Vol.2016-DBS-163 No.11
Vol.2016-IFAT-123 No.11
2016/9/14

マテリアルズ・インフォマティクスのための大規模多次元データベースシステムの提案

淺原 彰規[†1] 森田 秀和[†1] 林 秀樹[†1] 海野 英一郎[†2] 小野 寛太[†3]

概要：材料科学分野では最近「マテリアルズ・インフォマティクス」という IT を用い研究開発を効率化する取り組みが活発化してきている．計算機の黎明期より数値シミュレーションが行われていることが示すように本分野は元来 IT リテラシーの高い人材が多い分野でもあり，この取り組みによる研究開発の加速が期待される．ところが，現在，材料科学分野の研究者には，RDBMS をはじめとする最近のデータ工学の成果はほとんど普及していない状況にある．よって，これまで導入を阻んできた技術的および環境的要因を明確化し解決しなければ，その状況を打破することはできないと考えられる．そこで今回，特に技術的要因に着目し，実際の材料科学分野の研究現場において，RDBMS などのデータ管理システムがどう活用できるかを調査した．本報告では，そのひとつである大規模多次元データベースのデータ構造，およびそれに対する典型的なクエリをまとめ，そこに内在する技術的困難とその解決策について，既存研究との関連をふまえて報告する．また，その一例について，実際の材料科学研究のデータを管理する DB を構築し，そのフィージビリティを評価した結果について示す．

キーワード：リレーショナルデータベースシステム，マテリアルズ インフォマティクス，ビッグデータ

Large-scale-multi-dimensional-database management systems intended for “Materials Informatics”

AKINORI ASAHARA[†1] HIDEKAZU MORITA[†1] HIDEKI HAYASHI[†1]
EIICHIRO UMINO[†2] KANTA ONO[†3]

Abstract: Materials Informatics (MI) is an information technology intended for accelerating material science researches, worked on eagerly in recent years. Many material science researchers are skillful to information technology, as shown by the fact that numerical simulations were developed in the early days of information technologies. So it is being expected that MI will drastically contribute to researches of material science. However, results of data engineering such as database systems are seldom used for material science in contradiction to the situation. Technical and social factors causing the situation, accordingly, should be clarified and addressed on to make a break through. We summarized such factors in this report, especially focusing on the technical factors, and we also suggested several use cases of data management systems such as RDBMS. A large-scale multi-dimensional-array database system, shown as a representative use case, is detailed in this report. The data structure and typical queries of it are summarized to clarify the requirements for the system Latent issues relevant to database systems for MI are related to the existing researches as the conclusion. An actual dataset, moreover, was imported into a database system. The performance of the system was evaluated to confirm the feasibility and the results are also shown in this report.

Keywords: Relational Database System, Materials Informatics, Bigdata

1. はじめに

材料科学分野では，最近「マテリアルズ・インフォマティクス」(以降，MI)という IT を用い研究開発を効率化する取り組みが活発化してきた[1]．MI は，従来，実験の試行錯誤によって発見されてきた材料科学分野の知見を，ICT 技術を用いることでより短期間に導き出そうという試みである. MI において最も主要な役割を占めるものの一つは第一原理計算をはじめとする，基礎的な物理法則にもとづく数値シミュレーションにより物理現象を再現する技術である．元来，新材料の開発においては，例えばベースとなる金属元素に対して何をどのくらい添加すればよりよい電気伝導性や塑性などを得られるか，などを実験的に評価する試みが繰り返し行われているが，これを例えば第一原理計算で置き換えることができれば，実験を繰り返すよりも時間的にも費用的にも効率がよくなると考えられる．

第一原理計算は基礎的な物理法則以外に仮定をおかず計算を積み重ねることで結果を得るため，近似計算ができず長い計算時間がかかってしまう．いくらかの仮定をおいたとしてもその傾向は同様であり，そのため大型並列計算機を用いて計算が行われるのが一般的である．その状況にあって，近年の GPGPU 等の並列計算の発展により，計算可能な量が増えてきたことが, MI が注目されはじめた要因の一つともいえる．しかし，MI に注目が集まり，多くの材料科学研究者が MI のために準備された豊富な計算リソースを用いて数値シミュレーションを行うようになると，次々とデータが生産され始める．特に大型計算機の計算結果と

†1 (株)日立製作所
Hitachi Ltd.
†2 (株)日立ソリューションズ
Hitachi Solutions Ltd.
†3 高エネルギー加速器研究機構
High Energy Accelerator Research Organization

Vol.2016-DBS-163 No.11
Vol.2016-IFAT-123 No.11
2016/9/14

もなると，そのデータはかなり大規模なものとなることが予想される．また，従来は仮説検証のために行っていた実験についても，数値計算結果の検証用途など，データの収集そのものが目的になることも想定される．さすれば網羅的にデータを収集しておくなど，データの再利用性を高めることが期待される．したがって，RDBMS(Relational Database Management System)やKVS(Key Value Store[2])などの，いわゆるビッグデータ向けのデータ管理システムが求められるようになってくると考えられる．

ところが現在，材料科学分野の研究者の多くは，第一原理計算のプログラムや可視化ツールなどを Fortran などのレガシーな環境を用いて開発しており，RDBMS などのデータ管理システムはほとんど活用されていない．数値シミュレーションが行われていることが示すように本分野はITリテラシーの高い人材も少なくないにもかかわらず，データ管理システムが普及しないのは，いくつかの技術的あるいは社会的要因が関連しているものと考えられる．

本報告では，この材料科学分野，特に金属材料を対象にICT 技術の活用たる MI のために必要となるデータ管理システムについて，本分野でどのようなデータが用いられているかという観点から要件を整理し，そこにある技術的，社会的要因について検討する．また，その一つの要素として，RDBMS にて多次元のデータを管理する方策について提案し，そのフィージビリティを評価するとともに，データ管理技術上の課題について示す．

2. 材料科学研究におけるデータ管理

2.1 材料科学分野の研究開発

図 1 に材料科学の対象についてまとめたものを示す．材料科学の対象は多岐に渡っているが，大きくはプラスチック，ゴム，繊維等の有機材料と，ガラス，セラミック，金属等の無機材料に分類される．有機材料は炭素 C を中心に，水素 H や窒素 N が組み合わさってできる材料の総称である．一方の無機材料は有機物でない材料全般を指し，金属材料とセラミックやガラスなどの SiO_2 を主原料とする非金属材料に分類されることが多い．特に金属材料は電気，電子部品や自動車航空機船舶等の機体など機械類の材料として広く用いられており，材料科学分野でも大きな位置を占める．本報告では，特に金属材料に着目し，データ管理技術の適用について検討する．

金属材料の開発では，金属材料の組成（使用されている金属元素）や製造プロセス（加熱処理など）を決定して実際にサンプルを作成し，その材料特性を評価することが行われている．これらの手順はどのような目的の材料開発かによって異なっている．表 1 に材料開発における目的となりうる材料特性の例を示す．ここに示したとおり，材料特性としては力学的なもの，電気的なものや化学的なものなど種々存在しているが，知られている限りにおいては万能な金属というものは存在しない．ただし，用途ごとに，経験則としてどの種の金属が適切か，大まかに知られており，それを基本として試行錯誤により材料特性の改善を図っていくこととなる．この試行錯誤の過程では，時として一つに数百から数千時間を要するケースもあり（例えば，金属の経年劣化，疲労特性を評価するには相応の時間をかける必要がある），試行錯誤の回数を少しでも削減することが材料開発の効率化につながるとされている．

そこで，数値シミュレーション等を用いて，その回数を低減するのが MI の取り組みである．したがって，材料の組成，加工プロセスのモデルを立て，そのモデルにしたがって数値計算を行い，その結果として材料特性を得ることにより，実物の試料を用いた試行錯誤の代替とするというのが最も基本的な手続きとなる．ここで，この手順では実は試行錯誤の回数自体は減っておらず，一部を数値計算によって仮想的に実行した，という点に注意を要する．すなわち，試行錯誤の回数を低減する取り組みは別に必要である．そこで例えば，他の研究者が同様の計算を過去に行った事例がないかを検索できるようにするなど，データ管理システムを活用して試行錯誤を低減できる可能性がある．

2.2 データ管理システム導入の阻害要因

前節の議論の通り，データ管理システムを導入することで，数値シミュレーション自体の回数なども低減できる可能性がある．しかし，実際の材料科学の研究者らによると，データ管理システムを導入している材料科学の研究者はかなり少数であった．その要因は決して明確ではないが，以降に報告者の分析をまとめたものを示す．

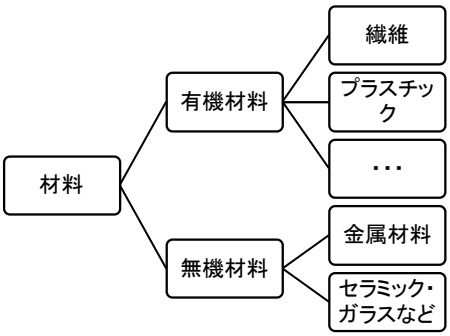

図 1　材料の分類
Figure 1　Category of materials

表 1　材料特性の例
Table 1　Example of material parameters.

物理的特性		機械的特性	
密度	誘電率	強度	疲労特性
透磁率	電気伝導率	硬度	加工性
熱伝導率	比熱	塑性	展性
膨張率	弾性係数		

2.2.1 社会的阻害要因

(1) 情報システムの活用可能性への懐疑

よくみられる意見の一つは，情報システムによる研究効率化の可否が半信半疑である，との意見である．特に多いのは，新材料の開発は未知の知見を求めるものであり，情報システムにそれを求めるのは困難ではないか，というものである．実際，続々新発見する情報システムは本稿執筆時点ではまだ困難である．一方，特に数値シミュレーション等を活用してきた研究者からは，これまでほぼ管理されていなかった材料科学研究のデータを活用していくことへの期待の声もきかれた．情報システムへの過大な喧伝が不信を招いている反面，実体験にもとづく有益さも理解されているようである．今後 MI が材料科学の発展に大きな貢献を果たすならば，このような齟齬は解消していくとも考えられる．

(2) IT スキルの偏在性

材料科学の研究者には，数値シミュレーションの開発者など，IT スキルの高い研究者も多く，潜在的なデータ管理システムのユーザと考えられる．ところが，実際には材料科学の研究者には，まったく情報システムに馴染みがないという研究者も多い．多くのケースでは，材料科学の研究室に１，２名程度特に IT スキルの高い研究者がおり，IT システムの管理を担当している．すると，IT に関連するタスクは彼らに集中しがちになる．RDBMS などの情報システムには，複数のユーザがコンピュータを使用することを前提としたトランザクションやアカウントの管理機能があるが，このようなケースではそれを使う必要がほとんどなく，ファイルシステム上のデータ管理で十分である．一般的には，複数人でデータを共有することが求められる場合には，表計算ソフトウェアのスプレッドシート等が使用される例が数多くあるが，それも使用されないことも多い．

一般的な組織では, 外部業者に委託して IT システムを構築することも想定されるが, IT スキルの高い研究者は独力でもかなりの部分対応できるため，外部のエンジニアに頼る必要性があまりないということも，かえってデータ管理システムの導入を阻む一つの要因となると考えられる．

新たな IT スキルを求めずとも使用可能なシステムを導入することがこの問題の解消の糸口になると考えられる．

2.2.2 技術的阻害要因

(1) データの非構造性

材料科学分野のデータは一般的に計測データに類するものであって，必ずしも表形式などの構造化された形式になっていない．そのため，RDBMS 等では扱いにくい面がある．ただし，近年では画像データ等の非構造データも対象にした分散処理プラットフォームなど，この課題に取り組んでいるデータ管理システムが多くなってきており，それらを活用していくことも一つの方法である．

(2) 分析の複雑さ

材料科学の基礎として，量子力学や統計力学などの理論が存在する．これらは複雑な数式にしたがっており，数値シミュレーションではそれらを単純な演算に帰着させて解いていることが多い（例えば有限要素法，差分法など）．逆にいえば，材料科学データを管理するデータ管理システムにもその種の処理が容易に実装可能な機構が求められる．というのも，データ管理システムで管理されているデータを処理のたびに逐一他システムへ移行していたのでは利便性を損なうためである．望むらくはデータ管理システム内で上記のような複雑な数式にもとづく処理でも実行できるようにすべきである．一見するとそれは困難にも見受けられるが，実際にはすでに数値シミュレーション等で単純な演算への帰着がなされており，データ管理システムは数値シミュレーションの結果をできるかぎりありのまま扱えばよいと考えられる．ただし，この方策は，数値シミュレーションにおける課題を同様に内包する．すなわち，処理効率（処理時間）に課題があると想定される．

2.3 材料科学研究におけるデータ

データ管理システムを導入する技術的阻害要因は，データとその処理の特異性に影響されることが示唆された．そこで，以降では材料科学のデータについて述べる．

2.3.1 数値計算

第一原理計算を始めとする数値シミュレーションの結果は，数理モデルにもとづく形式である．物理現象を記述する上で特に基礎的な要素は時空間座標と運動量の組である．物理法則の多くは座標空間の偏微分方程式として記述され，その解が運動量と対応づいている．そこで時空間を有限の微小な要素に分け，近似的に偏微分方程式を解く方法がよく用いられる（有限要素法）．例えば，ポテンシャルエネルギー$\mathrm{V}(\mathbf{x},\mathrm{t})$のもとで運動する粒子の Schrödinger 方程式

$$\left(-\frac{\hbar}{2\mathrm{m}}\nabla^2+\mathrm{V}(\mathbf{x},\mathrm{t})\right)\psi(\mathrm{t},\mathbf{x})=\mathrm{i}\hbar\frac{\partial}{\partial\mathrm{t}}\psi(\mathrm{t},\mathbf{x})$$

の解は, 時刻 t における波動関数$\psi(\mathrm{t},\mathbf{x})$の分布であり(なお, $\mathbf{x}=(\mathrm{x},\mathrm{y},\mathrm{z})$の 3 次元位置ベクトル), 6 次元の配列$(\mathrm{t},\mathrm{x},\mathrm{y},\mathrm{z},\mathrm{Re}\,\psi,\mathrm{Im}\,\psi)$で記述される(波動関数は複素関数である). 他にも例えば粒子の運動は座標の時間変化$\mathbf{x}(\mathrm{t})$で記述でき，(t,x,y,z)の 4 次元配列となる．このように有限要素法のシミュレーション結果は通常多次元の配列として記述できる．それゆえ NetCDF [3]等の配列データ向けのデータ形式が用いられることもある．

2.3.2 計測結果

材料の計測データにはいくつかの種類がありえるが，大まかには，試料の状態の計測と，試料の材料特性の評価，の 2 種類に分類できる．

試料の状態の計測というのは試料がどのような状態にあるかを画像等で判断できるための計測である．代表的なも

のは試料表面を電子顕微鏡で観測した画像や，X 線等を試料に照射して得られる散乱[4]のパターンの画像があげられる．他方の材料特性の評価は，試料に対する結果としての特性であり，電気伝導率などを直接的に計測したり，試料を加熱したり,酸に曝したり,圧力を加えたりしてから，その特性を計測することが行われている．

これに関連する情報としては，いわゆる実験ノート等の実験経過の記録がある．実験経過の記録は古来より実験において最も重要なものの一つとして扱われているが，多くは実験中に手書きで記述されており，後日，スプレッドシート等を用いて電子化されるケースが多い．

2.3.3 文献情報

論文や特許等の公開文献も材料科学で扱われるデータの一つとみなすことができる．このデータは他の研究者の論文に対する追試や対照実験として参照されるため，ある研究者が現在対象としている材料に関連する情報を推薦するなどのレコメンドシステムが必要になると考えられる．

3. 多次元データベースシステムの提案

3.1 多次元データベースシステム

前節までの議論によると，数値シミュレーションの結果や散乱実験，電子顕微鏡画像の実験データなど，多くの実験データが多次元の配列として表現される．そしてまた，データの分析が可能であるためには，数値シミュレーションのデータモデルをできる限りそのまま扱うことが求められる．したがって，この多次元の数値データを管理するデータ管理システムが必要となる．ただし，多次元のデータをそのまま管理することは多大なオーバーヘッドを要求するため，それを効率的に行う仕組みが必要となる．

本報告では，このような多次元のデータを扱うデータ管理システムを多次元データベースと呼ぶ．多次元データベースは Multi Dimensional Array Database として知られている実装がいくつかある[5][6]が，ここでは一般的な RDBMS を用いて多次元データベースを実装する方針で検討した．

3.2 フィージビリティスタディ

今回は，磁性体の理論 Landau-Liftshitz-Gilbert equations (LLG)の数値シミュレーションを対象とし多次元データベースのフィージビリティスタディとして，RDBMS による管理を行うことを想定した検索性能評価を行なった．LLG シミュレーションは 3 次元空間の格子上に配された要素によって表現される，有限要素法の数値シミュレーションである．整数で指定される空間座標値(x,y,z)に物理量がそれぞれ付与されており，それが時間 t とともに変化していく．データベースシステムはこのデータ構造をなるべくそのまま引き継ぐべきである，という観点からスキーマ設計を行った．表 2 にその結果を示す．

表 2　今回用いた LLG データの構造
Table 2 LLG data structure used in the experiment.

(a) 空間データテーブル(SDP)

カラム名	型	意味
s_data_id	integer	主キー
exp_id	integer	実験の ID
grid_id	integer	格子点の ID
x	integer	x 座標値
y	integer	y 座標値
z	integer	z 座標値
grain	integer	粒子番号

(b) 時空間データテーブル(STDP)

カラム名	型	意味
st_data_id	integer	主キー
exp_id	integer	実験の ID
grid_id	integer	格子点の ID
t	integer	計算開始からのステップ数
m1_x	double precision	元素 1 の磁化 x 方向
m1_y	double precision	元素 1 の磁化 y 方向
m1_z	double precision	元素 1 の磁化 z 方向
m1_theta	double precision	元素 1 の磁化 θ 方向
m1_phi	double precision	元素 1 の磁化 ϕ 方向
m2_x	double precision	元素 2 の磁化 x 方向
m2_y	double precision	元素 2 の磁化 y 方向
m2_z	double precision	元素 2 の磁化 z 方向
m2_theta	double precision	元素 2 の磁化 θ 方向
m2_phi	double precision	元素 2 の磁化 ϕ 方向
a_x	double precision	一軸異方性ベクトルの x 方向
a_y	double precision	一軸異方性ベクトルの y 方向
a_z	double precision	一軸異方性ベクトルの z 方向
E_tot	double precision	エネルギー合計
E_zee	double precision	ゼーマンエネルギー
E_exc	double precision	励起エネルギー
E_ani	double precision	異方性エネルギー
E_dip	double precision	双極子エネルギー

SDP は空間的な分布のテーブルである．空間座標をそれぞれ 512 分割したセルで表現されており,3 次元ではその 3 乗，134,217,728 件のデータとなっている．各座標はいずれかの粒子に所属しており，その粒子を一意に識別する番号が grain_id に格納されている．一方，STDP は各空間座標のデータの時間変化を管理するテーブルであり，それぞれ時間変化する物理量を示している．今回は 4 ステップ分格納しているため，データの件数としては 4 倍の 536,870,912 件となる．なお，PostgreSQL での SDP のテーブルサイズは

Vol.2016-DBS-163 No.11
Vol.2016-IFAT-123 No.11
2016/9/14

12GB，STDP のテーブルサイズは 111GB であった．各レコードのデータサイズは決して大きくはないため，データサイズはさほどでもないが，レコード数は極端に大きいデータベースである．この SDP と STDP はシミュレーション上，特に区別されていないが，データの件数が多いことから，重複を排除し正規形のスキーマ構造とすることでデータサイズの増大を抑制している．SDP には空間座標での検索を高速化するため，(exp_id, grid_id)，(x, y, z, exp_id, grid_id)，(y, z, x, exp_id, grid_id)，(z, x, y, exp_id, grid_id) にマルチカラム B-tree インデックスを構築した．また，STDP には時刻に関する検索を高速化するため，(exp_id, grid_id, t) にマルチカラム B-tree インデックスを構築した．また，利便性を確保するため両者を grid_id および exp_id にて inner join したビュー v_data を構築して用いた．

表 3 にこのデータベースに対して想定される検索クエリをいくつか示す．本報告ではこのうち最も単純で基礎的なクエリの一つである(a)について性能を評価した．(a)はある x 軸に垂直なある平面に関する断面を取得するクエリである．このクエリは，たとえば断面画像を表示する際に用いることが想定される．断面を作る面として，同様のものを y 軸，z 軸についても生成できる．本報告では x 軸，y 軸，z 軸それぞれについての断面を生成するクエリ，つまり x=0 の条件のクエリ 1，y=0 の条件のクエリ 2，z=0 の条件のクエリ 3 の三つについて性能評価を行った．いずれのクエリも得られるレコード数は 262,144 件である．

表 3 検索クエリの例
Table 3 Query examples.

(a) 表面の M1_Z の取得

```
SELECT
  x, y, z, m1_z
FROM v_data
WHERE
  exp_id = 1    AND    t = 0    AND x = 1;
```

(b) ステップごとに m1 の磁化の z 成分が正方向になっている数を評価

```
SELECT
  t, count(*)
FROM v_data WHERE exp_id = 1 AND m1_z    > 0
GROUP BY t;
```

(c) 粒子ごとに m1 の磁化の z 成分向きが異なるものが混じっているかをステップごとに評価

```
SELECT
  grain_id, t,
  count(distinct ROUND(m1_z /abs(m1_z)))
FROM v_data WHERE exp_id = 1
GROUP BY grain_id, t
```

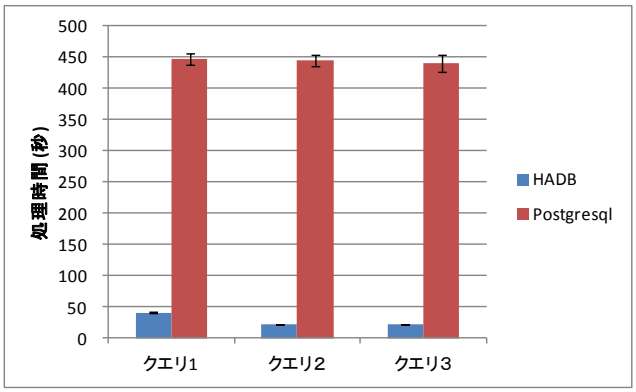

図 2 検索処理の時間
Figure 2 Query processing time

今回の評価では一般的な RDBMS である PostgreSQL 9.2.15 およびストレージへの並列アクセスの機能を持つ Hitachi Advanced Data Binder 03-01-/B [7] (以降 HADB[a])を用いた．計測環境としては，CPU が Xeon E7-8860v3 2.20GHz(16 コア)を 4 つ，メモリ 256 GB を搭載しており，OS は Red Hat Enterprise Linux 6.7 を用い，毎分 10,000 回転，容量 1.2TB の HDD を最大 32 台並列アクセスできるストレージをファイバチャネル接続し（理論データ転送性能は 12 Gbps），そこにデータベースを構築した．本データベースを構築するにあたって，それぞれのシステムに用意されている CSV データインポートの機能を用いて LLG シミュレーションのデータをインポートし，その後インデックスを構築した．なお，データのインポートに際し，各軸のデータに偏りが生じないようランダムに順序をシャッフルした CSV ファイルを生成してからインポートした．また，性能評価は毎回キャッシュ等を削除しながら 12 回繰り返し，その平均を用いた．

3.3 実験結果

図 2 に各クエリの処理に要した時間を示す．クエリ 1, 2, 3 はそれぞれの空間軸での検索であるが，いずれもほぼ均等な処理性能であった．PostgreSQL では各検索におよそ 440～450 秒，HADB では 20～40 秒程度で検索を終了した．この処理時間は，繰りかえしデータを分析することを考慮に入れても実用的な処理時間である．ただし，今回用いたデータは，一つのシミュレーションデータで，かつ時刻は 4 ステップしか扱っていない．実際には多数のシミュレーションデータで，それぞれ数百から数千ステップのデータを含むものを管理せねばならない．その場合，ごく一般的な RDBMS である PostgreSQL をそのまま用いた場合では，非常に処理時間が長くなってしまうことが予測される．し

a) Hitachi Advanced Data Binder プラットフォームは，内閣府の最先端研究開発支援プログラム「超巨大データベース時代に向けた最高速データベースエンジンの開発と当該エンジンを核とする戦略的社会サービスの実証・評価」（中心研究者：喜連川 東大教授／国立情報学研究所所長）の成果を利用しています．

たがって，HADB のような並列アクセスする機能を備えた RDBMS を用いることが必要となる．ただし，今回評価したクエリ以外のクエリ，例えば(b)(c)については索引が効きにくいことからさらに処理時間を要することが予測され，今後，多次元配列データに対する集計演算の高速化が求められるようになると考えられる．

3.4 技術課題と関連研究

本実験で示されたとおり，数値計算データを表現する多次元配列への条件付集計は現状の技術でも実行可能であるが，さらにデータが増加した場合やクエリが複雑化した場合の高速化が課題の一つと考えられる．数値シミュレーションの結果や実験データの多くは多次元配列の形式で記述されるため同様の課題を備えていると考えられ，配列志向のデータベースシステム[5][6]の並列処理により，集計クエリのスケーラビリティを向上していくことが求められる．また，今回の実験ではシミュレーションデータを対象としたが，実験データの場合は計測誤差等を含むので誤差の不確定性を加味した検索も必要になると考えられる．なお，IT スキルの偏在性に対応するため，これらの集計を容易にするためのユーザインターフェースも重要であり，標準 SQL[8]をサポートすることや，BI ツール等を用いて可視化，集計する機能[9]を提供することにより，より多くの研究者がデータ管理システムを扱えるようにすることが望ましい．

それ以外のデータとしては，文献データが挙げられる．文献データについては，レコメンドシステム[10][11][12]の技術が適用可能と想定されるが，材料科学分野の文献に限ると使用される元素や目的となる材料特性を的確に抽出することが必要となる．

4. おわりに

本報告では，材料科学分野におけるマテリアルズ・インフォマティクスという IT を用いた研究開発の効率化の取り組みが推進された際に，どのようにしてデータ管理システムを導入していくべきかという観点から，主に金属材料の研究開発を対象として検討を行った．今回は特に技術的な導入阻害要因に着目し，大規模な多次元データベースの必要性と，そこで管理されるべきデータの構造や規模，およびそれに対する典型的なクエリを整理した．また，その一例として，実際の材料科学研究のデータを管理する DB を構築し，そのフィージビリティを評価した．その結果，性能面の課題は大きいものの，一般的な RDBMS でも工夫次第で性能面での課題は解決できる余地があり，また SQL を用いた集計演算がデータの再利用や分析に非常に有益であることが示された．

今後，KVS や分散型のデータ管理システムや，インメモリデータベース等，大規模データの分析に適した基盤の検討を進めていくことで，より効果的なシステム提案ができるようになる．この種の知見は材料科学分野のみならず，自然科学の研究においては汎用的に用いられる可能性も高く，データ工学の知見が自然科学全般の発展に貢献するものと考えられる．

謝辞 この成果の一部は，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)の委託事業未来開拓研究プログラム「次世代自動車向け高効率モーター用磁性材料技術開発」の結果得られたものである．

参考文献

[1] Krishna Rajan, “Materials Informatics”, Materials Today, Vol. 8, Issue 10, pp.38-45, Oct. (2005)
[2] DeCandia, Giuseppe, et al. "Dynamo: amazon's highly available key-value store." ACM SIGOPS Operating Systems Review 41.6 (2007): 205-220.
[3] Open Geospatial Consortium, “OGC Network Common Data Form (NetCDF) Core Encoding Standard version 1.0 (10-090r3)”, http://www.opengeospatial.org/standards/netcdf
[4] 松岡秀樹, "意外に多い小角散乱実験からの情報 (1) 小角散乱の基礎 X 線・中性子の小角散乱から何がわかるか." 日本結晶学会誌 41.4 pp. 213-226 (1999)
[5] Michael Stonebraker, Jennie Duggan, Leilani Battle and Olga Papaemmanouil, “SciDB DBMS Research at M.I.T”, IEEE Data Eng. Bull., vol. 36- 4, pp. 21—30 (2013)
[6] rasdaman community open-source project, “rasdaman”, http://www.rasdaman.org/ (2016)
[7] 藤原 真二, 茂木 和彦, 田中 美智子, 田中 剛, 合田 和生, 喜連川 優, “TPC-H ベンチマークの 100TB クラスを用いた商用アウトオブオーダ型 データベースエンジンの評価と同クラスへの世界初登録”, DEIM Forum 2014 D8-5 (2014)
[8] ISO,”ISO/IEC CD 9075-15 Information technology -- Database languages -- SQL -- Part 15: Multi dimensional arrays” (2016)
[9] 武田浩一. "ビジネス・インテリジェンスと人工知能技術." 情報処理 47.7 (2006): 723-728.
[10] 鳥羽美奈子, 森靖英, & 田代大輔. (2012). 知識共有型レコメンドシステム “Knowledge Recommender” の提案とビル省エネ管理事業への適用. 情報処理学会論文誌, 53(1), 149-162.
[11] Chong Wang and David M. Blei. 2011. Collaborative topic modeling for recommending scientific articles. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '11). ACM, New York, NY, USA, 448-456.
[12] Takashi Yoneya, Hiroshi Mamitsuka , “PURE: A PUBMED ARTICLE RECOMMENDATION SYSTEM BASED ON CONTENT-BASED FILTERING”, Genome Informatics, Vol. 18 (2007) P 267-276