◆連　載◆

《隔月連載 全 5 回》

第 3 回　R & D 部門におけるデータ共有，AI 活用のためのデータの記録，蓄積，分析方法

上島 豊　　（株）キャトルアイ・サイエンス　代表取締役

《PROFILE》

略歴：

1992 年 3 月　大阪大学工学部　原子力工学科　卒業
1997 年 3 月　大阪大学大学院工学研究科　電磁エネルギー工学専攻　博士課程修了
1997 年 4 月　日本原子力研究所　博士研究員
2000 年 4 月　日本原子力研究所　研究職員
2006 年 3 月　日本原子力研究開発機構（旧日本原子力研究所）　退職
2006 年 4 月　キャトルアイ・サイエンス設立　代表取締役　就任

主な参加国家プロジェクト：

文部科学省 e-Japan プロジェクト「ITBL プロジェクト」,「バイオグリッドプロジェクト」
総務省 JGN プロジェクト「JGN を使った遠隔分散環境構築」
文部科学省リーディングプロジェクト「生体細胞機能シミュレーション」

主な受賞歴：

1999 年　6 月　日本原子力研究所　有功賞　「高並列計算機を用いたギガ粒子シミュレーションコードの開発」
2003 年　4 月　第 7 回サイエンス展示・実験ショーアイデアコンテスト文部科学大臣賞「光速の世界へご招待」
2004 年 12 月　第 1 回理研ベンチマークコンテスト　無差別部門　優勝

主な著作：

培風館「PSE book ―シミュレーション科学における問題解決のための環境（基礎編）」ISBN：456301558X
培風館「PSE book ―シミュレーション科学における問題解決のための環境（応用編）」ISBN：4563015598
培風館『ペタフロップス コンピューティング』ISBN978-4-563-01571-8
臨川書店『視覚とマンガ表現』ISBN978-4-653-04012-5

8　R & D 部門におけるデータ蓄積基盤としてデータベースがなぜ適切か？

前章では，「R & D 部門におけるデータ蓄積を Excel で行えるか？」と題し，R & D 部門におけるデータ蓄積が他の部門のデータ蓄積と大きく異なる点，Excel でデータ蓄積していく場合の課題に関して，説明を行った。本章では，「R & D 部門におけるデータ蓄積基盤としてデータベースがなぜ適切か？」と題し，R & D 部門のデータ蓄積基盤として，データベースを採用すべき理由を具体的な事例とともに解説する。

データ共有・利活用をするためには，実験，分析の再現ができるレベルの記録すべき項目を明確化し，それら項目を使って，どのようにデータを探査し，データ処理をするのかを第三者が実施できるレベルの手順書にまとめておく必要がある。実は，これだけでもデータ共有・利活用は可能である。そういう意味でいうと，データベース化は本質でないと言えばその通りである。実際，前章で説明したように Excel を使って，データ蓄積，共有することも可能である。ただ，Excel では，複数の人が同時に値を変更できないばかりか，変更や確認する必要のない実験（行）や項目（列）が表示されており，誤入力，誤編集リスクが大きく，データ品質が維持できない問題を抱えている。

データベースは，Excel のこのような問題を解決し，複数の人が同時編集及び誤入力，誤編集リスク抑止を実現する。実はデータベースの利点はこれだけではない。複数段の工程による実験のデータ蓄積，共有は，Excel では課題がある。例えば，合成工程でエチレン，プロピレンなどを混ぜてポリマーを合成し，配合工程でそれらポリマーの短鎖，長鎖の 2 種配合し，配合材料を作成するような実験があったとする。そして，それは確かに次表のように合成工程の Excel と配合工程の Excel で記録できる。

合成 ID	エチレン濃度	プロピレン濃度	ブタン濃度	引張強度
Poly1	10	90	0	100
Poly2	0	30	70	150
Poly3	20	30	50	175

配合 ID	短鎖ポリマー	短鎖ポリマー濃度	長鎖ポリマー	長鎖ポリマー濃度	粘度
Sample1	Poly1	10	Poly2	90	102
Sample2	Poly1	30	Poly2	70	203
Sample3	Poly1	20	Poly3	80	304
Sample4	Poly2	20	Poly3	80	403
Sample5	Poly2	60	Poly3	40	505

この時，ある配合材料の短鎖ポリマーのエチレン濃度を知りたいときにどうするのだろうか？配合工程の Excel において，当該配合材料で使われている短鎖ポリマーがどの合成工程の実験で作られたものか，つまり，合成 ID を確認し，その後，合成工程の Excel において，その合成 ID の合成樹脂でどの程度濃度のエチレンを合成に使ったのかを確認すれば，情報は得られる。しかし，知りたい「ある配合材料」というのが 100 個あれば，上記 2 つの Excel を行き来する作業を 100 回行わなければならないわけである。そして，当然そこで得た情報は覚えていられないので，別 Excel に短鎖ポリマーのエチレン濃度などの情報を一時記録しておくに違いない。そして，100 回も調べるときっと何回かは間違って一時記録してしまうものもあるはずだ。短鎖ポリマーのエチレン濃度は一時記録しておくといったが，後で他の量との比較やグラフを描いた分析をすることを考えると配合工程の Excel を複写したものに短鎖ポリマーのエチレン濃度の項目を追加するのが，便利なはずである。

実は，同じように短鎖ポリマーのプロピレン濃度やブタン濃度，調査ポリマーのエチレン濃度やプロピレン濃度やブタン濃度も同じようにして，配合工程の Excel を複写したものに項目追加しておけば，その Excel 一つで他の量との比較やグラフを描いたりし，分析をすることは可能である。それなら最初から工程を分けずに 1 つの Excel で記録すればいいのではないか？と思うかもしれない。

しかし，1 つの Excel で記録しようとすると合成工程で大量に作ったポリマーをたくさんの配合工程で少しずつ配合比率を変えて行う実験を行う場合，固定であるはずの当該ポリマーのエチレン濃度，プロピレン濃度，ブタン濃度を配合実験毎に同じ値を記入していく必要がある。また，配合は，短鎖ポリマーと長鎖ポリマーを配合するが，エチレン濃度，プロピレン濃度，ブタン濃度と書いてしまうとどちらのポリマーの組成のことを示しているのかがわからなくなるため，短鎖ポリマー内エチレン濃度，短鎖ポリマー内プロピレン濃度，短鎖ポリマー内ブタン濃度，長鎖ポリマー内エチレン濃度，長鎖ポリマー内プロピレン濃度，長鎖ポリマー内ブタン濃度のような項目にして，記録するようにしなければならない。実際，エチレン濃度，プロピレン濃度，ブタン濃度という項目名だけでは不十分で，何のポリマーの組成の濃度かの情報がないとデータ分析などはできないのである。

実際，「実験の独立変数（実験パラメータ）は何ですか？」と聞かれた場合，合成工程と配合工程の実験パラメータを列挙して，エチレン濃度，プロピレン濃度，ブタン濃度，短鎖ポリマー濃度，長鎖ポリマー濃度の 5 変数と答えてしまいそうだが，それは間違いである。「実験の独立変数（実験パラメータ）は何ですか？」の正確な答えは，短鎖ポリマー濃度，短鎖ポリマー内エチレン濃度，短鎖ポリマー内プロピレン濃度，短鎖ポリマー内ブタン濃度，長鎖ポリマー濃度，長鎖ポリマー内エチレン濃度，長鎖ポリマー内プロピレン濃度，長鎖ポリマー内ブタン濃度の 8 変数である。つまり，記録の煩雑さや誤記を減らそうとすると合成工程と配合工程を別々の Excel で記録していく方がいいが，データ絞り込みやデータ分析を行うために独立変数というものを意識しなければならない場合は，配合工程に合成工程を繰り込んだ 1 つの Excel で記録する方が良いのである。

配合 ID	短鎖ポリマー濃度	短鎖ポリマー内 エチレン濃度	短鎖ポリマー内 プロピレン濃度	短鎖ポリマー内 ブタン濃度
Sample1	10	10	90	0
Sample2	30	10	90	0
Sample3	20	10	90	0
Sample4	20	0	30	70
Sample5	60	0	30	70

長鎖ポリマー濃度	長鎖ポリマー内 エチレン濃度	長鎖ポリマー内 プロピレン濃度	長鎖ポリマー内 ブタン濃度	粘度
90	0	30	70	102
70	0	30	70	203
80	20	30	50	304
80	20	30	50	403
40	20	30	50	505

工程ごとの項目を分けて記録をすると見かけ上，独立変数が減ってしまうので，実際データ分析をするときは，この 8 項目に展開しないと X-Y Plot さえかけないのである。当たり前だが，X-Y Plot を書くということは，データが単純な 2 次元表形式になっている必要があり，2 つの表のようなものは，そのままでは X-Y Plot さえ，書けないのである。複数工程実験を複数の表で記録している人は，X-Y Plot を書くたびに複数の表を一つの表にまとめる作業をしているはずである。ただし，全ての項目を一つの表にするのは大変なので，X-Y Plot で使う，X 軸，Y 軸，凡例だけを取り出して，一つの表にまとめる作業をしていると思う。この作業が面倒な人は，良く使う項目だけをマクロを使って，自動的に一つの表にまとめる機構を作っている人もいると思う。しかし，このようにすることで，それ以外のデータ絞り込み，データ分析はより億劫になってしまい，分析範囲が無意識に狭まってしまうので，実は，研究としては望ましい状況ではない。

もちろん，最初から短鎖ポリマー濃度，短鎖ポリマー内エチレン濃度，短鎖ポリマー内プロピレン濃度，短鎖ポリマー内ブタン濃度，長鎖ポリマー濃度，長鎖ポリマー内エチレン濃度，長鎖ポリマー内プロピレン濃度，長鎖ポリマー内ブタン濃度という項目名の 1 工程実験として，データを記録すれば問題ないのだが，通常，そういう記録は煩雑で好まれないはずである。また，2 工程程度ならまだいいのだが，5,6 工程にもなれば，項目名が幾何級数的に増えることになってしまい，恐らく正しく記録していくこと自体が困難になる。また，多工程実験のデータでは，項目名の揺れを抑制するだけでなく，工程間の紐づけ（短鎖ポリマー濃度のポリマーは合成工程のどのポリマーに対応するのか？）揺れ抑制も重要となってくる。つまり，下表のように工程間の紐づけに揺れがある記録では，データ絞り込みや分析が正確にできないので，工程間の紐づけ揺れの抑止機構も必須である。

合成 ID	エチレン濃度	プロピレン濃度	ブタン濃度	引張強度
Poly1	10	90	0	100
Poly2	0	30	70	150
Poly3	20	30	50	175

配合 ID	短鎖ポリマー	短鎖ポリマー濃度	長鎖ポリマー	長鎖ポリマー濃度	粘度
Sample1	Poly1	10	Poly2	90	102
Sample2	poly1	30	poly2	70	203
Sample3	ポリマー 1	20	ポリマー 3	80	304
Sample4	Poly2	20	Poly3	80	403
Sample5	Poly2	60	Poly3	40	505

つまり，データを共有し，利活用するためには，記録は工程ごとの Excel で，データ絞り込みや分析はそれらを 1 工程に繰り込んだ Excel で行える必要があるのである。そして，複数工程を 1 工程に繰り込むときには，工程間の紐づけ情報を分析し，項目名を内部で自動生成する機構が必要であり，そういう機構がないと複数段工程のデータ検索，分析はできないのである。データベースはそういうことを内部で行えるようになっており，もうひとつのデータベースの大きな利点である。つまり，データ共有，利活用を考えると「複数の人が同時編集及び誤入力，誤編集リスク抑止」，「多工程データの紐づけと 1 工程への繰り込み」が最初から考慮されているデータベースを基盤に据えることが適切だということである。ちなみに，Excel でそういうことをできる仕組みを作り込むことも不可能ではないが，これら機構を Excel に作り込むというのは，ある意味，Excel をベースとしてデータベースを開発するのと同じことになり，車輪の再発明になるため費用対効果やソフトウェア品質の観点から止めておいた方が良い。

先ほど話をした，多工程データを 2 次元表に射影するには，どのような機構が必要なのかを少し説明しておく。例えば，モノマー調整工程，樹脂合成工程，配合工程，測定工程の 4 工程からなる実験を考える。例えば，測定工程の配合材料において，内部で使われているモノマーの分子量を表記したり，条件指定したい場合，同じモノマーが異なる樹脂に使われていたり，同じ樹脂が異なる配合材に使われていたりすると，計測に使った配合材内のどういう樹脂に使われたモノマーの分子量ということを指定できる方法（指定できる項目名）が必要になる。したがって，『計測で使われた計測使用配合材 ID ＝“F01”の中で使われた配合内樹脂 ID ＝“P02”の中で使われた樹脂内モノマー ID ＝“M03”の分子量』と指定することで初めて，項目として意味のあるものになるのである。したがって，以下のような項目名を自動生成し，表現できる必要があるということになる。

計測使用配合材 |F01| 配合内樹脂 |P02| 樹脂内モノマー |M03| 分子量

9 そもそもデータ蓄積，共有は何のために行うのか？

前章では，「R & D 部門におけるデータ蓄積基盤としてデータベースがなぜ適切か？」と題し，R & D 部門におけるデータ蓄積，利活用の観点からデータ蓄積基盤として，Excel では限界があり，データベースを採用すべき理由を説明した。本章では，「そもそもデータ蓄積，共有は何のために行うのか？」と題し，データ蓄積，共有の動機に関して，考察を行う。

まず，「なぜ，データを蓄積し，共有しようと考えたのか？」に戻って動機を見つめなおしてみよう。それは，恐らく，「自分の記憶の実験結果だけでは得られない有用な情報が，他者の実験や記憶から消えてしまっている自分の過去実験にあるのではないか？」，「そういう実験結果を簡単に確実に閲覧し，分析できれば，新しい傾向が見つかり，研究が加速するのではないか？」といったものではないだろうか？双方，漠然と参照できるデータが増えれば，「研究にポジティブインパクトがあるのではないか？」とか，「機械学習などの MI が使えるようになり，それらから良いアイデアが提示されるのではないか？」と思っているのではないだろうか？それは，大きな認識間違いである。参照できるデータが増えれば，自分の考えや推測に合う有用な情報に見えるデータを選び出すことがより容易になり，それはある意味データを間違って解釈してしまう可能性を増やすことにつながる。機械学習などの MI に関しては，6 章で話をしたように必要とされる精度，推定力を確保できる適用領域を明確化すること自体が難しく，「良いアイデアが提示されるのではないか？」と思って，利用すると期待を裏切られ，利用者は強い MI アレルギーになってしまいかねない。

「有用な情報」とか，「良いアイデア」というのは，結果論であり，「他者の実験や記憶から消えてしまっている自分の過去実験」自身や「機械学習などの MI の推定値」自身は，常に「有用な情報」とか，「良いアイデア」であるわけではない。つまり，「他者の実験や記憶から消えてしまっている自分の過去実験」や「機械学習などの MI の推定値」が「有用な情報」とか，「良いアイデア」かを判定する装置なようなものがあれば，もちろん，データ共有は研究進展を確実に改善できるのであるが，そんな判定装置がない以上，そういうことに大きな期待を寄せ

てはいけないのである。それであれば，データ共有では研究進展を確実に改善できることはないのだろうか？

結論から言うと，データ共有で研究進展を確実に改善できることはある。それはデータ探査，データ処理に網羅性，均質性，再現性を付与することである。そして，まさしく，データベースの価値はここに見出すべきである。逆に言うと，研究過程において網羅性，均質性，再現性のないデータ探査，データ処理は，研究の品質を著しく損ねるので，行ってはいけないはずである。素粒子実験や天体観測などの大規模実験では，「網羅性，均質性，再現性のないデータ探査，データ処理」は厳しく禁止され，監査されている。それは網羅性，均質性，再現性のないデータ探査，データ処理の結果から導き出された仮説から次の実験を行うと無駄実験になる可能性が高いからであり，大規模実験では無駄実験は金額的にも相当大きい無駄になるからである。しかしながら，これは大規模実験にしか適用されるべきことというわけではない，例え，再実験が安価なものであったとしても本当に間違ったデータ分析から導かれた無駄実験をしてもいいのだろうか？もちろん，データ分析時点でのデータが少なく，結果的に無駄実験になってしまうのは致し方なしであるが，実際にはデータが十分にあったり，全く同じ実験をしているのにデータ抽出，分析方法を間違って，無駄実験をすることはあってはならないはずである。小規模実験では，実際実験をする過程で，一番高価なものは研究者の時間のはずである。無駄実験で研究者の時間は相当無駄に消費されることになり，それがそもそも研究開発の進展速度を遅らせている大きな原因の一つのはずである。データベースにより網羅性，均質性，再現性を担保されたデータ探査，データ処理を使い，研究を実施することで，このような無駄実験はなくすことができるはずなのである。

データベース化は，これ以外にも作業者負担軽減，作業時間短縮というメリットもある。これは，手動処理で網羅性，均質性，再現性を達成できるような業務では主たるメリットであるが，R & D 部門に限って言うと，手動処理（ばらばらに保存されているデータが記録されたファイルをフォルダを駆け回りながら探し，ファイルを開いて・・・）で網羅性，均質性，再現性を達成できることはまずないため，作業者負担軽減，作業時間短縮はあくまで副次的なメリットであり，データベース化の目的は，データ探査，データ処理に網羅性，均質性，再現性をもたらすことと考えるべきである。以下で，網羅性，均質性，再現性に関して，具体的にどういうものか解説を加えておく。

網羅性：手動で全データを対象に探査，処理を行うことは現実的ではないが，データベースだと全データを網羅した探査，処理が可能である。

例）熱伝導度が○○以上の材料を探したい場合に，記憶からいつぐらいの実験にあったはずと目星をつけて，いくつかのデータを確認する。しかし，目星が間違っていることもあるし，そもそも「熱伝導度が○○以上の全ての材料」に目星が付けられるわけではない。このような網羅的でない抽出データは傾向が偏っている可能性もあり，それを分析した結論は，信頼性を大きく棄損している。

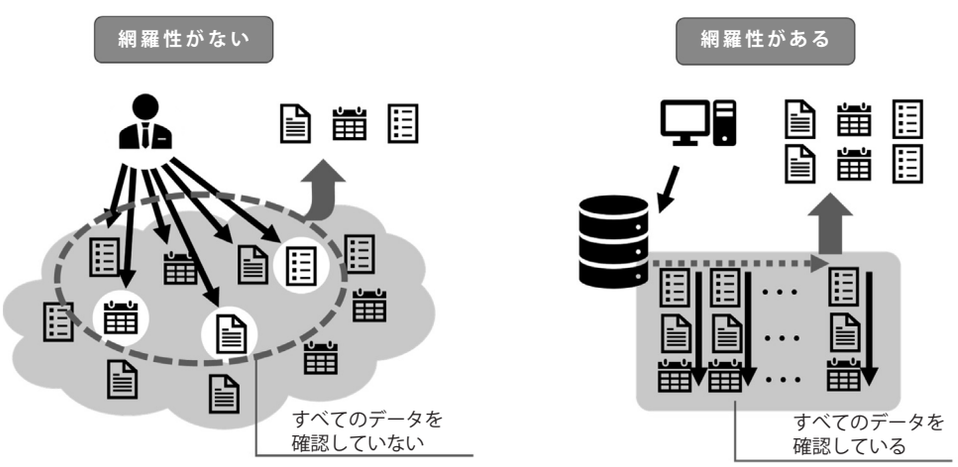

図 17　データベースの利点の網羅性とは

均質性：手動で探査，処理すると意識バイアスがかかったり，初期と終期での探査，処理の同一性が維持できないが，データベースだと完全に均質な探査，処理が可能である。

例）耐電圧が 500V 以上の材料を探している場合に，いくつかのデータを見つけた後，あまりにも該当データが少なかったので，480V 以上のものも含めるようにした。ただ，最初の方から探査をやり直していないので，いくつかのデータは取りこぼしている可能性がある。また，明らかに他の特性が合致しないデータは，500V 以上でも抽出しないことにしたが，明示的なルールで一貫しているかといわれると自信をもって Yes と言い切れないのではないだろうか。このような均質でない抽出データを分析した結論は，信頼性を大きく棄損している。

再現性：半年前のデータ探査や処理は，手動ではまず再現できないが，データベースであれば，データ探査や処理の再現が可能である。

例）手動処理では，上記で述べたように網羅性も均質性もほとんどの場合で担保されていないので，当然，再現性も期待できない。こういう再現性のない抽出データを分析した結論は，信頼性を大きく棄損している。

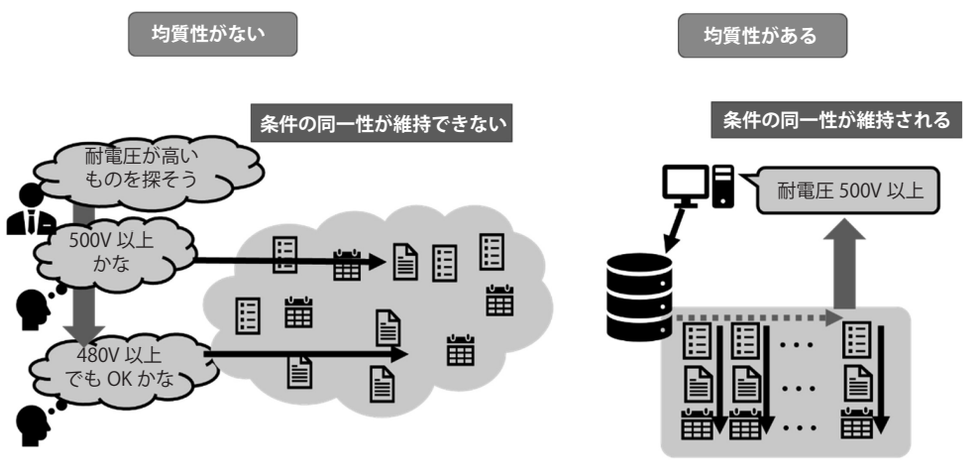

図 18　データベースの利点の均質性とは

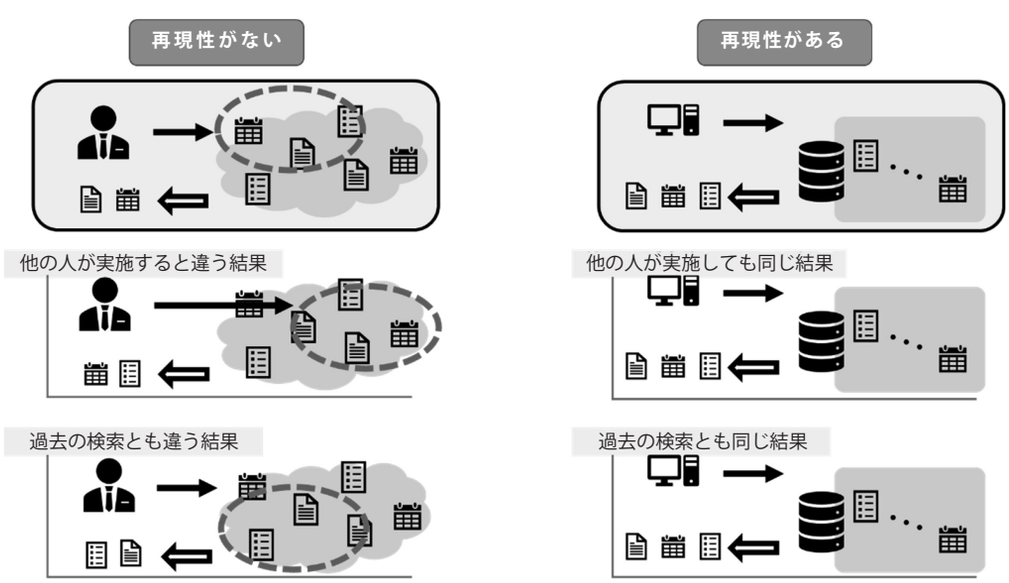

図 19　データベースの利点の再現性とは

10 R & D 部門においてデータベースを機能拡張していく場合の注意点

前章では，「そもそもデータ蓄積，共有は何のために行うのか？」と題し，データ蓄積，共有の動機に関して，考察を行い，蓄積されたデータをどのように利活用すべきかを説いた。本章では，「R & D 部門においてデータベースを機能拡張していく場合の注意点」と題し，データベースを機能拡張（例えば，機械学習機能を組み込むなど）していく場合に注意が必要な点に関して，論じる。

まず，R & D のデータベースには何を期待していいのか？何は期待してはいけないのかに関して，触れておく。結論を一言で書くと，「R & D のデータベースは，魔法の箱ではなく，オルゴール！」である。データベースというものを何でもできる魔法の箱のように思っている人が多い。それはある意味では正しいが，ある意味では間違っていると言える。

業務系のように既に業務内容が詳細に決まっており，それが単純な機能で長期間変化せず多くの人が使い続ける場合は，データベースは魔法の箱のように作りあげることができる。つまり，項目や作業手順（データ絞り込み，データ処理）が明確になっていて，そのデータ絞り込み，データ処理が比較的簡単で，長期間変化しない場合は，そういうソフトウェアを開発し，データベース内にあらかじめ組み込んでおけば，データベースは魔法の箱になる。ただし，非常に簡単な業務内容を実施できるようにするだけでも，データベース開発費は数千万から数億円になってしまうはずである。この業務内容が部門（経理，購買など）で会社に依存しない標準的なものでいいなら，数千万から数億円で開発したものを数百社に販売すれば，数十万円から数百万円の価格になる。そういう意味で，業務系データベースは，自社専用にこだわらない限り，非常に安価な魔法の箱が実現可能なのである。

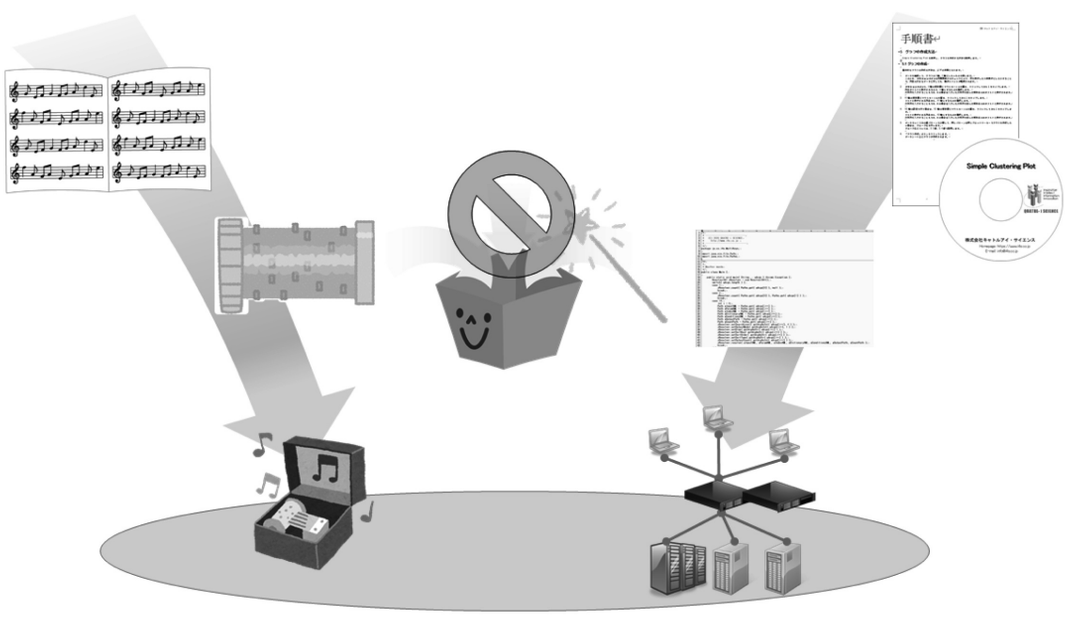

図 20　データベースやシステムは魔法の箱ではなくオルゴール！

一方，R & D 部門では，業務系に比べ業務内容が詳細に決まっておらず，データ分析も一定ではなく人によって様々，つまり，属人的である。また，データ分析には既存の様々なツールを使う必要があり，ツールの機能も業務系のデータ処理とは比較にならないほど多様かつ複雑である。R & D 部門のデータ分析に使われるツールは，多様かつ複雑な機能を持つため業務系データベースのようにデータベース内に開発したものを組み込むことは，現実的でない。例えば，一般的なグラフ機能や画像処理等機能のすべてをデータベースに組み込むなら，それらツールを製品として開発するのと同等以上の費用と時間がかかるようになる。製品として販売されているツールの開発費は，どんなちっぽけなツールであっても，仕様作成，設計，開発，テスト，マニュアル作成を考えると開発費は，数千万円を下らない（数百本売れば，1 本当たり十万円程度になる）。このことを考えれば，R & D 部門のデータベースに一般的なグラフ機能や画像処理等機能のすべてをデータベースに組む込むことが如何に現実的でないかは明白である。

したがって，R & D 部門のデータベースでは，データベース内にツールの機能を作りこむのではなく，データベースから既存のツールを自動起動，自動処理する形で連携できるよう（オルゴールの外箱とドラムのよう）にすることで，開発費の低減とメンテナンス性（ツールの変更や追加）の向上を図る必要がある。

ここまで，データベースというものとツールというものを使い分けて説明をしているが，それらの違いが判るだろうか？誤解をしたまま読み進めないためにも，ここで本書内での使い分けに関して，説明をしておく。

ツール：研究者が欲するデータ処理機能を有するスタンドアローンソフトウェア
様々なデータ処理の機能を有するスタンドアローン（各 PC やサーバ等にインストールが必要な）ソフトウェアで，データの蓄積，共有機能やそれらを複数人で同時に利用するのに必要な運用機能が組み込まれていないもの。最近は，ツールがスタンドアローンではなく，Web サービスになっているものもあり，ツールかデータベースかの判断がし難いものもあるが，ここでは永続的にデータ蓄積，共有する機能の有無でツールかどうかを区別することとする。
例）エクセル，グラフツール，機械学習ツールなど

データベース：共有，検索，保全性等の裏方的機能を有するサーバソフトウェア
データの登録・更新，バックアップ，利用の開始・停止などの複数の人間が同時に同一のソフトウェア（データベース）を利用しても問題が発生しないように設計されているもので，データの保全性や運用管理者と一般利用者などの役割を分けて機能利用ができる機構などの運用機能を備え，検索，処理に網羅性，均質性，再現性を保証することが大きな目的となっているもの。

本来，データベースで解決できることは，データ探査とデータ処理の自動化及び，それらに網羅性，均質性，再現性という品質保証と作業者の省力化を付与することだけである。つまり，データベースは魔法の箱ではなく，データを蓄積し，データをツールへ引き渡す部分はオルゴールの外箱に対応し，どのような探査，処理をするのかはオルゴールのドラムに対応する。どのような音（探査と処理）を奏でるかは，オルゴールを作る前に詳細に決めてられていなければならないのである。言い換えるとデータベースがない状態で，既存ツールのみを使いデータ探査とデータ処理ができるところまでは達成できていなければデータベースは作れないということである。「どのような音を奏でるかは，オルゴールを作る前に詳細に決めておく」ということは，楽譜レベルに落とし込んでおく必要があるということで，実際，オルゴールのドラムに凸凹を刻み込む作業は楽譜がなければできない。「データベースがない状態で，データ探査とデータ処理ができるところまで」というのは，データ探査とデータ処理を誰でも読めば実施できるレベルの手順書にまとめられているということと同義である。そして，ドラムに凸凹を刻み込む作業は，手順書をデータベースに組み込む作業に対応するのである。

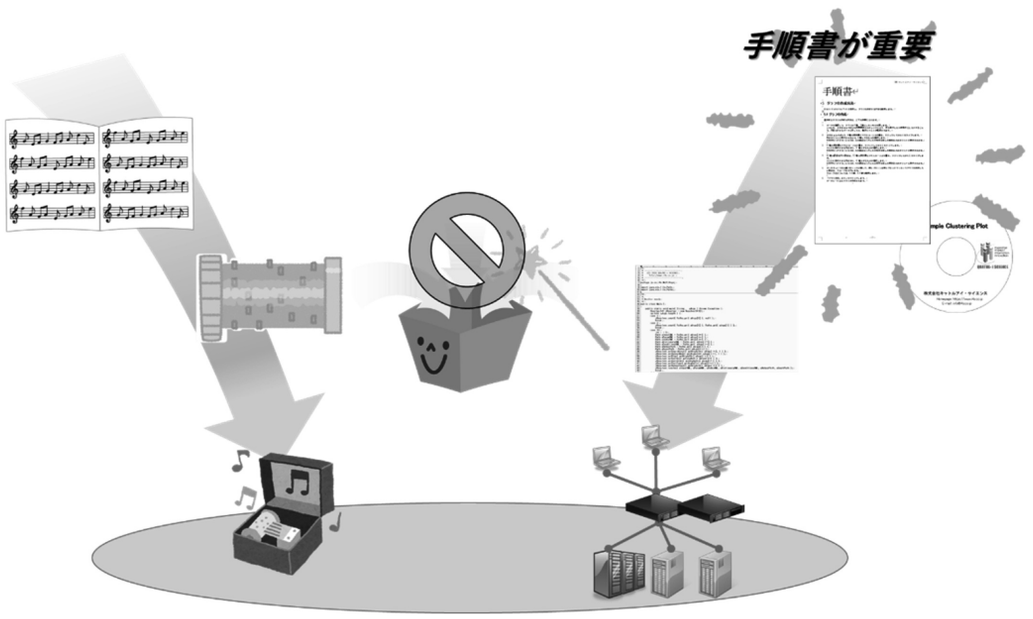

図 21　手動で実施するための手順書がないとデータベースに自動処理は組み込めない

したがって，データベース化で課題を解決するためには，データベースがない状態（ツールは使ってよい）で，研究者でない第三者が手動処理で確実に実施できる手順が確立されている必要がある。実際，この手順書がなければ，データベースの仕様が決められないし，意図通りに作られたかどうかも確認できない。つまり，この手順書を書き下すことがデータ共有・利活用の最初の一歩ということである。

データベース化に便乗して，今まで行っていないデータ処理をこの機にデータベースに実装してしまいたいという要望が挙がることがよくある。しかし，これは絶対に行ってはならないことの一つである。以下，なぜこのようなことは行ってはいけないのかを説明しておく。

1）第三者が手動処理で確実に実施できる手順書がない

データベースは，オルゴールと同じで，楽譜（手順書）があって，それをドラムに刻み込ま（プログラムを組み込む）なければデータベースは完成しないことは前述した。手順書がないということは，楽譜がないということである。楽譜なしに音楽をドラムに刻み込むのは不可能であり，また，即興で楽譜を書いたとしても天才でない限り，まともな音楽になっていることはまずない。実際，すでに手動実施していることを手順書に書き下すことでさえ，悪戦苦闘する実態を考えると，如何に現実的でないか理解できると思う。当然，そのような楽譜や音楽は，オルゴールに組み込むべきではない。

2）その作業が本当に価値あるものかどうかが疑わしい

即興で実際に実施可能な手順書が書けたとしても，「その手順通り実施した結果が価値の高いものか？」や「その実施頻度がどの程度か？」は，保証されない。つまり，そもそも手動でまだ，実施していない程度の作業は，それほど重要度が高くないはずである。そもそも，R & D 専用のデータベースであったとしてもデータベースへの作業自動化の組み込み自体もそれなりに手間のかかるものである。組み込んで使ってみて，「何か思っていたのと違う」，「使いにくいので，ここをこうしてほしい」などを言い出すようでは，いけない。そもそも，データベースが存在していなくてもその作業を行っていて，それを正確に手順書の書き下せていたのであれば，「何か思っていたのと違う」，「使いにくいので，ここをこうしてほしい」は，発生しえない。実際には，「使いにくいので，ここをこうしてほしい」は発生しえるが，データベース無しで実施していた時よりはましなはずで，作業時間も処理の網羅性，均質性，再現性も上がっているはずで，十分費用対効果があるはずである。その上で，「使いにくさ」を改善するのは，費用対効果が高いとは言えないことであり，

よほど他に改善する余地のないＲ＆Ｄ以外では，行うべきでない課題である。

実際，「今まで行っていないデータ処理をこの機にデータベースに実装してしまいたい」という要望が上がってくるのは，データ探査，データ処理のためのツールが存在しないからであることがほとんどである。そもそもデータベースは検索した結果をツールに引き渡すだけぐらいの作業しかしないので，データベースに組み込まれていなくとも検索結果をツールに引き渡してあげる作業を手動で行うだけで，さほど大きな手間にはならないはずである。したがって，まずは，スタンドローンツールとして，ツール単独で利用者に使える状態にし，その状態で利用者が十分に利用するまではツール開発として行っていくべきである。そして，本当によく使うものは，「データベースは検索した結果を当該ツールに引き渡す」機能拡張を行えばよい。

「今まで行っていないデータ処理をこの機にデータベースに実装してしまいたいという要望」の中で，ツール自体もまだ存在していない場合もある。当然このような作業のデータベース化は，上記理由より行うべきではないが，「ツールが存在していない」状態では，データベース化を行ってはいけないもう一つの重要な理由を以下で説明しておく。

3）ツールとデータベースの境界が曖昧になる

Ｒ＆Ｄ部門では項目追加，変更だけでなく，データ処理に関しても追加，変更が頻出する。これが業務系と異なり，データベース化を難しくしている原因である。このようなＲ＆Ｄ部門でもデータベース化を可能にするためには，ツールとデータベースの分離が重要で，データベースからツールを呼び出す形のデータベースでなければならない。そのようにすることで，データベースなしにツールを使い，手順を確立することが可能になる。しかし，ツールとデータベースを同時に開発すると，手動手順の確立が難しくなってしまう。また，同時開発はツールとデータベースの境界を曖昧にしてしまい，その結果，ツール部のみの改良やデータベースを改良したときにツール部に問題が生じないかの確認，検証が難しくなる。つまり，ツールやデータベースのバージョンアップ時にデータベース自体の長期の停止が必要になるなど，メンテナンス性が悪くなるのである。

4）ツール部，データベース部及び手順の問題の切り分けができない

ツールとデータベースを同時に開発すると，何か問題があったときに，手順に問題があったのか？，ツール仕様に問題があったのか？，データベースに問題があったのか？の切り分けが非常に難しくなる。そもそも，データベースを使わない状態での正解が示されないとその判別自体が原理的にはできないということになる。データベースの問題であることを排除するためには，手動でツールのみを使った手順の確立が必須である。もちろん，ツールの問題であることを排除するためには，ツールを使わない手動手順の確立が重要である。したがって，ツールを自作する場合には，他の複数の商用ツールなどでも同じことができる手順を確立しておくべきである。また，データベース開発会社にすべてを丸投げしてしまう場合は，丸投げ体質が染みつき，もし，ツール部やデータベース部に問題があっても，気づかず間違った分析をしたままになる可能性が高くなる。実際，正確な仕様書と手動手順書がないとデータベース会社も何が正しいのかを把握できないことになるので，正解がわからないところに丸投げをするという非常に危機的な状態に陥る。

「それでは，将来展望など，現時点でできないことを実現するためにはどうすればいいのか？」という質問がよくある。すでに答えは上述しているが，以下でまとめておく。

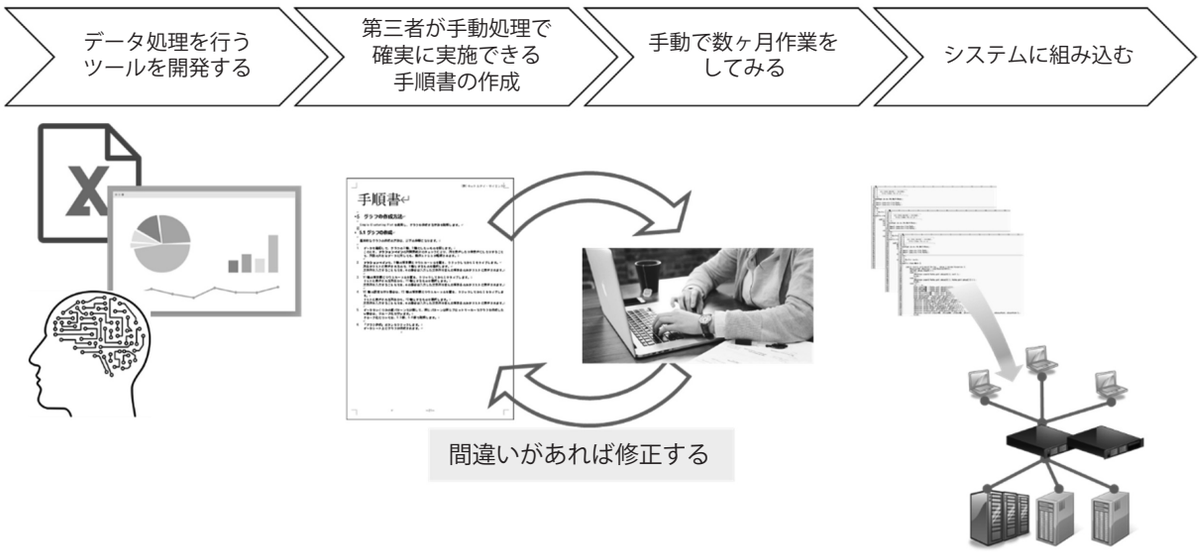

図 22　手動で行えていないことをシステムに組み込むまでの手順

1）データ処理を行うツールを開発する

データ処理を行うツールが存在しないのであれば，まず，手動手順での実施を可能にするためのツール作成を行う。ツールの動作を確認する為のサンプルデータとその結果も整えておくこと。実際，ツールに不具合がないことを保証するのは，相当大変だが，データベースに組み込む前に十分確認しておかないと，皆がツールには問題がないとして使いだすので，十分注意が必要である。もし，ツールをデータベースに組み込んだ状態でツール機能をそのまま利用開始した 1 年後などに問題が見つかった場合は，間違った結果もデータベースに蓄積されてしまうので，多くのデータ訂正が必要になる。したがって，その連絡および関係部署への影響範囲の確認などが発生し，結構大きな問題になる。これが，単独ツールでデータ蓄積がない状態との大きな違いであり，多くの人が意識できていないことである。

2）第三者が手動処理で確実に実施できる手順書の作成

ツールを使って，手動で実施可能な手順を確立し，手順書を作成する。また，この過程で，データの項目名や項目値の単位がばらばらだとか，数値の項目値に数値以外の値が入っている（>1，～ 10，〃とか，目視では意味が分かるが，データベースには意味判別ができない記述）ものは，適切な数値に修正をしたり，新たな項目（下限値，上限値，信頼幅など）を作り，その値に変更するなどの作業をする。この作業は一般的にデータクレンジングと呼ぶ。また，この手順書作成過程で，ツールの機能が利用者のニーズにマッチしているかどうかの確認が行われることになる。一般的に，この段階でツールが有用と判断されているならばツールは何度もバージョンアップされているはずである。

3）手動手順書に基づき手動実施を数ヶ月行う

手動手順書に基づき，手動実施を数ヶ月間行う。ここで，手動実施が数ヶ月間で数回程度のものは，ツールの機能不十分さや利用方法の教育が十分でないため，使われない場合もあり得るので，ツールの改善及び教育の強化を行うべきである。

4）データベース組み込み意義が高いと判断されたらデータベースに組み込む

ツールの改善ペースが落ち着き，手動実施回数も多くなってきたら，今後の利用想定頻度を提示してもらい，また，データベース連携することで何がどの程度改善するのかを提示し，データベースに組み込むかの合意形成をする。実際，3）のままに比べ，ツールをデータベースに組み込むことをして，格段の改善がある事は稀なので，そこを冷静に客観的に判断することが重要である。一般論として，データベース組み込みをしたほうがいいのは，ツールの出力結果をデータベースに保存することまで，行う場合である。この場合は，手動でデータベースに登録すると入力ミスや紐づけ間違いが発生するので，そ

の頻度やそれら間違いを確認，訂正する工数を定量的に評価すると，データベースに組み込んだ方が良いと判断できる場合が多いのである。

参考文献

1）川田重夫，田子精男，梅谷征雄，南多善，上島豊，他
PSE book ―シミュレーション科学における問題解決のための環境（応用編），
川田重夫，田子精男，梅谷征雄，南多善　共編，培風館，(2005)，p69-82

2）谷啓二，奥田洋司，福井義成，上島豊
ペタフロップスコンピューティング，
矢川元基　監修，培風館，(2007)，p183-202

3）牧野圭一，上島豊，視覚とマンガ表現，臨川書店，(2007)，p1-5, 221-229

4）上島豊，月刊「研究開発リーダー」8 月号，技術情報協会，(2020)，p33-37

5）上島豊，月刊「研究開発リーダー」9 月号，技術情報協会，(2020)，p53-57

6）上島豊，月刊「研究開発リーダー」1 月号，技術情報協会，(2022)，p58-63

7）上島豊，月刊「研究開発リーダー」2 月号，技術情報協会，(2022)，p46-50

8）上島豊，月刊「研究開発リーダー」3 月号，技術情報協会，(2022)，p62-65

9）上島豊，月刊「研究開発リーダー」6 月号，技術情報協会，(2023)，p63-68

10）上島豊，月刊「研究開発リーダー」7 月号，技術情報協会，(2023)，p86-91

11）上島豊，月刊「研究開発リーダー」8 月号，技術情報協会，(2023)，p78-82

12）上島豊，他
研究開発部門への DX 導入による R & D の効率化，実験の短縮化，技術情報協会，(2022)，p195-221

13）上島豊，他
ケモインフォマティクスにおけるデータ収集の最適化と解析手法，技術情報協会，(2023)，p39-74

14）上島豊，他
実験の自動化・自律化による R & D の効率化と運用方法，
技術情報協会，(2023)，p159-199