機械可読性を高める計測データのメタ情報抽出ツール（M-DaC）の開発と一般提供

～装置やメーカーで異なるデータ形式を統一　データ科学による新材料開発の促進に期待～

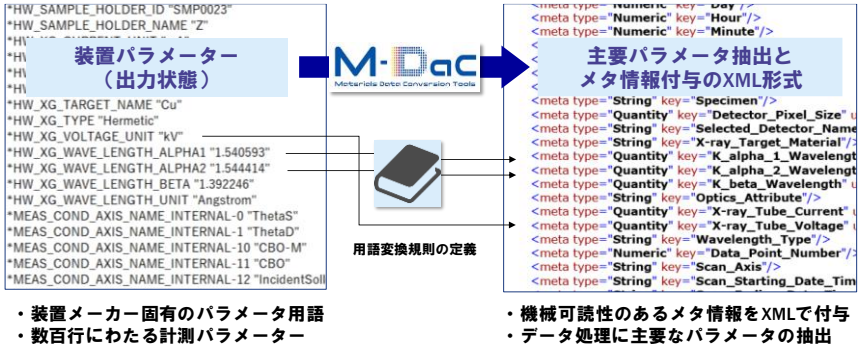
配布日時：2019 年 1 月 29 日 14 時

解禁日時：2019 年 1 月 30 日 14 時

国立研究開発法人　物質・材料研究機構

概要

1．物質・材料研究機構　（NIMS）　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（DPFC）は、計測装置より出力されるデータから、計測条件や試料情報等のメタ情報※1 を抽出し、機械可読性※2 の高い XML ファイル※3 へと変換するツール（以下、メタ情報抽出ツール）を、計測機器メーカー2 社（アルバック・ファイ株式会社、株式会社リガク）の協力のもとで開発しました。AI や機械学習で利用しやすいデータの創出・蓄積を効率的に行えるようになり、データ科学を活用した材料開発の促進が期待されます。



2．現在、材料データを機械学習によって統計処理し新材料の開発を目指すデータ駆動型の材料開発が注目を集めています。しかし、統計処理の元となる計測データの多くは、同一のメーカー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りました。また、ファイルに計測条件などのメタ情報が記録されていないため、対象とするデータの検索も難しく、機械学習で利活用しやすいデータ形式へ変換するツールの開発が求められていました。

3．今回、材料評価で広く用いられている 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、計測メーカーの協力のもと、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開発しました。まず第 1 弾として、アルバック・ファイ社 Quantera SXM 等のファイル形式で生成された XPS スペクトル、およびリガク社 SmartLab のファイル形式で生成された粉末 XRD パターンの計測データに対応しています。今後も引き続き計測メーカーの枠を広げるとともに、XRD や XPS に限定せず、対応する装置や対象とする計測技術分野の拡大を図っていく予定です。

4．メタ情報抽出ツールのほか、バイナリデータのテキスト変換ツール※4 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への視覚化変換ツールをあわせ、“M-DaC（Materials Data Conversion Tools）”と命名して NIMS-DPFC のウェブサイトにて公開します。M-DaC のソースコードの一部は MIT ライセンスのもと、利用者自身で改良することも可能です。また、装置が出力したサンプル用生データも公開しており、「クリエイティブ・コモンズ・ライセンスの表示 - 非営利 4.0 国際（CC BY-NC 4.0）」のもとでの利用が可能です。

M-DaC 公開ページ：https://www.nims.go.jp/MaDIS/about/M-DaC.html

5．本成果は 2019 年 1 月 30 日から東京ビッグサイトで開催される nano tech2019 及び同時開催の MaDIS シンポジウム 2019~AI で加速する材料開発とデータプラットフォーム戦略~にて発表されます。

研究の背景

国立研究開発法人 物質・材料研究機構 （NIMS） 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（DPFC）では、一般公開を目指した材料研究データバンクとその活用プラットフォームの構築を 2017 年 4 月に開始しました。目的の一つは、機械学習や AI による材料探索の効率化をはかることです。そのためには、コンピュータ群を中心に据えたハードウェアのプラットフォームが必要であると同時に、そこで公開利用されるデータには、FAIR 原則と言われる Findable（検索可能）、Accessible（アクセス可能）、Interoperable（相互運用可能）、Reusable（再利用可能）が要求されます。

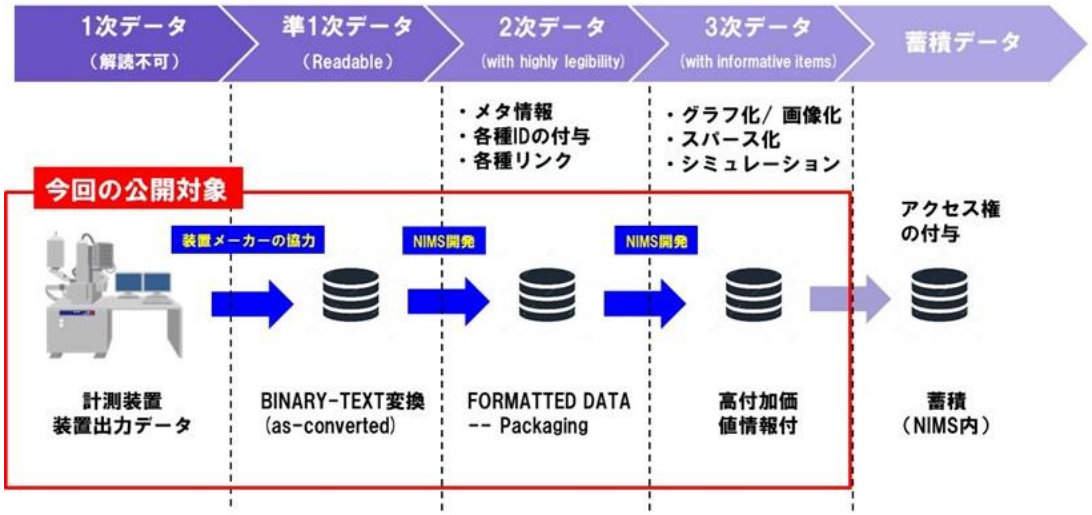
計測データは、バイナリ形式やテキスト形式※4、またはその混合形式で装置から出力されます。それらを共通の形式に変換したとしても、そこで使用されている用語や数値の扱いは、装置（またはメーカー）独自の表現が用いられており、そのままでは第三者可読性は非常に低いものとなっています。

その形式を第三者可読性が高いものに変換する際には、ISO 規格に代表されるような *de jure* 標準と呼ばれる特定のコミュニティの関係者間でコンセンサスが得られた方式に従うのが非常に扱いやすいのですが、標準の合意形成には長期間を要するとともに技術の進歩にあわせた更新（アップデート、メンテナンス）が難しいのが実情です。

そこで、NIMS-DPFC では、ファイル様式に自由度が高く装置メーカーの負担も少ない Schema-on-Read 方式※5 で計測データを変換する研究開発を進めています。

研究内容と成果

協力を得られた装置メーカーには、バイナリデータの変換ツールの提供や、データファイルに記述された測定にかかるパラメータの意味を開示していただいています。NIMS 側では各々のデータに高速な検索を可能にするメタ情報※1 を付加し、相互運用（IEEE の定義では「2 つかそれ以上のシステムまたはコンポーネントが情報交換でき、また交換した情報を使用できる能力」）するために数値データを人間が容易に理解し、かつ機械（コンピュータ）が解読できる形式に変換するツール、さらにデータの生成や変換に無関係な第三者が独自の目的のために再利用することを容易にするために主要な計測パラメータを抽出するツールなどを Materials Data Conversion Tools（M-DaC）と命名し開発しています。



異なる形式とデータを比較する必要のない環境でも、メタ情報を抽出することで膨大な数の計測データ群からの特定データファイルの検索を容易にすることに利用することも可能です。

現状では、アルバック・ファイ社 Quantera SXM の X 線光電子（XPS）スペクトル、XPS の深さ方向プロファイル、およびリガク社 SmartLab の粉末 X 線回折（XRD）パターンに対応している変換ツールを NIMS-DPFC のウェブサイト（https://www.nims.go.jp/MaDIS/about/M-DaC.html）から誰でもダウンロードできます。

今後の展開

データ変換を可能にする技術分野を装置メーカーの協力やオープンソースソフトウェアを活用して拡大し、変換ツールを順次公開していくことにしています。また、現在 DPFC で開発を進めている新しいデータリポジトリに搭載されることにより、データ創出から蓄積までが人手を介さずに一気通貫で行えるようになり、今後、AI に資する計測データの利活用に供されることにより、より効率的でより高度な材料開発の促進が期待されます。

M-DaC の公開サイト

https://www.nims.go.jp/MaDIS/about/M-DaC.html

用語解説

※1　メタ情報

メタ情報とは、データについて、そのデータ属性や関連情報などを与える情報を指します。データのデータと言われることもあります。図書館の書誌情報を例にとれば、著者名、書名、出版社、発行年などが書籍を表すメタ情報です。膨大なデータを適切なメタ情報を付与させて管理することで、データベースではこのメタ情報を活用させることにより、効率的にかつ精度よく目的のデータを検索することが可能となります。

※2　機械可読性

機械可読性とはコンピュータによる文書構造の認識しやすさを示し、機械可読性に配慮した文書やデータはより検索されやすく，利用されやすくなります[1]。文章やデータのファイルの形式では、よく知られた txt 形式や csv 形式がありますが、これらの形式は機械でファイルそのものは読み込むことはできても、機械がその中身の文章構造やデータ構造を認識すること、すなわち機械可読をしていることにはなっていません。機械可読性のあるファイル形式にするにはメタ情報を与え、XML 形式※3 などへ変換する必要があります。

[1] 中西 秀彦 : ”人間可読性から機械可読性の時代へ　XML 組版への製作現場からの提言”, 情報管理, 57 巻, 3 号, p. 149-156, 2014.

※3　XML ファイル

広くインターネットで使われている HTML（Hyper Text Markup Language）は web ページの表示の特化した言語ですが、表示させる文章の意味や重要度などを機械は理解できるようにできていません。一方、XML(Extensible Markup Language)は、データの要素や属性などを定義することで重要度も機械が認識できるようにしたデータ記述用の言語です。その記述方法は、<XXX>と</XXX>のようにタグと呼ばれる山括弧で文章や用語をくくることにより、その文章や用語の属性もしくは修飾情報をタグの中の XXX で規定させることができる仕組みです。このタグの属性情報としてメタ情報を与え、データ構造を XML で記述すれば、人間に可読性のあるテキスト形式のデータがそのままコンピュータで処理できることも備えさせられるため、柔軟性のあるデータ管理とデータベースの構築のほかにコンピュータ間やデータベース間でのデータ互換性も容易になります。

※4　バイナリ形式とテキスト形式

バイナリ形式とは「0」と「1」で表記される 2 進数で記録されたデータ様式を指し、コンピュータの処理はこのバイナリ形式で行われます。バイナリで記録された形式のファイルは、コンピュータにとっては容易に解釈できますが、人間にとってはその意味や内容を解釈することは極めて困難な形式です。そのため、データは、人間が解釈できるようにバイナリ形式からテキスト形式へと変換されます。テキスト形式とは、まさに名前の通り人間が可読できる文字データからなる形式で、アルファベットや数字などのほか、日本語の場合にはその種類にはひらがな、カタカナなども含まれます。

※5　Schema-on-Read 方式

データのスキーマ（構造）を事前に明確に構造化してデータベースを構築してゆくことを Schema-on-Write

方式と呼び、MySQL のようなリレーショナル・データベースの構造化されたデータベースでは広く使われている方式です。一方、本ツールで採用している Schema-on-Read 方式は、データを保存する際にデータの加工は不要、または最低限でよく、非構造化データに対応し、要素の定義も柔軟にできる利点があります。両者は二律背反的ではなく、データの構造が分かっているときには、Schema-on-Write 方式は圧倒的に処理が容易です。ただし、全ての要素を厳密に定義する必要があるため、Schema-on-Write 方式の確立には相当の時間が必要であり、本ツールが対象とする複雑な装置データの場合には、メーカーへの負担も大きくなるというデメリットがあります。

本件に関するお問い合わせ先

（研究内容に関すること）
国立研究開発法人物質・材料研究機構統合型材料開発・情報基盤部門
材料データプラットフォームセンター　副センター長
吉川英樹（よしかわ　ひでき）
E-mail: YOSHIKAWA.Hideki@nim.go.jp
TEL: 029-859-2451

（報道・広報に関すること）
国立研究開発法人 物質・材料研究機構　経営企画部門　広報室
〒305-0047　茨城県つくば市千現 1-2-1
TEL: 029-859-2026、 FAX: 029-859-2017
E-mail: pressrelease@ml.nims.go.jp