺伝子配列のような文字列も記載できる。

また、実際のデータの不確かさ（例えば、標準偏差。他にも適格範囲など）が計測された場合、<uncertainty>タグを用いて、不確かさを記載できる。これは計測分析データには変動性、揺らぎがあることが明らかであることから、データの信頼性を担保することが目的である。

データの意味を表現するには、key 属性、及び子要素として<description>要素を用いる。ここで、key 属性は、名前空間を用いて、値の意味の一意性を保証する。図 7 では、ns1 及び ISO 18115-1 という名前空間を定義し、ns1：Temperature、ISO18115-1：Beam Particle など、key 属性値を設定している。例えば、ISO 18115-1（Surface chemical analysis ― Vocabulary ― Part 1: General terms and terms used in spectroscopy）に関連した URI を用いて名前空間を定義した場合を考えてみよう。ISO18115-1：BeamParticle は、ISO 118115-1 の用語 8 . 9 において、electron、positron、ion、atomic、molecular、or cluster species contained in the incident beam と規定され、入射ビームの種類を示すことが分かる。これにより、専門用語の発散を防ぐことが可能である。

一方で、ns1 は、www.yyy.com が提供する URI と関連付けられている事例である。このように企業やユーザが定義した場合でも、URI と連携しているため一意性を確保できる。この場合も、オントロジーの述語を用いて、いずれかの ISO や JIS の規格との関係を記していくことで、用語の標準化作業の辞書編纂へと繋がることが期待できる。

4 計測分析のプロセスの包括的モデル化

図 6 に示した MaiML の 4 つのカテゴリーのうち、<protocol>層の子孫要素は、計測分析に関わる一連の工程、その前処理・後処理、また、同一試料の複数の異なる計測分析などの操作のワークフローのデザイ

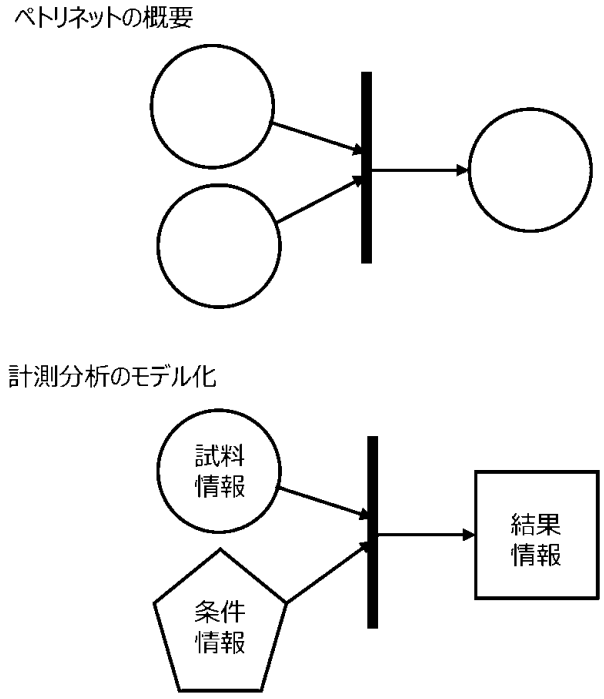

図 8 ：ペトリネットの概要と計測分析のモデル化
○がプレース、｜がトランジッション、及び→をアークと呼ぶ。さらに、計測分析で取り扱う情報を 3 種の情報、計測分析の操作(縦棒)及びその情報の流れ(矢印)として表現した。計測分析における 3 種の情報は、計測分析の対象となる試料情報(丸)、計測分析のパラメータとなる条件情報(五角形)、及び計測分析の結果情報(四角形)としてモデル化した。これにより、計測分析ワークフローをデザインできる。

ンを記載する要素である。計測分析の共通の性質として「計測又は分析の一連の過程（プロセス）」であると捉えることができることから、離散分散システムとして捉えることとした。この離散分散システムの数学的表現であるペトリネット（**図 8**）を用い、計測分析のワークフローを記載するとした。実際のデータファイル内の記載としては、XML の拡張である PNML（Petri Net Markup Language）に準拠したワークフローの記載とした。また、ペトリネットでは、離散分散システムの事象にあたるトランジッション（遷移：transition）を生起すための状態、及び事象の結果として生じる状態をプレース（place）と呼ぶ。プレースとトランジッションとをアーク（arc）と呼ばれる矢印によって結ぶことにより、離散分散システムの動作を表現する。

今回のモデル化では、遷移を計測分析の操作に対応させ、プレースを、計測分析の試料情報、条件（パラメータ）の情報、及び結果の情報の 3 つに分類し、計測分析の一連の過程をモデル化した。それぞれの情報は、図 7 で示した汎用データ表現を用いて、デジタル表現するとしてモデル化した。試料情報とは、計測分析毎に異なる試料やデータとして、入力データなることを想定している。一方で、条件情報とは、計測分析のパラメータ（設定条件）に当たるものである。MaiML によって記載されたデータファイルでは、計測分析でデザインされたワークフロー（＜protocol＞層：デザインされた入力値及び出力予定のデータ種などが記載）、及びインスタンス（＜data＞層：実際の数値などが記載）が記載されている。これらの情報を利用すれば、例えば、計測分析を繰り返し行う場合、または他の計測分析で用いられた計測分析条件と同一の条件で計測する場合にも用いることができる。

図 9 は、分光光度計を用いて、濃度計測の事例を示したものである。まず、標準試料と標準計測条件を用いて濃度計測の校正線（校正計測条件）を生成する。次に、実際の計測条件と校正計測条件を使って、試験試料の濃度を計測し、校正された濃度が結果として出力されることを示している。いずれも吸光度の生データも残すことを推奨し、データの信頼性保証を行うこととしている。標準化された校正方法を用いる場合は、ワークフローと標準計測条件及び試験試

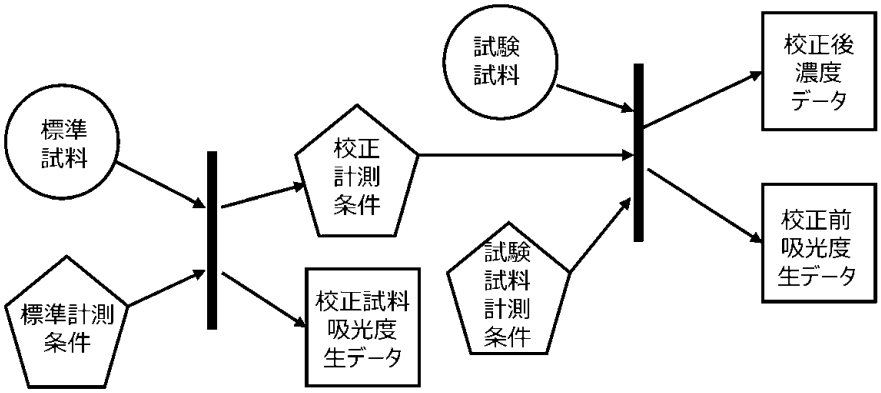

図 9 ：校正線を用いた分光光度計による濃度計測におけるワークフロー

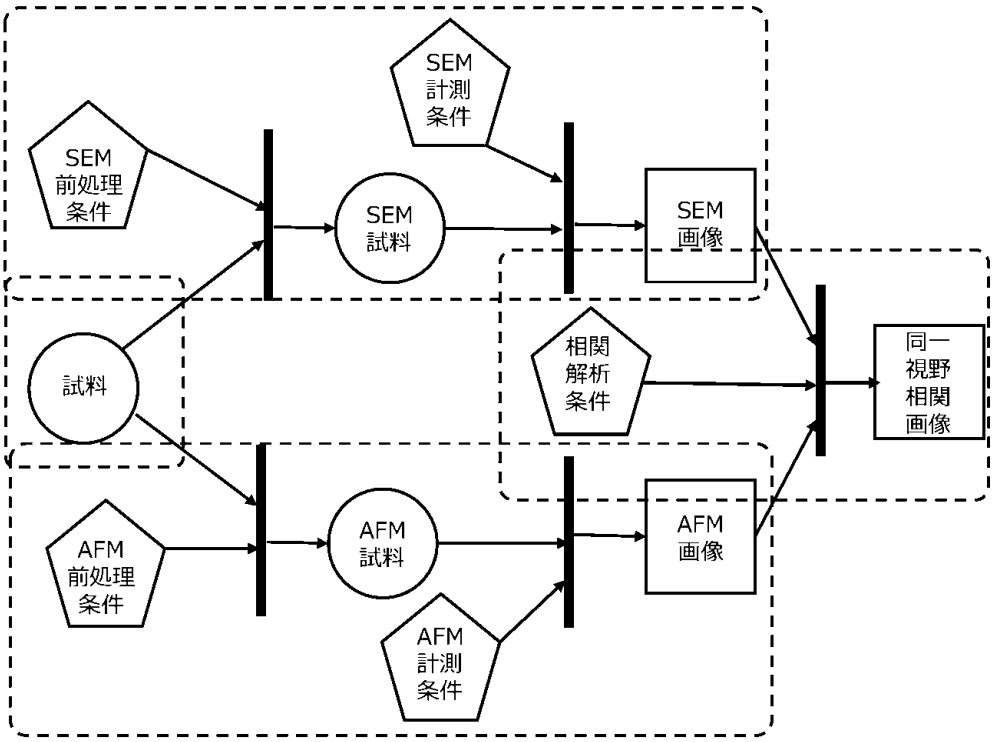

図10：同一試料を、複数のシステムで、計測され、その後、相関分析する場合の事例（SEM/AFM）。複数のMaiMLファイルに記載することも可能である。

料計測条件がデザイン（設計され）されたMaiMLファイル（＜document＞層と＜protocol＞）とだけをもつファイル）を利用して、標準規格に則った計測分析を実施することもできる。さらに、MaiMLファイル（<document>層及び＜protocol＞層に加え、＜data＞層及び＜eventLog＞層をもつファイル）を使うことで、過去の計測分析を参考にした計測分析も可能となる。

ここで示した図9は、直列のワークフローであるが、同一の試験試料を、複数の計測の操作を行って、その後、比較・相関分析などを行う場合にも利用できる。**図10**は、同一試料をSEM及びAFMで観察した後、その相関画像を計算した事例を示している。同一のファイルに記載することも可能であるが、計測分析毎、またその後の相関解析を異なるMaiMLファイルに記載することも許容する。その場合には、同一の試料、SEM画像及びAFM画像は、同一のUUIDをもつとしていることから、データレイクにおいて、接続される。

ここで示したプレース（試料情報、計測条件、結果）は、ワークフローとしては記載することを必須化している。ただし、その内容全体を、XMLのもつ暗号化技術をもちいて、秘匿化することが可能である。それぞれのプレースを秘匿化するプレースと公開するプレースを分けることで、データのオープン／クローズ戦略を採り得る。また、UUIDによる一意性の保証を利用することでの秘匿化も可能である。例えば、図10の場合、試料の詳細情報は、秘匿したい場合には、試料の詳細情報をもつファイルを別に用意しておき、UUIDのみを、計測分析の専門業者におくることにより、情報を秘匿化することもできる。

図10で示した試料の情報と最終の同一視野相関画像について、多数のデータがビッグデータとして集まってくれば、AIなどを用いて学習することで、相関画像と試料情報の相関関係が改めて見出される可能性がある。その際、ワークフローにより、時系列が明確になっているので、単なる相関関係ではなく、因果関係を自動的に見出すことになるであろう。もちろん研究者・開発者からみれば、因果関係と判断することは可能であるが、AIなどのコンピュータは、時系列が分かっていなければ、因果関係なのか、単なる相関関係なのかは、自動的に判断することはできない。

さいごに、**図11**に基づいて、＜eventLog＞層について、述べておく。ここには、実際に行われた計測分析の操作の「終了時刻（complete）」を記載するとした。これはデータ・計測分析の「トレーサビリティ」を保証するためのものである。さらに、図11に示したように、計測分析がスケジュールされた時刻、開始時刻

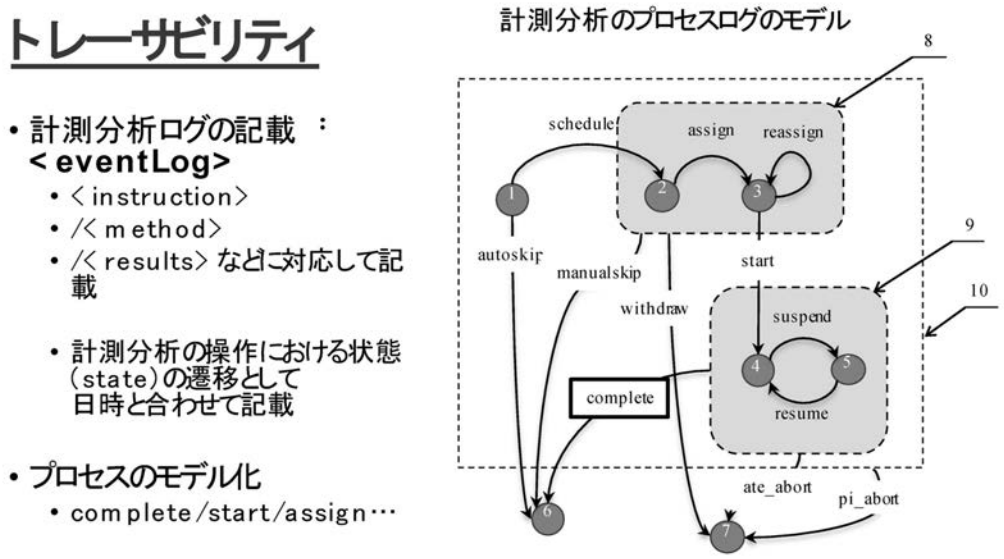

図11：計測分析の操作のログのモデル化。ワークフローのイベント及びイベントフローに関わるログの取扱について、IEEEで標準化が進んでいるXES（Extensible Event Stream）に準拠している。

など、操作の事象の状態遷移を記載することも可能である。この遷移モデルは、XES（Extensible Event Stream）とよばれるデータフォーマットで提案されているものである。

＜eventLog＞層のデータは、プロセス・マイニングと呼ばれる情報工学分野との接続を期待している。データマイニングが、データの情報から相関関係などを検出することが目的であるのに対して、プロセスの情報から有益な情報を抽出しようとする分野である。元々は、ホテルの従業員の働き方など、人間のワークフローのボトルネックの検出、改善などを目的としてスタートした。今回、計測分析そのもの計測分析のプロセスが自動化され、高速に多量のデータが排出されるにつれ、このプロセスの把握、改善、異常検知などが、データが多量であるが故に困難になってくる。例えば、異常値が出た際、または、さらにプロセスの改善を行う際に、どの事象・行為の遷移が問題であるのか、ペトリネットによるワークフローとこのログ情報を組み合わせることで、発見的なアプローチが可能になることを期待している。

5 ま と め

今回、計測標準フォーラム講演会において、我々の計測分析に関わる包括的なデータフォーマットの標準化への考え方を通して、未来に繋がる計測分析データの在り方を、著者自身も改めて考え直す機会を得た。しかし、JISの本体が100頁を超えている規格であるために全体を示すことは困難があり、本稿も読者に採ってみると物足りないと感じる向きがあるだろう。また、包括的であること、及び汎用的であることは、ややもすると、結局、使い物にならないと感じられる可能性がでてくる。また、データの汎用表現においても、key属性のメタデータを縛ることを期待するユーザーが多いことも確かである。これらの課題を少しでも排除するために、名前空間を活用しオントロジーに繋ぐことにより、将来の専門用語の標準化へと寄与できることを目指した。また、計測分析としてモデル化し、相関関係から因果関係を見出すことへの期待などを述べた。ガイドラインとして公開するものには、ユーザー側、計測分析機器メーカ側の具体的な利用法を含めて、事例紹介をしたものも作成し、また、試験利用できる環境も提供する予定である。

皆さんからの多くの意見などを頂きながら、今回の共通データフォーマットをプラットフォームとして、計測分析自身が、ビッグデータとして、Soceity 5.0で利活用される、新たな段階に進むことを期待している。

6 謝　辞

本稿で紹介したMaiMLは、産学協力研究委員会（計測分析プラットフォーム第193委員会）、一般社団法人日本分析機器工業会、国際標準化調査委員会（計測分析データ共通フォーマット）のメンバーの皆様のご協力により、ここまで成熟させることができました。この場を借りて、感謝申し上げます。