材料データベース構築とデータの統合活用

（物材機構）石井 真史
Material database construction and integrated utilization of data (*MaDIS, National Institute for Materials Science* ○Masashi Ishii

NIMS has been constructing various databases on the background of data-driven R&D. In addition to the strengthening of the historical database called MatNavi, we are planning to promote data integration and knowledge creation to go beyond the limit of activities of a single institute, aiming to provide an overview of materials science and induce topical innovations everywhere. The goal is to induce localized innovations everywhere. We present our efforts to realize data structure and data linkage including huge background knowledge, which is the base of machine learning and artificial intelligence.
Keywords : Database; Data-driven; Data Structuring; Data Integration; Knowledge Base

NIMS で公開しているデータベース MatNavi（https://mits.nims.go.jp/）は、遠く過去を振り返ると、NIMS の前身である金属材料技術研究所（National Institute for Materials, NRIM）の原子力材料、超伝導材料、構造材料 (クリープ、疲労、宇宙材料関連強度および腐食)に関するプロジェクト、新技術開発事業団（現　科学技術振興事業団(JST)）の高機能物質、研究情報のデータベース化事業を受け継ぎ、2003 年に一つの材料データ基盤として web 公開され、現在まで構築を続けてきた。この間のコンピュータ、データベース、通信技術の発展は、データベースの提供方法はもちろん、データ構造自体にも大きな影響を与え、巨大データになればなるほど、過去と現在の技術的な整合に苦労し、一方では新しい技術の導入方法を模索してきたといえる。特に最近のデータ駆動型研究への対応は、データベースの在り方に、大きな変化をもたらした。ここでは、いくつかのトピックスを挙げて、最近の NIMS の取り組みを紹介する。

(1) MDR XAFS データベース（https://doi.org/10.48505/nims.1447）

どのインフォマティクスにおいても良質なデータの確保は、その後の数理学的なデータ処理の大前提であるが、容易ではないと考えられている。化学分析分野でもデータの流通・再利用を検討する動きは少なくないが、X 線吸収分光（X-ray Absorption Fine Structure, XAFS）については、手法の性質上データの比較が必須であり、データ共有の声が高いため、NIMS の材料データリポジトリ（Materials Data Repository, MDR）を高度に利用したデータベース MDR XAFS DB の構築が進んでいる。ここでは、放射光施設や関連の大学から XAFS スペクトルデータを提供いただき、横断検索・一括ダウンロードなどができるような仕組みを構築した。データをそのまま登録するだけでは単なるオンラインストレージに過ぎず、解読不能なスペクトルが使うあてがないまま放置されてゆく傾向があるが、本取り組みでは機械可読なメタデータの付与、材料名の辞書による名寄せを実現し、機関の違いを感じることがない検索性が極めて高いデータベースを実現した [1]。一旦こうした仕組みが出来上がると、同様な取り組みをしている外部のデータベースとの連携が可能になる。実際、HAXPES データベー

ス、XANES Bibliography などと連携し、内殻を俯瞰する知識基盤構築に向けて動き始めている。こうした統合データベースの考え方は、今後の化学分析におけるデータ駆動の基礎技術になると考えている。

(2) 高分子データベース PoLyInfo の知識ベース化

PoLyInfo（https://polymer.nims.go.jp/）は、2023 年 1 月現在データ数では、ホモポリマー数 18,697、コポリマー数 7,736、物性ポイント数 494,837、文献データ数 21,055 の世界でもあまり例がない、まとまった高分子データを収録したデータベースである。平易な PoLyInfo と機械学習の解説は、ちょうど発刊される参照文献 [2]をご覧いただくのが良いと思うが、その中で述べた通り PoLyInfo をデータ駆動研究に活用するには、いくつか課題があると考えている。PoLyInfo は主に一次構造を重視したデータベースであり、化学的に新しいポリマーの探査などには向いている反面、高次構造の機械可読化は十分ではない。また、複合材料といった産業上重要な領域になると、おそらく外部のデータベースのほうが優れていると考えている。こうした状況にあって本データベースは、これまでのデータ収集方針は維持しつつ、将来的な展開を試みている。言い換えるならば PoLyInfo の独自性は保ちつつ、連携できる部分では連携し拡張的にデータ駆動研究の基盤を作ることを考えている。具体的には

- 生分解など、環境への負荷を低減するためのデータ提供
  生分解性を持つ細菌と分解対象のポリマーの組み合わせの定式化。細菌を含む生物科学系の外部の大きなデータベース（例えば NBCI, The National Center for Biotechnology Information）および関連データベースと PoLyInfo との統合検索環境の構築と検証
- バイオポリマーとの共創的研究基盤の構築
  Spider silk などバイオポリマーと PoLyInfo の合成ポリマーとの比較からの課題抽出など広い観点での材料科学の俯瞰。複合材料を含めた生体材料との親和性の向上。
- データベースの知識基盤化
  PoLyInfo が持つ高分子物性や構造情報の網羅性を機械可読化することで高分子オントロジーを構築し、データを知識として扱えるようにすること

を進めている。

最初に述べた通り、データベースは技術の進歩とともに形を変えてきた。情報爆発ともいわれた、大量のデータや情報の全世界的な発信は、明らかに現在のデータベースに新しい役割や再生を求めている。殊に日本はデータ活用において、一歩遅れを取っているという意見も多い。データを創出する側も、公開する側も、統合や連携を意識した新しい視点が求められているであろう。

参考文献

1) M. Ishii, et. al., “Integration of X-ray absorption fine structure databases for data-driven materials science”, to be published.
2) 高分子科学最近の進歩「PoLyInfo と機械学習」石井真史、高分子 Vol. 72、2 月号 (2023).