利用性の高い材料データベース構築の仕方

徐 一斌　　物質・材料研究機構（NIMS）統合型材料開発・情報基盤部門　副部門長

《PROFILE》

略歴：

1994 年	中国科学院工学博士
1995 年	名古屋工業技術研究所（現 AIST）STA フェロー
2000 年	（株）シーティーアイを経て
2002 年より	NIMS
2007 年	名古屋大学情報学博士を取得 専門分野は材料データベースと材料熱物性

1　はじめに

近年，マテリアルズ・インフォマティクス（MI）の急速発展により，材料データに対する考え方および扱い方は大きく変わってきた。長期的なデータ蓄積と広い範囲のデータ交換を可能にする材料データプラットフォームの構築は，研究基盤の不可欠な一環として，多くの研究プロジェクトおよび材料研究者と技術者の日々の研究活動に取り込まれるようになってきた。しかし，これまでに開発された材料データベースを見ると，長期的に持続可能なデータベースはそれほど多くはない，プロジェクトの終了と共に廃棄されたデータベースも多数存在する。その原因は様々であり，例えば，予算や人員確保できないこと，データの使い道がなくなったことなどが考えられるが，なかでも，データの利用性の低さが最も主要な原因と考えられる。なぜなら，仮に様々な材料分野と目的で利用できる利用性の高いデータベースがあった場合，ユーザからのニーズが高いため，予算の確保や開発チームの維持が容易であり，長期的な存続が可能になるからである。実は，データの付加価値と利用性は，データの収集・編集の仕方により大きく変化する。本稿は，NIMS 無機材料データベース AtomWork-Adv[1] を一例として，材料定義と識別の視点から付加価値と利用性の高いデータ収集とデータベース構築の仕方について解説する。

2　AtomWork-Adv データとデータベース

AtomWork-Adv は，NIMS とスイスの MPDS 社が著作権共有している Pauling File[2] というデータセットを利用した NIMS が開発・公開している無機材料データベースである。Pauling File は，1995 ～ 2002 年までは科学技術振興機構（JST）と MPDS が共同開発し，2003 ～ 2015 年は MPDS が独自開発，2016 年以降は，NIMS と MPDS の協力による開発を行っている。そのデータは，多数のハンドブック，（Landolt-Bornstein，Handbook of Inorganic Substances など）とデータベース製品（Materials Platform for Data Science，PDF-4+/Web 2020，ASM Alloy Phase Diagram Database など）が含まれ，幅広く利用されている。AtomWork-Adv は，1900 年以来，1000 種を超える科学雑誌で発表された無機材料の状態図，結晶構造，および，約 500 種類の物性データを網羅的に収めている。2021 年 5 月現在の公開データ件数は，状態図 44,554 件，結晶構造 334,450，特性 410,401 件である。AtomWork-Adv の最大の特徴は，全ての材料が，物質レベル（化学式＋結晶構造）で識別できることである。この特徴により，異なる文献で発表された異なる材料の結晶構造や特性データなどを物質レベルで有機的にリンクすることができる。さらに，そのリンクは，AtomWork-Adv 内のデータだけではなく，外部データベースとの連携，例えば，ユーザ独自の材料データを AtomWork-Adv の結晶構造データとリンクすることも可能となる。

3　利用性の高いデータベース要件

MI は，ビッグデータ解析手法を用いた新材料探索と材料性能最適化を目的とした新しい研究手法であり，その原理は，目標となる材料の特性と，材料の化学組成，構造や作製条件などの影響要因との相関性を見つけ出し，材料設計による目標特性を実現することである。それを実現するために，材料の組成，構造，作製条件と特性データを有機的に統合する必要がある。しかし，研究目的や計測条件などの制限により，同じ材料に対して，全ての構造解析と特性測定を行うことはできない，よって，類似性のある材料から取得したデータを統合して利用することが必要である。そのため，一つの材料で取得したデータを多くの外部データと統合できるようにシステム化する必要がある。

統合可能な材料データベースの基本要件として，共通の材料識別システムが不可欠となる。材料の識別は，材料データベースの最も根本的な部分であるが，この材料の定義は意外と難しい。一つの材料を厳密に定義するためには，その化学組成，分子構造，結晶構造，ナノスケールからマクロスケールまでの組織構造を全部記述しなければならないが，それらを全て解析した材料はほとんどない。一方，異なる実験や生産プロセスで作製された材料は，全く同じものが存在しないので，材料データの統合は，ある程度の類似性を持つ材料を同じものとみなす前提で行うしかない。その類似性を判断するために，共通の材料識別システムが必要である。一部のデータベースは，独自の材料識別子，例えば，サンプル番号，ロット番号などを用いているが，それでは，外部データとの統合が不可能である。現在，材料データベースで最も多く使われている材料識別システムは，化学組成と化学式である。しかしながら，それだけでは材料特性に大きい影響を及ぼす相構成や，結晶構造などの情報が識別できないので，材料特性の記述子としては不充分である。

一般的に，無機材料の識別は，構成元素 → 化合物（化学式）→ 物質（化合物＋結晶構造）→ 材料（物質＋組織構造）の 4 レベルに分けられる。多くの材料に対して，厳密に識別できるのは，物質までである。物質レベルで材料を識別できれば，物性理論による物性計算や，単結晶の実験データとの比較が可能となり，材料類似性の判断基準として，合理的かつ現実的だと考えられる。また，結晶構造など大規模な物質データベースは幾つか存在するので，それらをハブとして物質レベルでリンクを付けるのが，材料データ共有の有効的な方法である。例えば，AtomWork-Adv に収録している物質は，約 15 万種類があり，これまでの実験で検証された結晶構造は，ほぼ網羅している。材料データベースを作成する時に，AtomWork-Adv を参照し，AtomWork-Adv の物質情報とリンクすれば，AtomWork-Adv にある結晶構造，特性，状態図，更に AtomWork-Adv とリンクしている他のデータベースにある同じ物質から構成する材料のデータを自分のデータと統合することが可能になる（図 1）。我々は，AtomWork-Adv の物質情報を抽出し，物質辞書を作成した。さらに，化学式や結晶構造を用いた物質 ID の自動生成のアルゴリズムとプログラムも開発した。このプログラムを利用して，同じ化学式と結晶構造であれば，必ず同じ ID が生成され，物質の同定とリンク付けが容易になる。現在，物質辞書と物質 ID 生成プログラムは，テスト運用として研究協力者に無料提供されている。

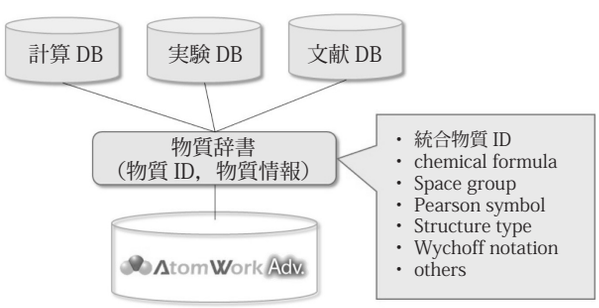

図 1　AtomWork-Adv をハブとした材料データネットワーク

複相材料や複合材料などの複雑材料システムに対して，図 2 のような階層構造を用いて物質レベルで材料を識別，類似性評価，またリンクすることもできる。

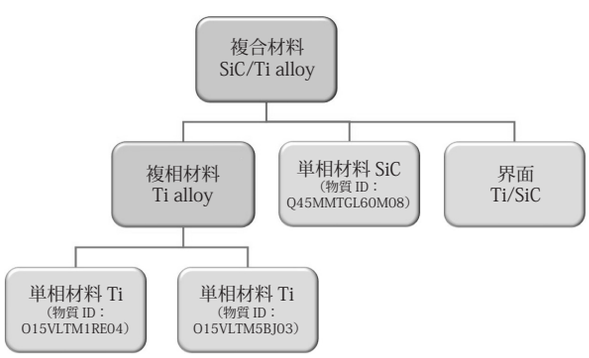

図 2　階層構造を用いた複合材料の記述例

材料の組織構造，作製条件や測定条件なども材料特性データに重要な記述子であるが，共通記述子の定義が困難であるため，各データベースの独自項目で定義するしかない。その場合は，できるだけ他のユーザが理解しやすく，また，データを再現できるように実験，計算などのデータ取得条件を記述することが重要である。

4　おわりに

本稿は，材料情報記述の視点から，外部データベースと統合しやすい，利用性の高い材料データベースの構築の仕方について説明した。近年，材料データの収集とデータベースの構築には，多大な人力と予算が投入されているが，それらのデータを長期に渡り，幅広く利用できるようにシステム化することが重要である。共通の材料識別システムのほかに，データの品質管理や内容の充実など，工夫すべきポイントはまだ沢山ある。今後，データ処理とデータベース技術の進歩により，データ収集と管理の方法も大きく変化していくと予想しているが，10年，20年後の材料科学のニーズにも適応できる利用性の高いデータベース設計と構築の基本思想は変わらないであろう。

参考文献

1）無機材料データベース AtomWork-Adv : https://atomwork-adv.nims.go.jp/

2）Pauling File : https://paulingfile.com/index.php?p=home