2022 年 3 月

マテリアル DX における計算データリポジトリに関する提言書

計算物質科学協議会（CMSF）は、我が国の最先端の計算物質科学技術を振興し、世界最高水準の成果創出と、シミュレーション技術、材料情報科学技術の社会実装を早期に実現するために、2020 年 5 月に設立されました。CMSF は計算物質科学技術に関わる全ての方々に開かれており、2021 年 12 月末時点で産学官の機関からの参加者 120 人（内企業会員 20 人）、参画機関数 56 機関（大学・国研等：38 機関、民間企業：18 社）の会員で構成されています。運営は、物質科学分野向けのスーパーコンピュータ共同利用・共同研究拠点である、東北大学金属材料研究所、東京大学物性研究所、自然科学研究機構分子科学研究所と、教育拠点である大阪大学ナノサイエンス教育研究センターの 4 機関が担っています。主な活動としては、我が国の学術振興と産業発展において、計算物質科学界が果たす役割と進むべき方向性について議論を重ねて意見集約を行い、国や関係諸機関への提言活動を実施しています。

スーパーコンピュータ「富岳」が稼働し、成果創出加速プログラムの発足、さらには共用利用が開始され、物質科学シミュレーションと AI・データサイエンスの融合により、新たな知識・価値の創造が益々期待されていた 2020 年度の 8 月に、CMSF より文科省に提言書を提出しました。そこでは、「富岳」をはじめとした高性能・大規模計算資源を活用した膨大な高精度物性データの創出、日本で開発されている最先端の計算物質科学シミュレーション手法やソフトウェア群、また、データ駆動の材料情報科学技術を、学界や産業界が一体となって迅速に開発し、活用することの重要性について述べました。

その後、2021 年 4 月に示された第 6 期科学技術・イノベーション基本計画には、超スマート社会（Society5.0）の実現に向けた政策のひとつとして、マテリアル革新力強化戦略が挙げられています。また、2021 年度に新たに「富岳」の成果創出加速プログラムに採択された 3 課題中 2 課題がデータ駆動型の物質材料分野であり、計算物質科学を基盤とするデータ駆動型の物質材料研究への期待が高まっています。世界一の計算性能を誇る「富岳」を中核とした HPCI 体制のアドバンテージがある今こそ、計算データの活用で世界をリードしていかなければなりません。そこで、CMSF では実際に最前線でデータ駆動型マテリアル研究を担っている若手や企業の研究者を中心としたワーキンググループを立ち上げ、計算物質科学界における研究開発の方向性とデータリポジトリのあり方について検討を行い、CMSF 運営委員会での審議を経て、以下の項目として新たな提言書をまとめました。これから加速されるデータ駆動型マテリアル研究の国際競争に、どのようにすれば世界をリードしていけるかをまとめています。

1．計算物質科学界におけるデータリポジトリの世界および日本の情勢
2．「富岳」成果創出加速プログラムでの研究データマネジメント状況
3．計算、合成、計測のデータ融合と利活用を考慮したデータ同化技術
4．産官学で活用される計算物質科学データリポジトリの在り方
5．国産ソフトウェアパッケージ開発の重要性
6．「富岳」を頂点とする大規模計算機に立脚した材料データの自動創出
7．計算物質科学コンソーシアムの育成とそれを基盤とするデータリポジトリの創出

1. 計算物質科学界におけるデータリポジトリの世界および日本の情勢

マテリアル DX に求められるデータとデジタル技術を活用したマテリアル研究の革新を効果的に進めるためには、高品質な研究データの収集・格納・管理・活用が必要となる。そのため、データ駆動型の研究様式を担保する利便性・拡張性・セキュリティを兼ね備えたリポジトリ（データプラットフォーム）の構築は、マテリアル DX の基幹となるインフラ事業であり、戦略的に進めなければならない。国外に目を向けると Dryad や Zenodo 等の研究領域を問わない膨大なデータ量を持つ一般的リポジトリが存在する。また、計算物質科学に関係した分野別リポジトリでは、汎用ソフトウェアによるデータフォーマットに特化した NOMAD、AFLOW、OQMD、Materials Project 等が世界的なスタンダードになりつつある。これらは Application Programming Interface（API）機能を提供しており、ユーザーは容易に材料データをリポジトリからダウンロードし機械学習やスクリーニングを実施することができる。一方、それらのデータベースで公開されている電子状態に関するデータは限定的であるため、データ駆動型の物質探査に十分とはいえない状況にある。我が国でも、汎用データは開示するが新規マテリアル開発につながるような貴重なデータは一般には開示しないのが一般的であり、データの価値ごとにオープン・クローズ戦略を慎重に検討していく必要がある。

我が国では、「マテリアル DX プラットフォーム構想実現のための取組」において、NIMS を中核拠点とし全国の材料データを集約する取り組みを進めている。しかしながら、マテリアルに関わる計算データをどのように収集していくかは、議論の余地を残している。上述した海外の分野別リポジトリと差別化を図り、より利用価値が高いリポジトリを構築するためには、独自の視点や方法論によるユニークなマテリアルデータを創出し、収集していく必要がある。海外の分野別リポジトリは単体物質や単純化合物に対する構造・電子状態データを中心に収集している。幸い、我が国では「京」コンピュータ関連事業などにおいて開発されている大規模高精度計算が可能な国産のソフトウェアと、「富岳」を頂点とする「HPCI」、「共同利用スパコン」等の大型計算機を有する。このソフトとハードを組み合わせ、ユニークなマテリアルデータを創出することができれば、世界的にも価値の高いデータを蓄積することが可能である。さらに戦略的に開発するマテリアルに関したデータベースを構築し、そのデータとソフトウェアを適切にオープン・クローズ戦略により運用することで、我が国の計算データリポジトリの利用価値を高めると同時に、国際的なマテリアル研究・開発を牽引することが期待できる。

2. 「富岳」成果創出加速プログラムでの研究データマネジメント状況

2020 年に開始した「富岳」成果創出加速プログラムプロジェクトでは、得られた計算結果等のデータマネジメント方針（データの保存・管理・利活用）について、内閣府の「研究データリポジトリ整備・運用ガイドライン」に沿ったデータマネジメントプランの作成が求め

られている。CMSF はマネジメントプランの作成・運用状況を把握し分析するため、2020 年度スタートの物質・材料系 5 課題に対してアンケート調査を行った。その結果、策定したマネジメントプランの運用を実際に開始しているのは「量子物質の創発と機能のための基礎科学 ―「富岳」と最先端実験の密連携による革新的強相関電子科学」と「大規模計算とデータ駆動手法による高性能永久磁石の開発」の 2 課題のみであった。その理由を考察した結果、推測される原因として下記があげられる。

(I) 多様な機能性材料を扱っており、各材料によって着目する物性が異なる。さらに、材料の種類や着目する物性によって、計算に用いるソフトウェアが異なる。そのため、プロジェクト内においても共通フォーマットでのデータ収集が非常に困難となっている。
(II) 計算物質科学に関わる研究者間での、データ整理手法やメタデータ作成に対する統一見解が無い。すなわち、どのようにメタデータを作成すると機械学習の実行が容易になり、効率よく材料開発が加速するか等についての体系的な整理が成されていない。
(III) 時間・コストをかけデータを整理し、他の研究者、または、産業界を対象に公開したとしても、現状ではインセンティブが得られる仕組みがない。
(IV) マテリアル DX の掛け声が先行してデータマネジメントの重要性のみが伝えられ、現場レベルでのデータ利活用の意義や出口などが十分理解されておらず、データ収集に向けたモチベーションが上がらない。

これらの問題を解決するためには、中長期的に継続した DX 人材の育成・教育、データ創出者の業績への適切な評価、オープン・クローズ戦略の明確化と周知、政策立案者または統括者のマテリアル DX に対する理解向上とその共有が必須と考えられる。

また、2021 年度からは、新たに「データ駆動型高分子材料研究を変革するデータ基盤創出」と「「富岳」を活用した革新的光エネルギー変換材料の実現」の 2 課題が発足した。特に前者では高分子材料を対象とし、物性計算の自動化のフレームワークやデータ科学に資する高分子物性データベースの構築を目的としており、産学連携によるデータベース共創の実現を目指している。

3. 計算、合成、計測のデータ融合と利活用を考慮したデータ同化技術

データ駆動型のマテリアル研究やマテリアルズ・インフォマティクスの発展に伴い、研究室レベルで購入可能な計算機で時間のかからない、低コストの物性科学シミュレーション手法を用い、データの数を稼いで物性データベースを構築するケースが多くなっている。目的は、構築したデータベースを元に「機能」→「材料」へと至る逆問題を解くことである。し

かし、計算手法や計算に用いるモデルが低コストであればあるほど、物性科学シミュレーションで得られる近似・数値計算手法等に起因したエラーや現実との乖離がデータの中に内在していることを認識しなければならない。このエラーの問題は、計算手法の発展により改善することが可能である。さらに近年では、シミュレーションと実験のデータを融合させるデータ同化を用いるというアプローチも進展している。データ同化は、元来、気象学の分野で用いられ、計測データを利用することでシミュレーションの精度を向上させることが目的であった。この技術を物質・材料の分野に応用し、シミュレーションと実験データの相関性を見出し、未知の実験データを予測する試みが行われている。例えば、「富岳」成果創出加速プログラム「大規模計算とデータ駆動手法による高性能永久磁石の開発」では、計算と実験のデータ同化により、利用環境に応じて最適な特性を持つ永久磁石の物質構造や組成のオンデマンドなデザインを、効率的に行うことに成功している。

このデータ同化の手法を推進し、我が国の材料開発速度を大幅に加速するためには、「富岳」等の大規模計算機によって創出される物性データに加え、大規模実験施設群（SPring-8、J-PARC 等）等で計測される実験データとの融合が非常に重要となる。元素戦略プロジェクト＜拠点形成型＞においては、計算と計測のデータ同化による研究が加速されたが、その仕組みをシステム化するのは次の段階への期待となっている。このデータ同化の仕組みをシステム化するためには、国家基盤として整備されたデータ活用社会創成プラットフォーム（mdx）を活用した計算・実験データそれぞれのリポジトリを共有するシステムを構築し、さらに学術情報ネットワーク「SINET6」を用いてリアルタイムでデータ解析・同化が可能な仕組みを整備することが望ましい。このようなシステムが構築できれば、単にデータ同化に則った研究が実施されるだけでなく、計算、計測、合成のデータリポジトリを連携させたプラットフォーム上で、研究領域を超えたコミュニティーの醸成が促進され、世界に先駆けた独自のデータ駆動型マテリアル研究の推進が可能となる。

例えば、国立研究開発法人新エネルギー・産業技術総合開発機構（NEDO）が推進する超先端材料超高速開発基盤技術プロジェクト（超超プロジェクト）では、高機能材料・部材の研究開発支援を可能とする高度な計算科学、高速試作・革新プロセス技術、先端ナノ計測評価技術を駆使した革新的な材料開発基盤の構築を目指して、企業 18 社からなる先端素材高速開発技術研究組合（ADMAT）との産学連携でデータ創出及びデータ同化技術の利用が進みつつある。しかし、さらにデータ活用を物質材料分野全体で進めるためには、順次、リポジトリ及びデータベースをオープンとして研究コミュニティーの拡大を促す努力が必要不可欠である。

4. 産官学で活用される計算物質科学データリポジトリの在り方

計算物質科学により得られた物性データベースは価値ある研究資源であり、長期間保管可

能な堅牢、かつ、安定的な体制のもと管理する必要があり、そのためのハードウェアなどの維持費、および、それらを運営する人的コストは継続的に担保しなくてはならない。近年の計算物質科学分野では、国家プロジェクトなどを通し産官学が連携したデータベースの構築も進められており、秘匿性の高いデータの取り扱いにも注意を払う必要がある。全ての大学・機関がデータリポジトリを有し、これらに対応するには莫大なコストが求められ現実的ではない。各分野の代表機関が共同利用の枠組みの一環としてデータリポジトリを運営・提供し、各分野のニーズ収集、および、シーズの提供を行う仕組みの構築が求められる。

データベースの利活用という観点からは、データリポジトリは高いユーザビリティを備える必要があり、データベースへのアクセスのしやすさ、所望するデータの探しやすさ、取得したデータの解析のしやすさは特に注意を払うべき事項である。特に、データベースを検索するためのポータルサイト構築、データへのメタデータ・タグの付加、データを CUI・GUI ベースで取得できるシステムの構築、データ解析するためのツールの開発・提供などが重要となる。普及という観点からは、講習会やデータベースを活用した実習型の研究会の開催なども、データ駆動型マテリアル研究の振興推進という観点からは必要となる。

5. 国産ソフトウェアパッケージ開発の重要性

マテリアル DX が進む中、データ駆動により新機能材料の開発を効率よく実施するため、物性データの収集・整理が重要視されるようになったが、データを創出する基盤はソフトウェアであることを認識しなければならない。国が公的に支援する各種プロジェクトにおける計算物質科学界の役割の一つは、世界をリードする高度な数値計算アルゴリズム・ソフトウェアを開発し迅速に社会実装を行うことである。そして、産業界での運用とフィードバックを通じ、ニーズを的確に把握しソフトウェアのさらなる機能向上を行う。このような研究環境を構築し運用することで、計算物質科学界の強化と、我が国の産業発展への貢献が期待される。

しかし、現状では、産業界において国産のソフトウェアを用いて研究開発を実施している企業はごく少数であり、大多数は材料物性の解析やデータ収集の目的に VASP 等の国外ソフトウェアを用いている。ブラックボックス化されたソフトウェアでは、新機能材料や新奇物理・化学現象が発見された時、解析するためのアルゴリズムを迅速に実装するのは困難であり、鍵となるデータの創出も先行する国外グループに先を越されることが容易に想像できる。

計算物質科学界において、大学・研究所・企業で日常的に使用されている物質科学ソフトウェアは、欧米において開発されたものが多い。その理由は、ヨーロッパにおいて 1994 年に計算物質科学コミュニティーである「Psi-K ネットワーク」が立ち上がった例に見られるよ

うに、戦略的かつ長期的な視野を持って電子状態計算ソフトウェアの開発が実施されてきたことにある。現在、巨大コミュニティーに成長しており、チュートリアルやワークショップの開催、そして、若手キャリアパスのサポートを行ってきた実績を持つ。その結果、VASP等のソフトウェアが世に知れ渡り、計算物質科学界で世界的スタンダードになっている。また、古典分子動力学ソフトウェアの LAMMPS は、米国 DOE 傘下の研究所で 1990 年代半ばから開発が始まり、現在でもサンディア国立研究所を拠点にワークショップなどを通じてユーザコミュニティーを拡大してきた。機械学習やデータ科学が注目を集めている中、欧米では現在も地道に、かつ、着実に基盤ソフトウェアの開発が継続されている。

世界的に認識されうるソフトウェアの開発には数十年の期間を要し、さらに維持するためのコストも継続的に必要となる。我が国においても「京」、および、ポスト「京」関連プロジェクト等で、約 10 年間、国産ソフトウェアの研究開発に力を注いできた。しかし、現在は「成果創出加速プログラム」として成果の創出に力が注がれており、ソフトウェアの開発や普及に関して実施するリソースが無いのが現状である。

ソフトウェアは物質科学の進化と社会ニーズの劇的な変化に伴い複雑化し、開発を継続するためには物質科学・数値計算・大規模計算機に渡る幅広い専門的な知識・技術を持ったチームが必要である。つまり、普及を目的とするソフトウェア開発は個人や小規模グループでは不可能である。「成果創出加速プログラム」においては、それまで開発してきたソフトウェアを活用し、データ駆動型の研究開発を特定課題に対して実施しているが、このプロジェクトで培ったノウハウを広く材料開発全体に普及させていくには、「京」の際に実施した、研究課題と学術振興の課題を平行させ、複数の研究課題のノウハウを統合してソフトウェアを普及させていく 10 年規模のプロジェクトが必要と考えられる。この中で、戦略性を持った国産ソフトウェアを選定し、計算、計測、合成のデータ同化技術を駆使し、HPCI や共同利用スパコン、mdx、SINET6 の学術基盤施設を活用することで、海外からも着目されるシステムの構築が可能である。その中で、海外製のソフトウェアやデータベースの活用も検討していくことが望ましい。

さらには、既存ソフトウェアのアウトプットを利用することで、異なったスケールでの有効モデルを構築し、階層を跨いだ物理量を計算する、いわゆる多階層連結シミュレーションを実現するソフトウェアとそのインターフェースの整備も非常に重要である。このようなシミュレータはより現実の材料を模したシステムや材料開発に直接関係する物理量を計算できるため、マテリアル DX を推進するにあたって非常に有用なデータを創出可能となる。

一方で、これらのソフトウェアの多くは最先端の研究を目的に開発されたものであり、必ずしも多くのユーザーを想定したものにはなっていない。ソフトウェアのユーザビリティを

高めることで多くのユーザーを獲得し、より多くのデータの収集と充実したデータベースの構築へと繋げることができる。これらのソフトウェア開発・データベース構築を核にし、産学を含めたコミュニティーを形成・発展させることで、国際的な競争力向上も期待される。今後は研究向けのソフトウェア開発を応用・普及向けに発展させることにも注力し、そのソフトウェアを適切に支援していくことが強く求められる。

6. 「富岳」を頂点とする大規模計算機に立脚した材料データの自動創出

現在、計算、計測、合成を問わず、マテリアルの研究データ量はビッグデータと呼ぶには程遠い。データ量の少なさゆえマテリアルズ・インフォマティクスの手法を有効に活かせず、広範囲の材料空間を効率的に探索できていなかった。その課題を打破するための「富岳」のような超大型計算機と計算物質科学ソフトウェアを用いたデータ創出は、この「スモールデータ問題」を打破するのに非常に効果的であることを強調したい。「京」ではシステムサイズと計算時間の 2 軸に重点が置かれた運用方針であったが、「富岳」においては 3 つ目としてマテリアルパラメータ（構成原子種・組成、構造等）の軸が追加され、マテリアルデータの網羅創出を推奨する運用方針となっている。実際に、成果創出加速プログラムの「大規模計算とデータ駆動手法による高性能永久磁石の開発」では、「富岳」上で自動網羅計算ツールを適用することにより、約 15 万種類の磁性合金材料の電気、磁気、伝導特性に関し、一週間以内で収集することに成功している。

「富岳」の計算能力と特性を生かしデータ創出を戦略的に進めるためには、計算データの再現性を担保し、かつ、ユーザーフレンドリーな自動計算インフラストラクチャの整備が最重要課題であろう。また、大学と企業が一体となってマテリアルズ・インフォマティクスに資するデータを創出する際は、データに関する権利やルールの整備化は必須事項であり、データのオープン・クローズ戦略を練る必要がある。国外では、計算科学の複雑なワークフローの自動化を支援するインフラストラクチャ「AiiDA」を利用した網羅計算が積極的に行われている。「AiiDA」は、すべての計算履歴、ワークフロー、入力、出力、メタデータを自動的に追跡し、研究の再現性を FAIR 原則に則った形で実現している。このようなフレームワークの利用も、マテリアル DX を進めるにあたって選択肢の一つである。我が国では、AI の利用や研究の自動化に対して嫌悪感を抱く研究者も少なくない。しかし、これからの知識・デジタル社会においては、既存研究スタイルの変化（リサーチ・トランスフォーメーション）は避けて通れない。ハイスループット計算・材料創製・計測によりデータ創出・収集を行い、研究者はそれらで得られたデータを活用し、より創造性やオリジナリティに重点を置いた研究に時間を割くのが、マテリアル DX における国際競争力強化の鍵となる。

7. 計算物質科学コンソーシアムの育成とそれを基盤とするデータリポジトリの創出

近年、計算物質科学分野ではプロジェクトで創出された独自性の高い新たな国産ソフトウ

ェアを基盤とし、産学官のメンバーが参画したコンソーシアムの設立が活発になってきている。特に、開発ソフトウェアの活用によって、コンソーシアムに参画する研究者がそれぞれの異なる研究目的に従って多くのデータ創出を行うことで、データリポジトリの構築を進める活動も活発化してきている。このような流れは、*5.* で述べた独自性の高い国産ソフトウェアの開発が、計算物質科学分野におけるデータリポジトリの構築に必要不可欠であることを示している。また、コンソーシアムの設立は、ソフトウェアのユーザー数の増加のみならず、継続的なソフトウェア開発および新たなコミュニティーの形成・若手人材の育成へと結びつくことも大いに期待される。ただ現状では、プロジェクト終了後のデータの維持・管理、コンソーシアムの運営などは、ボランティア活動に頼らざるを得ない部分があることも事実である。その解決案の一つとしては、スーパーコンピュータ共同利用・共同研究拠点のような各分野の代表機関において、中長期的なプロジェクトを通した新たな国産ソフトウェア開発、それを中心としたコンソーシアムの構築によるデータ創出やデータ利用、そして共同利用・共同研究拠点のスーパーコンピュータを活用したデータ管理とデータ更新といった流れを構築することも、長期的な観点からは必要と考えられる。上記のような施策によって、*4.* で述べた長期的なデータリポジトリの管理・維持に加えて、国産ソフトウェア開発、人材育成なども長期的なビジョンを持って、実現が可能になると期待される。
さらに、上記のようなソフトウェアを中軸としたデータリポジトリに、SPring-8、J-PARC、SACLA、次世代放射光などによる計測データや実験データを組み入れた計算・計測・実験が一体となったデータリポジトリの構築は、個別のコンソーシアムのみでは難しい部分が多く、スーパーコンピュータ共同利用・共同研究拠点のような各分野の代表機関が、NIMSとも連携をとりながら、その役割を担っていくことが必要と思われる。

さらに、個別のソフトウェアを通した個別のコンソーシアムの設立やそれを通した個別のデータリポジトリの構築のみでは、分野融合や新たな研究分野の創成への発展が難しいという側面も懸念される。独自性の高い新たな国産ソフトウェアを基盤として設立されたコンソーシアム間の連携や融合を進めることで、分野融合や新たな研究分野の創成を促進することも、マテリアルズ・インフォマティクスにおける日本の国際競争力を強化するための重要な課題であると考えられる。スーパーコンピュータ共同利用・共同研究拠点のような各分野の代表機関が、NIMS とも連携をとりながら、このような分野融合や連携に重要な役割を果たすことが必要になってくると考えている。このような施策を通して、日本で開発された特徴的なソフトウェアに基づくマテリアルズ・インフォマティクスの存在感を世界に対して示していくとともに、長期的な視野で日本のマテリアル・インフォマティクス技術の育成、さらにはそれを発展・促進させてくれる若い人材の育成を進めていく必要がある。

【計算物質科学協議会・提言書作成ワーキンググループ】
福島 鉄也（東京大学）
吉見 一慶（東京大学）
大谷 優介（東北大学）
南谷 英美（分子科学研究所）
濱田 幾太郎（大阪大学）
佐伯 昭紀（大阪大学）
旭 良司（名古屋大学／元（株)豊田中央研究所）
古山 通久（信州大学）
藤井 幹也（奈良先端大／元 パナソニック(株)）
茂本 勇（東レ(株)）
岩崎 誉志紀（太陽誘電(株)）

【計算物質科学協議会・運営委員／相談役】
（運営委員）
江原 正博（自然科学研究機構）（協議会副代表）
小口 多美夫（大阪大学）
尾崎 泰助（東京大学）
川勝 年洋（東北大学）
川島 直輝（東京大学）
久保 百司（東北大学）（協議会代表）
斉藤 真司（自然科学研究機構）
森川 良忠（大阪大学）
旭 良司（名古屋大学／元（株)豊田中央研究所）
尾方 成信（大阪大学）
片桐 孝洋（名古屋大学）
茂本 勇（東レ(株)）
天能 精一郎（神戸大学）
藤井 幹也（奈良先端科学技術大学院大学／元 パナソニック(株)）
松林 伸幸（大阪大学）
三宅 隆（産業技術総合研究所）
（相談役）
赤井 久純（東京大学）
岡崎 進（東京大学）
田中 功（京都大学）
常行 真司（東京大学）