R & D 部門における効果的なデータ管理手法と管理体制の作り方（上）

上島 豊　　（株）キャトルアイ・サイエンス　代表取締役

《PROFILE》

略歴：

1994 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

1　はじめに

現在のR & D 領域では，データ分析や管理は，極めて属人的な扱いです。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ます。研究というもの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す。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，客観的かつトレーサブルであるべきです。実際, センサーや計算機の能力向上により，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え，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ます。

実際，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 年間行なっていました。その中でR & D データが属人的処理され，その管理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ました。

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 部門における効果的なデータ管理手法と管理体制の作り方」に関して，簡単に解説します。

2　R & D 部門におけるデータ管理の実情

R & D 部門におけるデータ管理の実情の話の前に，「データ管理」とは，何を意味しているのかを説明したいと思います。「データ管理」とは，読んで字のごとくデータを管理することなのですが, より具体的に言うと，データを生み出した実験及び解析を第 3 者が再現するために必要な情報を記録し，それを必要な時に迅速かつ確実に参照できる状態に保っておくことを意味します。そういう観点において，ほとんどのR & D 部門におけるデータの管理は管理というレベルには達しておらず, 単なる蓄積と呼ぶのが相応しいというのが実情です。公的，民間の様々なR & D 部門を見てきた結果，データがどのように蓄積されているのかを，以下でご紹介します。

一人で完結できるような実験や解析では，ほとんどの情報は研究者の頭の中のみにあり，注目しているパラメータのみを研究ノートにメモ書きをされているだけの場合があります。実験や解析結果の比較評価がある程度難しい課題に対しては，実験や解析の情報がエクセルに書き写され，比較しやすいように纏められ，個人 PC 内に保存されていることもあります。複数の人が関わった実験や解析の場合も上記状況と変わらない部分もありますが，他の人への実験や解析の引き渡し（依頼）に必要な情報のみは，フォーマットが揃えられた用紙もしくはエクセルが準備されていることが多いと思います。また，それら用紙はキングファイルなどにファイリングされたり，エクセルファイルなどのデジタルファイルは共有のファイルサーバなどが準備され，そこに保存するようにされているところもあります。これらは，「データを生み出した実験及び解析を第 3 者が再現するために必要な情報を記録し，それを必要な時に迅速かつ確実に参照できる状態に保っておく」という観点に立つと，「データ管理」ができていないということになります。これらの状況では，研究者本人は何らかの形で頭の中では整理ができていると思っています。したがって，この状態を「属人的なデータ管理」と呼ぶことにします。

稀ですがデータ管理の意識が高い組織では，しっかりしたデータベースシステムを導入し，実験データを蓄積，検索できるようになっているところも存在します。しかしながら，その蓄積されたデータは，まさに蓄積をする以外において利用，活用されていないことが多いのです。データベースシステムがあり，運用され，データが蓄積されているのに，データの利用，活用が進まないということは，実質的には「データ管理」ができていないということになります。この状態を「形骸化したデータ管理」と呼ぶことにします。

3 属人的なデータ管理，形骸化したデータ管理状況から生まれる問題点

前章で，ご説明しました通り，ほとんどの R ＆ D 部門では，「データを生み出した実験及び解析を第 3 者が再現するために必要な情報を記録し，それを必要な時に迅速かつ確実に参照できる状態に保っておく」ことは達成されておらず，「属人的な，もしくは形骸化したデータ管理」状況になっています。本章では，「属人的な，もしくは形骸化したデータ管理」状況がどのような問題を生み出すかについて論じます。

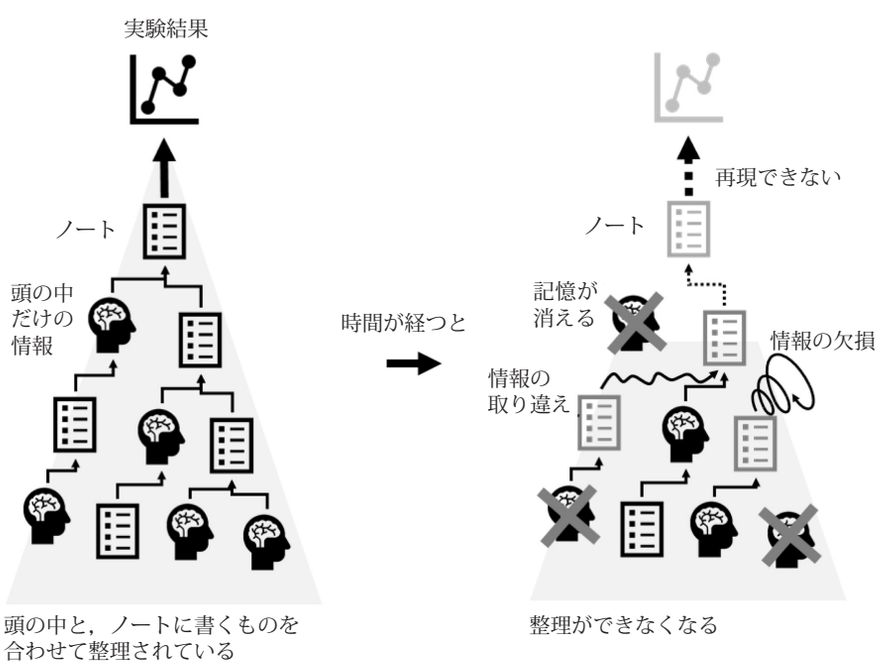

図 1　属人的なデータ管理とその特性

「属人的なデータ管理」状況では，実験及び解析の詳細なことは，実際の実施者しかわからない状況になります。当然のことながら実施者以外の人が実験条件や結果内容を知ろうとした場合，実施者にそれらを聞くしかありません。実施者から実験及び解析を第３者が再現するのに十分な情報を提供してもらえれば問題はないのですが，実験データをそういうことができるような状態で蓄積している研究者はほとんどいません。そもそも，どれだけ整理好きの研究者でも，「○○の実験・解析情報，データ」が欲しいと言われたところで，実施者とて該当するデータを探し出すこと自体，相当困難なことが多いはずです。該当するデータが漏れなく，間違いなく提供されることは，ほぼ不可能と考えてもいいかもしれません。このような状況の中，データの授受を行なうと，間違った情報を基に実験や解析を進めることが発生し，間違った結論が導かれたり，検討を進めたのちに始まりに立ち返って，再実験や再解析を行なわざるを得なくなることもあります。

そういう事態に何度か遭遇すると，研究者同士の信頼関係が崩れたり，他の研究者のデータを参照せず，自分が実施した実験や解析のみしか参照しないような状態が進んでいきます。このような状態は，研究自体の属人化を加速し，研究の蛸壺化，継承不可能化を進めてしまいます。

「形骸化したデータ管理」状況は，「形骸化」＝利用，活用されていないこと自体が問題です。どんなデータ蓄積システムでも構築自体にコストがかかり，運用，メンテナンスにもコストがかかります。また，そこにデータを登録することも研究者に余分な作業を発生させることになります。それだけコスト，労力をかけても，利用，活用されないのであれば，システム化しない方が良いのではないでしょうか？システムが構築され，それが稼働していると，それだけで何かが良くなったと思いがちですが，研究効率という観点ではシステム化以前より悪くなっていることもあるのです。「データを生み出した実験及び解析を第３者が再現するために必要な情報が記録されている」にも関わらず蓄積されたデータが利用，活用されていないのであれば，利用，活用を促す仕組みづくりを見直すべきです。しかしながら，「データを生み出した実験及び解析を第３者が再現するために必要な情報が記録されていない」場合が多いのも事実です。実際，データが蓄積されたシステムが利用，活用されていないので，その発覚が遅れ，大量の不完全なデータ蓄積，つまり，役に立たないデータ蓄積に終わってしまうことが多いのです。

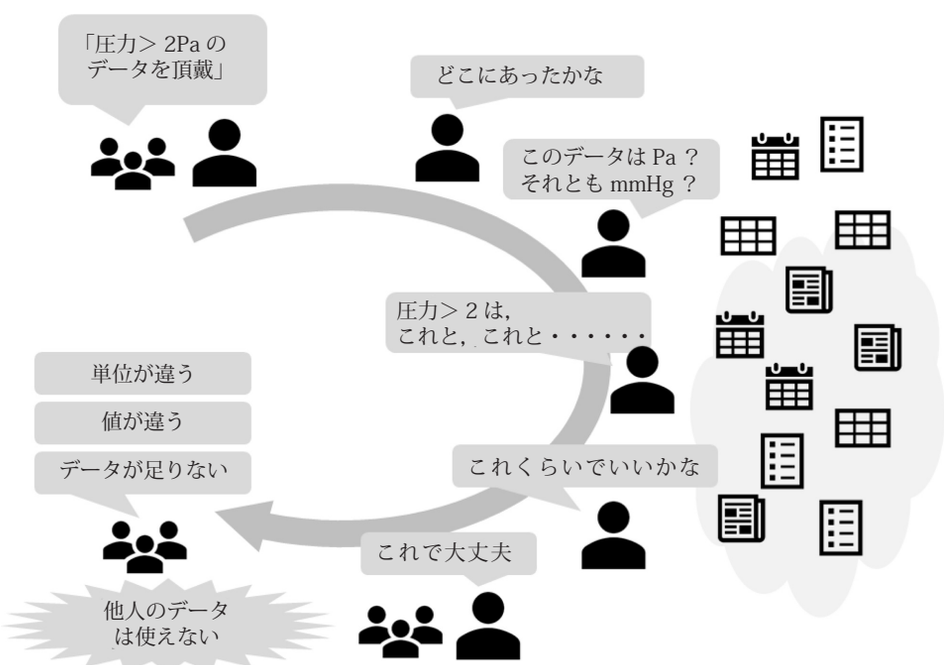

図２　属人的に管理されたデータが引き起こす問題

4 属人的なデータ管理，形骸化したデータ管理状況が生み出される原因

前章では，「属人的な，もしくは形骸化したデータ管理」状況が，どのような問題を生み出すかについてご説明しました。本章では，そのような問題があるにもかかわらず，なぜ「属人的な，もしくは形骸化したデータ管理」状況が生み出されるのか？，その原因に関して考察します。

「属人的なデータ管理」状況が生み出される直接的原因は，記録が必要な項目及びそのフォーマットが決まっていないからです。R & D 部門の業務は，新しい実験及び解析方法は日進月歩で生まれていくため，固定的に項目やフォーマットを決めても，すぐ使い物にならなくなります。したがって，「属人的なデータ管理」状況が生み出される真因は，「R & D の変化に追従できる項目及びフォーマット」と，その「利用徹底を遂行できる運用体制」がないことということになります。「変化に追従できる項目及びフォーマット」は，ある程度技術でカバーできますが，「利用徹底を遂行できる運用体制」というものは，なかなか難しい問題を孕んでいます。

実験及び解析を行なう場合，実験及び解析の全てのパラメータを変化させることは，まずありません。目標とする特性，性能値を達成するためのいくつかの主要因を想定して，ごく限られたパラメータのみを振って，実験及び解析を行なうはずです。その場合は，その他のパラメータは一定に保つものであり，いちいち記録に残さなくとも頭の中に残っているので，記録を省略することが多いのです。当然，実験及び解析の最中や直後は，記録がなくとも実施者自身が実験及び解析の再現は可能です。目の前の実験及び解析の結果に意識を集中しすぎるあまり，その時はその他の一定パラメータは忘れるわけがなく記録するまでもないと感じてしまうので，記録を取ることを省略してしまうのです。つまり，「変化に追従できる項目及びフォーマット」があったとしても，記録する，しないに属人性が入り込んでしまい，「実験及び解析を再現するのに十分な情報」が記録されない状態になってしまいます。また，そもそも実験及び解析の最中の研究者は，実験及び解析パラメータを忘れるわけがないという意識があり，「実験及び解析を再現できるだけの情報記録」をしていないその状態をそんなに悪い状態ではないと認識をしている研究者が多いことも属人性に拍車をかけています。

そもそも人間，というか生物一般は，論理的に物事を考えるよりも，多数の経験をすることで，パターンマッチング的な認識を形成することが得意なように作られています。人の顔を覚えたり，物の名前を覚えたり，歩行や自転車に乗ったり，水泳を覚えるときもうまくいった状況を再現するための全ての情報を筆記記録して，うまくできるようになるわけではありません。もちろん，それらで習得したパターンマッチング的な認識は，属人的なもので他人には共有できません。そして，当の本人には，共有することのメリットは感じないのが普通です。結局，この属人的なパターンマッチング的な認識能力があることと，そもそもその属人的な認識を他人と共有することのメリットを意識できていないことがあるので，「実験及び解析を再現できるだけの情報記録」をすることの必要性を意識しにくいのです。

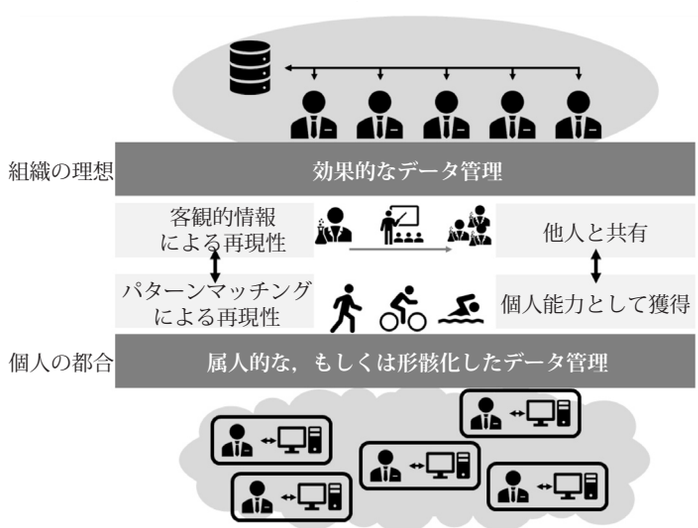

図 3　データ管理に関する組織と個人の乖離の原因

「形骸化したデータ管理」状況が生まれる原因は，「属人的なデータ管理」状況が生み出される原因である「属人的なパターンマッチング的な認識能力がある」ということと根本的な部分では同じではありますが，ここではもう少し掘り下げて，考察しておきます。「形骸化」＝利用，活用されていない直接的原因は，大きく２つあります。一つは，何かのデータを探したいと思った時に，どのように探せばいいのかわからない，そもそも探せるようにシステムが作られていないということです。どのように利用するのかを十分調査，分析をしないでシステムを作ってしまうことが直接的な原因です。「どのように利用するのか」を調査，分析することは，実はそんなに簡単なことではありません。実際，システムが存在しない状態で「どのように利用するのか」と聞いたところで，それは想像の中で「こう利用すると便利そうだ」と思っているだけで，実際にそのように利用して便利かどうかを保証した発言ではないのです。また，普段は「属人的なパターンマッチング的な認識能力」を使い（＝半分無意識で）データを探しているので，データの探し方は体系化，アルゴリズム化されているわけではなく，他人に説明できるようにはなっていないのです。そのような状況下での「どのように利用するのか」の調査，分析がいかに困難かは想像できると思います。

もう一つは，蓄積されたデータが不完全で，「実験及び解析を再現できるだけの情報記録」となっていないことです。このような状態では，前章でも述べた「間違った結論が導かれたり，検討を進めたのちに始まりに立ち返って，再実験及び再解析を行なわざるを得なくなる」ことが起こってしまい，システムを信頼しなくなり，使われなくなってしまうのです。研究者同士の信頼関係が崩れないだけまだましですが，データ管理という側面では実質的には失敗です。なぜ，「実験及び解析を再現できるだけの情報記録」とならないかは，「属人的なデータ管理」状況が生み出される原因と同じ理由です。また，それを未然に検知できなかったのは「実験及び解析を再現できるだけの情報記録」とは何かを調査，分析できていなかったことが原因です。

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