

《連載 全 3 回》

R & D 部門における データ共有システム構築の事前準備の要所（上）

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『ベタフロップス コンピューティング』ISBN：978-4-563-01571-8
 臨川書店『視覚とマンガ表現』ISBN：978-4-653-04012-5

1 はじめに

現在の R & D 領域では、データ分析や共有は、極めて属人的な扱いである。客観的なデータ生成、分析が要求される理学、工学領域で、この属人性は大きな問題を孕んでいる。研究というものとは創造的な活動であり、個人の才能、発想に起因する「なぜ、そう考えたか？」の部分に属人性が必要なことは当然である。しかし、どのようにデータを生成し、どのように分析し、結論を導いたかは、属人的では問題で、客観的かつトレサブルであるべきである。実際、センサーや計算機的能力向上により、データの生産性が向上し、扱うべきデータが膨大になり、詳細記録の欠如、偶発的データ取り違い、主観的データ操作が発生する余地が増大し、データ生成、分析プロセスの信頼性が大きく揺らいでいる。

私は弊社を設立する前は研究機関にてコンピュータ、ネットワークの最先端技術を駆使し、自然科学、工学研究を約 10 年間行っていた。その中で R & D データが属人的処理され、その共有状態がデータの信頼性及び有効活用性を大きく阻害し、共有化及びインフォマティクス分析、AI 化が進まないことを経験した。

本記事では、私自身の 10 年の R & D 経験と弊社の 15 年の R & D 支援実績から得た「R & D 部門におけるデータ共有システム構築の事前準備の要所」に関して、簡単に解説する。

2 属人的データ共有状況を脱するための事前準備の前に行うべきこと

本章では、データ共有システムを導入し、属人的共有状況を脱するために必要な事前準備の前に、どのようなことが必要かを説明する。

属人的、形骸化したデータ共有状態を脱する為には、データベース化 / システム化以前に行うことが沢山ある。データベース化 / システム化以前の単純なファイル共有レベルで、下記状況から脱却できている必要がある。

-1) 属人的データ共有状況からの脱却

-1.1 情報は記憶でなく、記録する。

⇒記憶は本人以外には見えないうえ、本人でさえ、時間が経つと忘却したり、改ざんされる。

-1.2 第三者が見て、認識に齟齬が出ないように項目名を定義し、合意を形成する。

⇒いくら記録をしても皆が同じ意味でとらえないなら、情報を共有していることにはならない。項目が増えてくると同じ項目に異なる項目名をつけたり、異なる項目を一つの項目として使ってしまうことが発生しうる。

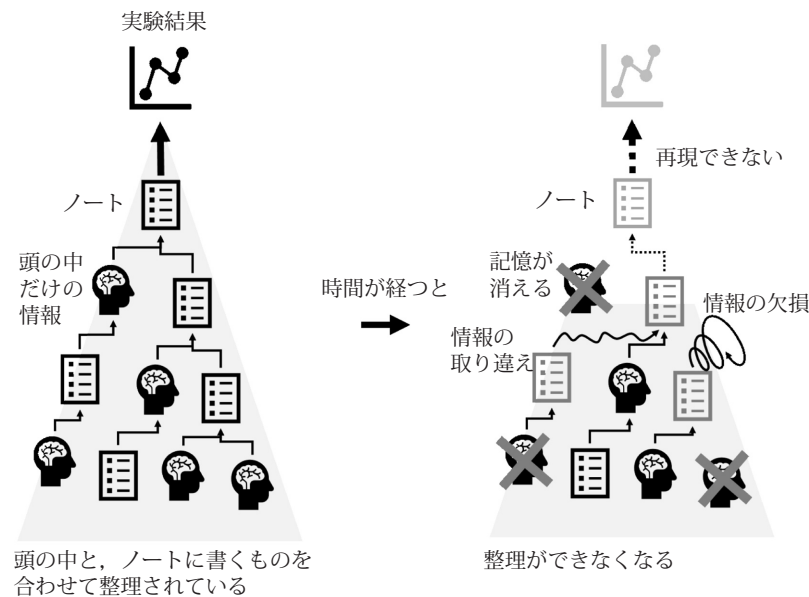


図1 情報は記憶でなく記録すること

0) 形骸化したデータ共有状況からの脱却

0.1 実験や結果分析の再現に十分な情報を特定・合意し、それを記録する

⇒ 誰しもが認識していない隠れパラメータは仕方ないが、そうでない限り、実験や結果分析の再現ができない記録はデータ共有の価値が大きく棄損し、実際、賢明な第三者は参照しなくなる。

0.2 利用、活用の観点から項目は適切に設定する

⇒ 例えば、ポリマ濃度 / モノマ濃度で絞り込むことが多いのであれば、ポリマ濃度、モノマ濃度という項目があったとしてもポリマ濃度 / モノマ濃度という項目がないと、全データを対象に演算を行うことが必要になる。それは、データ探査において、常に全データを取得する必要があるということであり、現実的でない。データベースも項

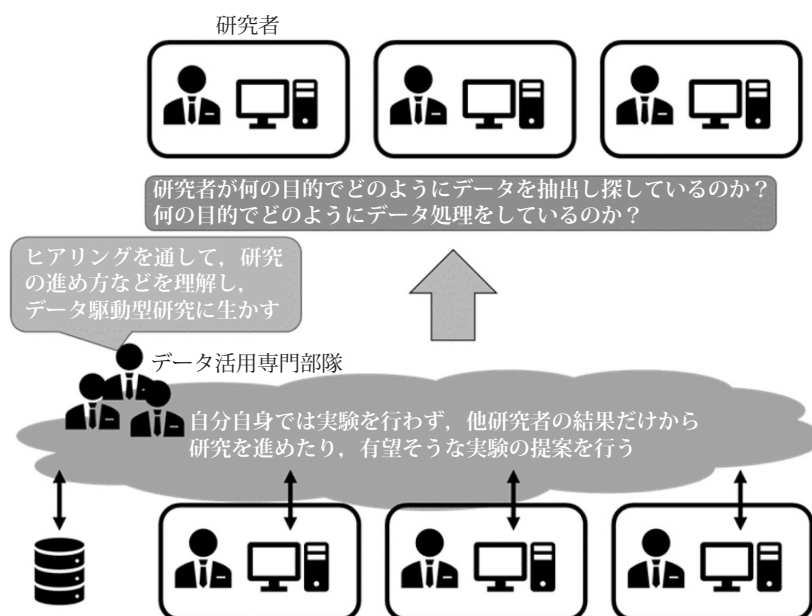


図2 データ活用専門部隊によるデータ探査、処理実態調査

目間の演算はできないので、データベース化したからといって解決する問題でもない。また、グラフ作成時の X,Y 系列に良く採用する項目に関しても、同じことが言える。つまり、データ共有をするからには、データ絞り込み、グラフ化でよく使うものは、あらかじめ項目化しておくべきということである。

これらは、データベース化、システム化という IT 観点で必要な事柄ではなく、データ共有のために必要な事柄である。したがって、それらを混同した議論や問題を起こさない為にも、これら対応が終わってからデータベース化、システム化の取り組みを行うことが肝要である。

3 R & D 部門におけるデータベース、システムは、魔法の箱ではない！

次にデータベース化、システム化のために直接的に必要な準備に関して、説明をしていこうと思うが、その前にデータベース、システムというものは、どういうものなのか？ R & D のデータベース、システム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、システムは、魔法の箱ではなく、オルゴール！」である。データベース、システム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、システム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、システム内にあらかじめ組み込んでおけば、システムは魔法の箱である。ただし、非常に簡単な業務内容を実施できるだけでも、システム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家に販売すれば、数十万円から数百万円の価格になる。そういう意味で、業務系システムは、自社専用にとどまらない限り、非常に安価な魔法の箱が実現可能なのである。

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。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システムのようにシステム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システムに組み込むなら、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 部門のシステムに一般的なグラフ機能や画像処理等機能のすべてをシステムに組み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 部門のシステムでは、システム内にツールの機能を作りこむのではなく、システム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、システム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

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、システム：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、システム）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、システム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、システム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し、データ処理を自動化する部分はオルゴールの外箱で、どのような探査、処理をするのかはオルゴールのドラムである。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。言い換えるとデータベース、システムがない状態で、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、システムは意味がない、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ならない。「データベース、システムのない状態で、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、楽譜ではなく、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、手順書をシステムに組み込む作業に対応するのである。

したがって、データベース化、システム化で課題を解決するためには、データベース、システムがない状態（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、システム仕様が決められないし、システムが意図通りに作られたかどうか確認できない。つまり、この手順書を書き下すことがデータ共有・利活用の最初の一步ということである。

4 データベース、システムの最大の利点とは！

ここで、そもそもなぜ、データ共有・利活用をするためには、データベース化、システム化をすべきなのかを考えておく。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、システム化は本質でないと言えはその通りである。しかし、オルゴールの例を思い返してみると、手順書止まりということは楽譜止まりで、データベース化、システム化をするとオルゴールになるのである。何が違うかというと、オルゴールは、ドラムを作った人の意図通りに常に同じ音楽を何度でも奏でるが、楽譜は奏者によるばらつきが生まれ、それこそ熟達した奏者でないとメトロノームがないと一定テンポで演奏し続けることさえ難しい。オルゴールの価値は、まさしく音楽を

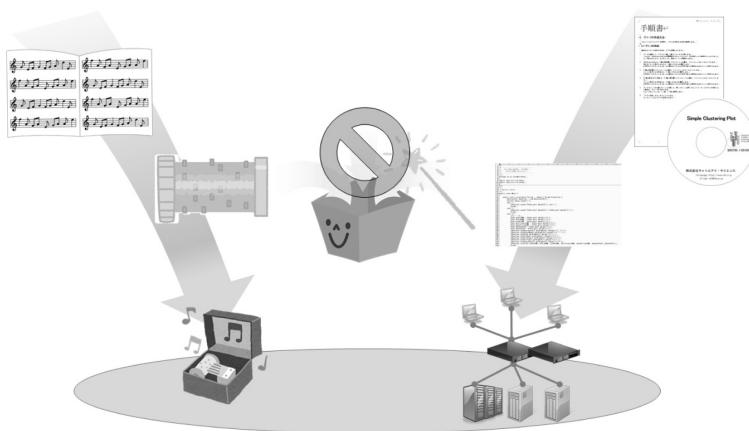


図3 データベースやシステムは魔法の箱ではなくオルゴール！

均質に何度も再現することができることである。つまり、データベース、システムは、オルゴールと同じようにデータ探査、データ処理に網羅性、均質性、再現性を付与するためのもので、まさしく、データベース、システムの価値はここに見出すべきである。データベース化、システム化は、これ以外にも作業負担軽減、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、手動処理で網羅性、均質性、再現性を達成することはまずないため、作業負担軽減、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、システム化の目的は、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。そう考えられない場合は、「データ探査、データ処理に網羅性、均質性、再現性」がないと、どのようなことが起こるかを再度思い返して

ほしい。以下で、網羅性、均質性、再現性に関して、少し解説を加えておく。

網羅性：手動で全データを対象に探査、処理を行うことが現実的ではないが、システムだと全データを網羅した探査、処理が可能である。

例) 熱伝導度が〇〇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、そもそも目星が「熱伝導度が〇〇以上の全ての材料」に着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。

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、初期と終期での同一性が維持できないが、システムだと完全に均質な探査、処理が可能である。

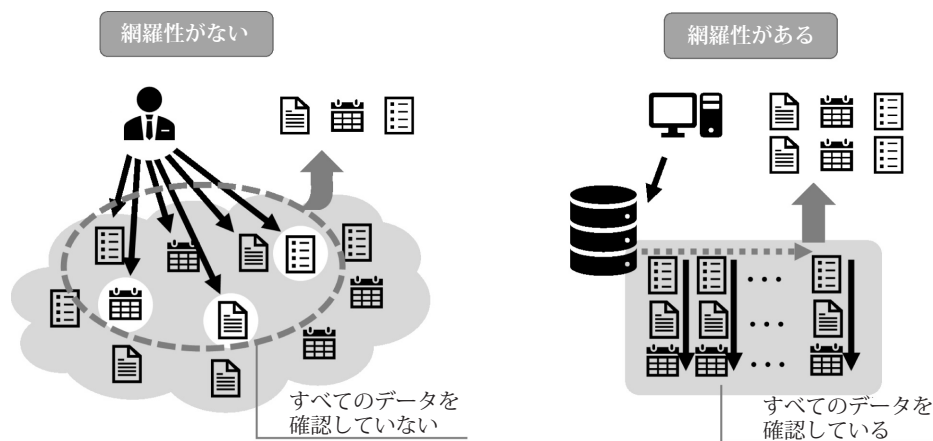


図4 データベースの利点の網羅性とは

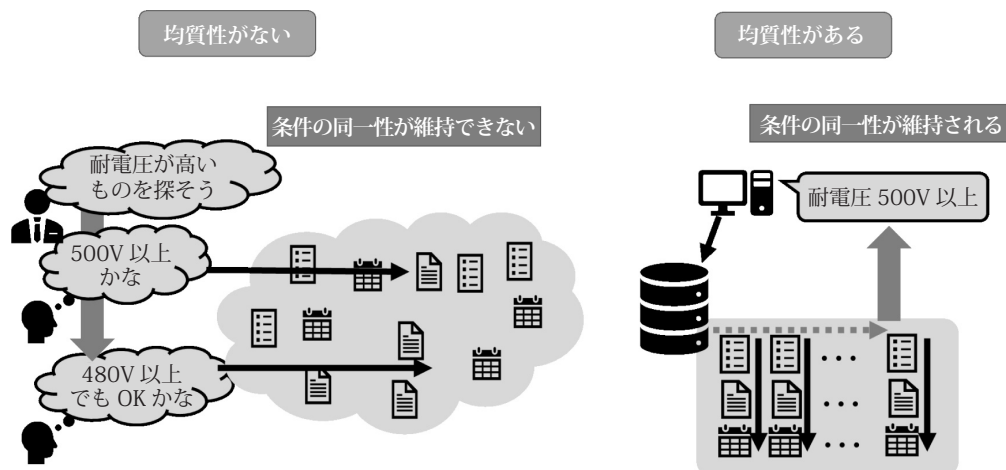


図5 データベースの利点の均質性とは

例) ガラス転移点が 212 度以上の材料を探している場合に、いくつか見つけた後、あまりにも該当データが少なかったため、210 度以上のものも含めるようにした。ただ、最初の方から探索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、212 度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グレーである。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：半年前のデータ探索や処理は、手動ではまず再現できないが、システムであれば、データ探索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。

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 ベ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

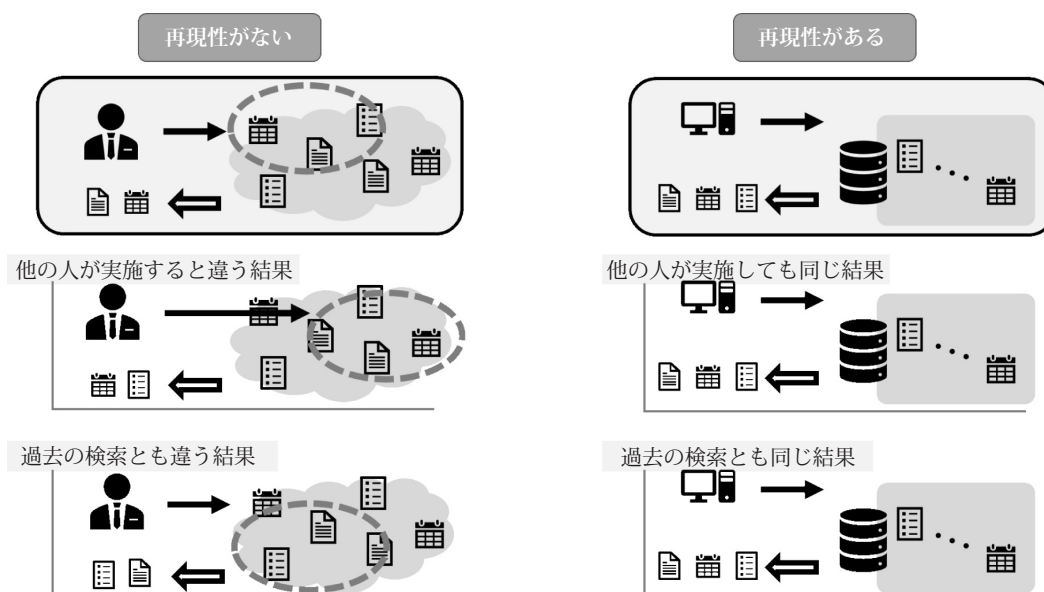


図6 データベースの利点の再現性とは