科学技術におけるデータベースの役割(11)

Role of Databases for Science and Technology (11)

馬場 哲也*、須田 幸子*、山下 雄一郎**,*
Tetsuya Baba, Yukiko Suda, Yuichiro Yamashita

１．データ科学とデータベース

今回は「データ科学」の視点から「データベース」の役割を論じたい。データ科学(data science)は統計学、機械学習などを核としてシミュレーションなどの計算機科学と連携し、個別の研究分野・技術分野にとらわれず、情報処理技術を駆使することにより、デジタル化された巨大なデータの集合を処理して知見を得るアプローチという認識が一般的であるように思われる。

一方、データ科学の優れた教科書のひとつである「データサイエンス講義」[1]においては

「学術的なデータサイエンティストは、社会科学から生物学までの何かに長けており、大量のデータを扱い、データの構造、サイズ、乱雑性、複雑性、性質によってもたらされる計算に立ち向かう必要があると同時に、現実世界の問題を解決する科学者であると言って良いかもしれません。」と述べている。同書では次いで、産業界(IT 業界)で使われる「データサイエンティスト」の意味を考察し、「さらに一般的にはデータサイエンティストはデータから意味を抽出し、解釈する方法を知っている人であり、それには統計や機械学習のツールや手法に加えて、【途中略】データの収集、クリーニング、マンジング（データの解析、解体、フォーマット）の作業にも多くの時間を費やします。」
と説明している。

このようにデータ科学に対する理解は多様であり、統計学や AI 技術などの情報処理技術のみならず、データベース開発に必要とされる知見や技術、データの収集から解析と解釈までの全体の流れを見通す視点が重要であると考えられる。

２．データ科学の物質・材料研究への適用

データ科学という用語は個別の専門分野に限定されない普遍的な認識法やデータ処理技術のニュアンスが正面にでるが、自然科学の専門分野への適用に際しては、生物学においてはバイオ・インフォマティクス(Bio-informatics), 化学においては化合物インフォマティクス(Chemo-informatics)など、専門的知見とデータ科学との融合により新しい研究分野が創出されてきた。バイオ・インフォマティクスにおいては巨大なゲノムデータベースが、化合物インフォマティクスにおいては化合物の命名法や表記法の標準化の取り組みや化合物データベースが確固とした基盤として研究を支えている。

材料インフォマティクス(Materials Informatics)は、図１に示されるように、データ科学を物質・材料研究に適用することによって生まれた研究分野である [2]。

材料インフォマティクスにおいては数 10 万物質を収録した結晶構造（実験的には X 線回折のデータにより決定）のデータベースが利用可能であるが [3, 4]、熱的特性などの物性値に関する実験データの集積とデジ

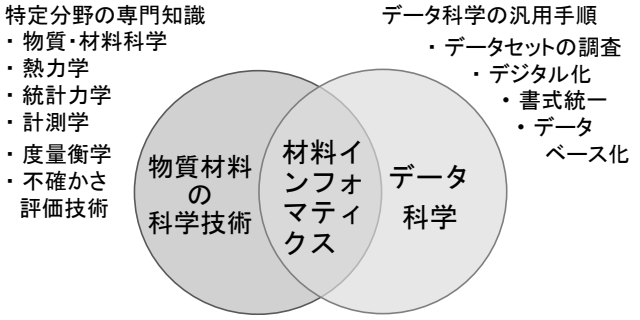


図１　データ科学と材料インフォマティクス [2]

* 未利用熱エネルギー革新的活用技術研究組合, 〒305-8568 茨城県つくば市梅園 1-1-1 中央第２, Thermal Management Materials and Technology Research Association (TherMAT), AIST Tsukuba Central 2, 1-1-1, Umezono, Tsukuba, Ibaraki 305-8568, JAPAN, FAX: 029-861-4236, E-mail:tbaba@thermat.jp
** 国立研究開発法人 産業技術総合研究所 物質計測標準研究部門 熱物性標準グループ, 〒305-8563 茨城県つくば市梅園 1-1-1 中央第３, Metrology Institute of Japan, National Institute of Advanced Industrial Science and Technology, AIST Tsukuba Central 3, 1-1-1, Umezono, Tsukuba, Ibaraki 305-8563, JAPAN

タル化は遅れている。産業技術総合研究所は熱物性データベースの開発を進め、物質・材料名や化学式などによる検索結果を表示するユーザインターフェイス、ＡＰＩによるアクセスを実現しているが、化合物データベースと比較するとデータの収録数は少ない[5]。

上記のように結晶構造以外の実測データに関して、データ科学の真価を発揮できる十分大きいデジタル化されたデータセットの利用が容易でない状況を反映して、第一原理計算により蓄積されたデータを主体にしたデータセットの活用が注目されている。(米国：Materials Project [6]，欧州：NOMAD, Novel Materials Discovery Laboratory [7]，日本：MI²I, Materials research by Information Integration Initiative, 情報統合型物質・材料開発イニシアティブ[2]など)

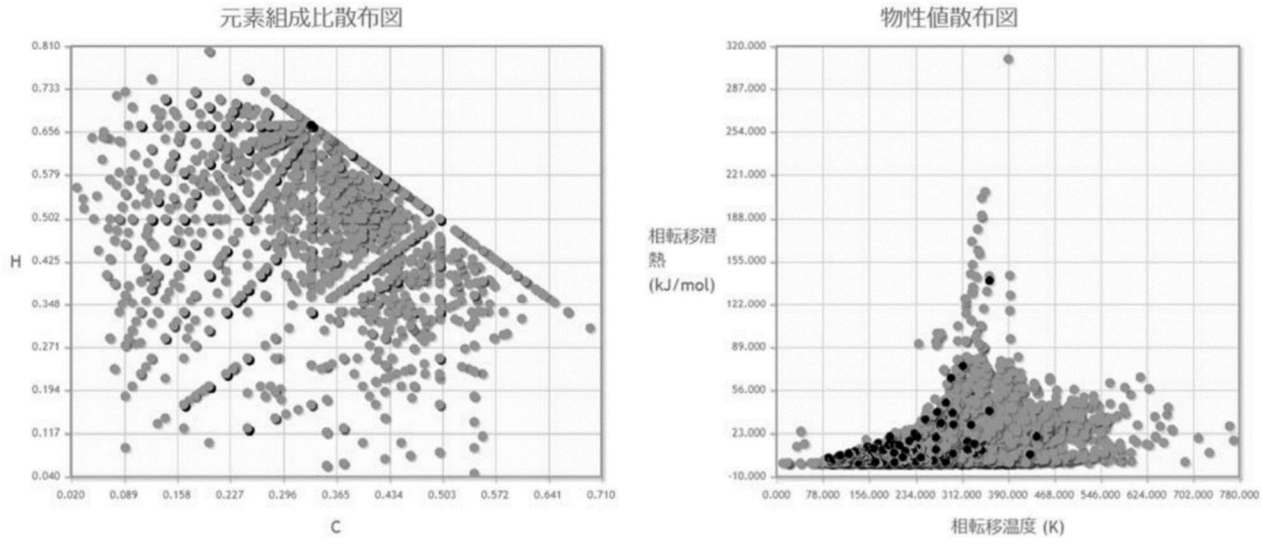
３．蓄熱材料のデータベース

国立研究開発法人新エネルギー・産業技術総合開発機構と未利用熱エネルギー革新的活用技術研究組合では、製造業や輸送機器などで排熱として失われている未利用の熱エネルギーを活用する技術の開発をめざして、「未利用熱エネルギー革新的活用技術研究プロジェクト」を 2013 年度から 2022 年度までの 10 年計画で実施している[8, 9]。プロジェクト全体を支える基盤技術として、未利用熱利用の高度化に不可欠な蓄熱材料、熱電材料などの熱的特性、化学組成を収録した「熱関連材料データベース」の開発が進められている[10]。

3-1 潜熱蓄熱に関するデータの蓄積

最も基本的な潜熱蓄熱材料は水（液相：水、固相：氷）であり、多量の結晶水を含むクラスレート・ハイドレートは水の融解熱に由来する大きな潜熱を有し、室温以下・室温付近の作動温度での有力な潜熱蓄熱材料として研究されている[11]。未利用熱エネルギーは主に排熱を想定しており、より高い融点を有する物質のなかで相転移熱が大きいものが求められる。

化合物は炭素原子が鎖状に結合し水素、酸素、窒素などが結合した分子により構成される有機化合物と、固



材料一覧

材料ID	材料名	化学式	分子量	生成	ギブ	エン	比熱	相転移温度 (K)	相転移潜熱 (kJ/mol)	C元素比 (%)	H元素比 (%)
2912	Cyclooctadecane	C18H36	252.480					346.000	9.870	0.333	0.667
2913	Dodecylcyclohexane	C18H36	252.480					285.800	45.840	0.333	0.667
3083	1,1,10,10-Tetramethylcyclooctadecane	C22H44	308.590					359.000	39.580	0.333	0.667
3145	Cyclotetracosane	C24H48	336.650					297.000	38.000	0.333	0.667
3145	Cyclotetracosane	C24H48	336.650					322.000	10.800	0.333	0.667
3146	Octadecylcyclohexane	C24H48	336.650					314.700	74.000	0.333	0.667
3146	Octadecylcyclohexane	C24H48	336.650					314.700	74.000	0.333	0.667
3146	Octadecylcyclohexane	C24H48	336.650					316.400		0.333	0.667
3146	Octadecylcyclohexane	C24H48	336.650					314.100		0.333	0.667
3197	1,1,4,4,10,10,13,13-Octamethylcycloocadecane	C26H52	364.700					427.000	6.740	0.333	0.667

図２　有機化合物の元素組成比と融解熱分布（モルあたり）、(固相/固相転移のデータを含む)
ＣとＨの元素組成比が 1:2 の場合（水素原子が酸素原子の２倍含まれ他の原子はない分子）を左上図で選択（黒丸）したときの融点と融解熱の分布が右上図に黒丸で示されている。
それに該当する化合物の一覧が下部に示されている。

体全体が共有結合、イオン結合、金属結合などで結合している無機物の結晶やガラスに大別される。両者は、それぞれ有機化学および無機化学という大きな学問分野として発展をとげてきた。

1) 無機化合物

上記の無機化合物の融点と融解熱（固相/固相転移も相図や熱力学の対象となるが潜熱蓄熱においては当面は考慮しないこととする。）は無機化学において相図と熱力学として体系的に研究されてきた。異なる物質を混合し相転移熱を制御することが潜熱材料の設計の要点となる。多様な材料に対して組成比、結晶構造、相図、熱測定の結果を材料熱力学の観点から総合的に理解する取り組みは“CalPhad 法”[12]などの「計算熱力学」として発展し，熱力学データは JANAF[13],SGTE[14]などに Gibbs energy として集積されてきた。

2) 有機化合物

有機化合物の材料インフォマティクスは化合物インフォマティクスを基盤として取り組むことができる。本プロジェクトにおいては主要な有機化合物 (2400 物質以上) の融点と融解熱の実測値を Landolt Börnstein [15] などのデータ集と論文から収録し、デジタル化して熱関連材料データベースに収録した。なお本データには固相/固相転移の相転移温度と相転移エンタルピー変化のデータも含まれているが、有機化合物においては融解（＝固相/液相転移）の相変化エンタルピー変化（融解熱）よりかなり小さい場合が一般的なので、蓄熱技術の観点からは、「相転移」という厳密な用語まで遡らず、融点・融解熱という用語を用いた。

熱関連材料データベースはデータの２次元的分布をインタラクティブに表示する機能を有しており、図 2 の左図は有機化合物中の炭素 C と水素 H の元素数組成比により収録された有機化合物の分布を表している。この図では X 軸が C の組成比、Y 軸が H の組成比を表しており、両者を加算して１になる場合が炭素と水素のみからなる化合物を表しており、左上から右下への直線上のプロットが対応している。他の構成元素を含めた組成比の和が１であるので直線の右上には化合物は存在しない。その中の黒丸は C と H 以外の原子が存在せず、H が C の２倍ある（CnH2n）エチレン、シクロヘキサンなどの炭化水素を表している。黒丸の点から左下に原点向かう直線は H が C の２倍あり他の元素も含む有機化合物を表している。

図２の右図は横軸を融点、縦軸を融解熱（単位質量あたり）として同じデータセットを表示している。融点が高いほど大きい融解熱が実現される。左図と右図は連動しており、左図で選択した黒丸の元素組成比の有機化合物の融点と融解熱の分布は右図の黒丸で示されている。図２の下部に炭素と水素の組成比で選択した化合物（図２、左図・右図の黒丸）に対応する化合物が表示される。

3-2 化学蓄熱に関するデータの蓄積

化学蓄熱の蓄熱量と対応する反応熱は Hess の法則により、反応経路に依存せず反応物と生成物の熱力学関数であるエンタルピーの差によって決定される。熱関連材料データベースには NIST-JANAF

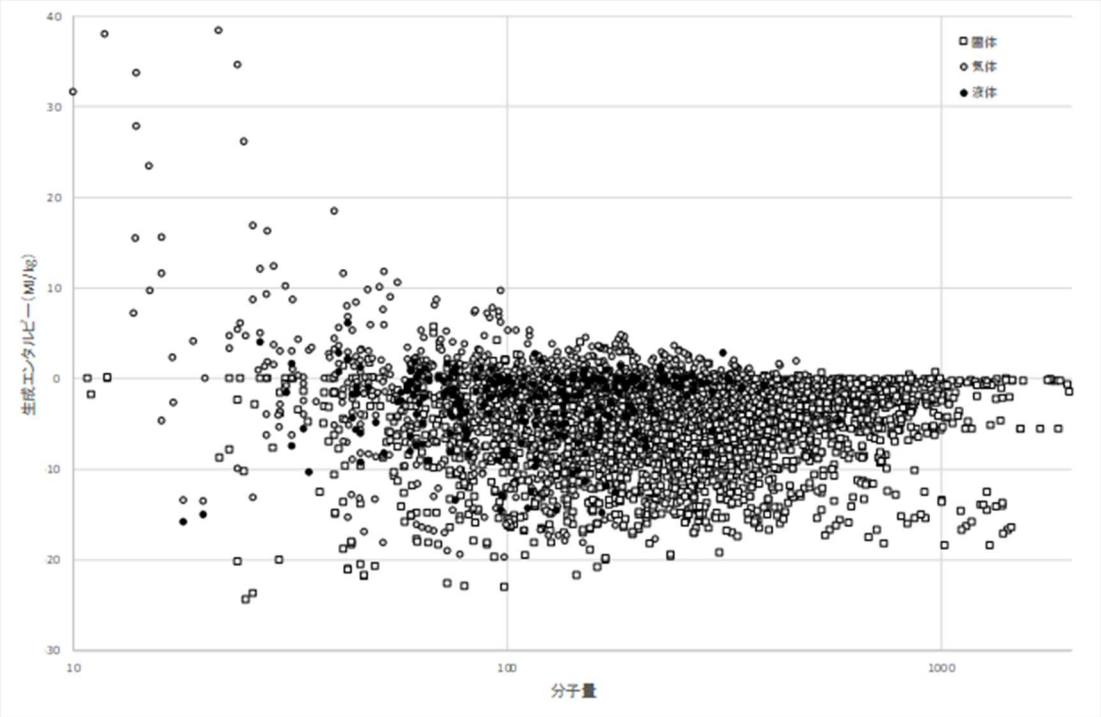


図３ 主要な無機化合物と炭素数１と２の有機化合物（約 12800 物質）の kg 当たりの標準状態における生成エンタルピー 気体はグレーの丸、液体は黒丸、固体は中抜き四角、のプロットで表示されている。

Thermochemical Tables[13]などから、主要な無機化合物および炭素数が２までの有機化合物（約 12800 物質）の標準状態（本稿では standard ambient temperature and pressure, SATP (= 298.15 K, 100.000 kPa)を採用する)における生成エンタルピー、生成エントロピー、ギブス自由エネルギーなどの熱力学関数の値が収録されている。このデータセットを参照することにより、「反応物の生成エンタルピーの和」と「生成物の生成エンタルピーの和」との差として反応熱が求まる [16]。

図３ に主要な無機化合物と炭素数１と２の有機化合物（計 12000 物質以上）の標準状態における生成エンタルピー（kg 当たり）を示す。プロットは気体（グレーの丸)、固体（中抜き四角)、液体（黒丸）の順で画面の奥から手前に描画されている。分子量は対数表示されている。記号が重なる場合があるので、液体の黒丸が集積している領域に存在する気体（グレーの丸）・固体（中抜き四角）が視認しにくい場合がある。化合物は標準状態においては分子量が大きくなるほど気相・液相より固相が安定に存在する傾向があるので、分子量の増加に伴い気体・液体・固体の順に分布がマイナス方向に広がる。

４．化合物インフォマティクスと構造活性相関

無機化合物の種類は構成する元素が１元系、２元系、３元系･･･と増加するに従い指数関数的に増大する。有機化合物は骨格が大きくなるにつれて膨大な種類が存在する。今日知られている化合物は一億種類のオーダーであるが、存在可能な化合物のごく一部にすぎないといわれている。化学の歴史においては化合物の組成・構造・名称をあきらかにして同定していくことが出発点でし、化合物の命名法、同定番号（ID）の整備は化学の進歩と相まって整備されてきた。

有機化学においては IUPAC による命名法が Bluebook に記述され､CAS 番号や日化辞番号などの ID, 構造を記述する Mol ファイル, SKF ファイル, SMILES などが整備され、教科書に詳述されている。このように分子の同定・記述が統一的に記述され、その情報から構造・性質が計算・推算できる状況の実現により、化合物インフォマティクス(Chemo-informatics)の発展が促進された。計量標準に関連する分野では新規化合物に対する安全性評価に構造活性相関（QSAR, Quantitative Structure Activity Relationship）が活用されている [17]。

５．TherMATにおける今後の取り組み

未利用熱エネルギー革新的活用技術研究プロジェクトでは蓄熱機能の視点から物質・材料の相転移温度、相転移エンタルピー、生成エンタルピーのデータベース化を進めたが、収録された化合物の大半は一般名のみが対応づけられている段階であり、SDF ファイルなどを特定し、化合物インフォマティクスを適用すべく取り組みを進めている。その状況を達成し、分子構造と蓄熱特性の相関を機械学習などにより解明し、優れた蓄熱性能を有する物質・材料の予測や設計手法の開発に取り組む計画である。

本成果は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) の委託業務により得られた。

参考文献

[1] Racel Schgutt,et. al. (瀬戸山雅人、他８名訳), “データサイエンス講義”, (オライリー・ジャパン, 2014).

[2] 情報統合型物質・材料開発イニシアティブ (MI2I : “Materials research by Information Integration” Initiative), http://www.nims.go.jp/MII-I/index.html.

[3] http://www2.fiz-karlsruhe.de/icsd_home.html

[4] https://www.nims.go.jp/news/press/2018/05/201805150.html

[5] https://tpds.db.aist.go.jp/

[6] Anubhav Jain, et. al., “Commentary: The Materials Project: A materials genome approach to accelerating materials innovation”, APL Materials 1, 011002 (2013); https://doi.org/10.1063/1.4812323

[7] The NOMAD (Novel Materials Discovery Laboratory), A European Centre of Excellence, https://nomad-coe.eu/

[8] http://www.nedo.go.jp/activities/ZZJP_100097.html

[9] http://www.thermat.jp/

[10] http://www.thermat.jp/project/base/index7.html

[11] 中島、平田、”水和物による高効率蓄熱技術の開発”, IHI 技報 Vol.49 No.4 (2009), pp. 210 – 218.

[12] CALPHAD (Computer Coupling of Phase Diagrams and Thermochemistry) , http://www.calphad.org/

[13] NIST-JANAF Thermochemical Tables, NIST Standard Reference Database 13, Last Update to Data Content: 1998, DOI: 10.18434/T42S31

[14] SGTE, Scientific Group Thermodata Europe, http://www.crct.polymtl.ca/sgte/

[15] Landolt Börnstein, Subvolume A, Enthalpies of Fusion and Transition of Organic Compounds.

[16] 馬場哲也, “科学技術におけるデータベースの役割 (3)”, 熱物性, 2015, pp.97 – 99.

[17] https://www.nite.go.jp/chem/qsar/qsartop.html

[Received January 27, 2019]