◆新連載◆

《連載 全 3 回》

第 1 回　R & D 部門におけるデータ共有，利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

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だ。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ある。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る。

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た。

本記事では，私自身の 10 年の R & D 経験と弊社の 17 年の R & D 支援実績から得た「データ共有，利活用のためのデータ記録，蓄積，分析方法」に関して，簡単に解説する。

2　R & D 部門におけるデータ共有，利活用の実情

R & D 部門に関わらず，データが共有，利活用されるためには「データが管理された状態」になっている必要がある。データ共有，利活用の実情の話の前に，「データが管理された状態」とは，何を意味しているのかを説明する。「データが管理された状態」とは，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を意味する。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は，管理された状態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ある。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，以下で紹介する。

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る。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 Excel 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る。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るが，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 Excel が準備されていることが多い。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る。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おり，以下ではこの状態を「属人的なデータ管理」と呼ぶことにする。

「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る。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ない。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ない。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，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ある。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ない。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る。そういうことを続けていく中で，データの共有，利活用は次第に廃れていってしまうのである。

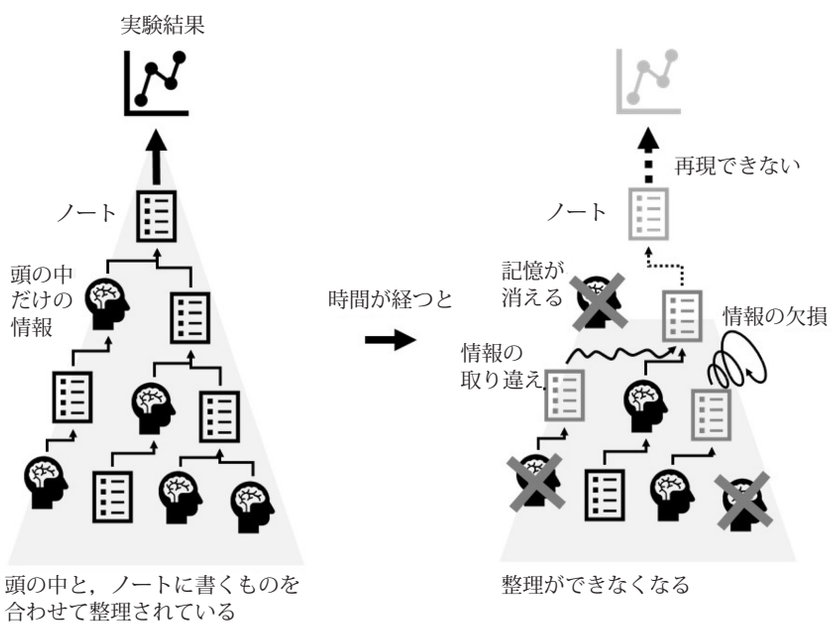

図 1　属人的なデータ管理とその特性

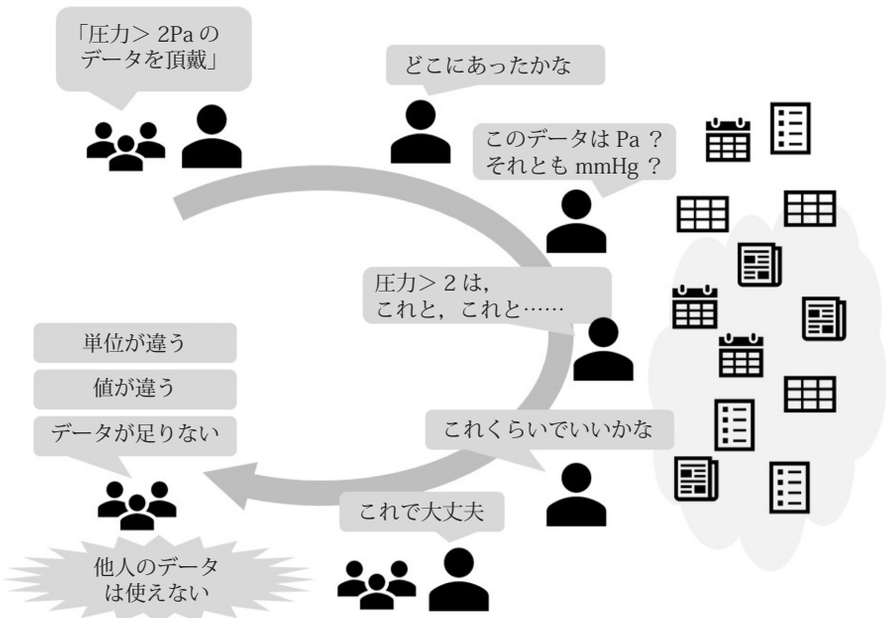

図 2　属人的に管理されたデータが引き起こす問題

「属人的なデータ管理」状況が生み出される直接的原因は，そもそも第 3 者が再現するのに十分な情報が記録されていないことと，記録されているものに関しても，それが何を示す値なのか，つまり，その値の項目名に関して，人によって，また，同じ人でも時期によって，その項目名の同一性が保たれていないことである。これらの解決方法は，研究開発リーダーのバックナンバー及び書籍「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」の私の執筆部分に詳しく記載しているので，そちらを参照してほしい。本連載では，実際にデータを蓄積していく段階で，どのように項目名を決め，どのようにデータを蓄積していき，どのようにデータを利活用するのか？つまり，どのようにデータ分析をしていくかを具体的な例とともに説明していく。

3　記録するデータの項目名は，意味が分かれば何でも良いという訳ではない

前章では，「R & D 部門におけるデータ共有，利活用の実情」に関して，説明を行った。本章では，「記録するデータの項目名は，意味が分かれば何でも良いという訳ではない」と題し，記録するデータの項目名は，どのように決めるべきかを論じる。

項目名決定において重要な点は，データ分析を前提にし，データ抽出 / 検索，分析を行いやすい項目名にしておかないといけないということである。データ検索，分析を行いにくい項目名だと，結局，データを探したり，分析したりするのに大きな手間がかかるため，データ共有を大きく阻害するのである。R & D 部門以外の他の業務系の場合は，項目名決定時にデータ分析のことはあまり考えず，どのようにデータを入力，登録するかを考え，項目名を決定する。つまり，データ入力のしやすさを前提にして，項目名を決定するのである。R & D 部門以外の他の業務系の場合は，データ検索，分析が複雑ではないので，データ入力のしやすさを前提にして，項目名を決定しても大きな問題にはならないのである。一方，R & D 部門では，項目が多く，項目の追加や変更も頻繁にあるなかで，複雑なデータ検索，分析が必要となるので，データ抽出 / 検索，分析を行いやすい項目名にしないとデータ抽出 / 検索，分析自体が困難になる。以下，R & D 部門でデータを記録していくための項目名を決めるときに注意すべき点を列挙しておく。

3.1　すべてのデータは，項目名－項目値という単純構造に落とし込む

これは，「データを分析するときには，2 次元の表形式データになっている必要がある」ことからくる条件である。Excel などで，グラフを描いてデータ分析をするときには，必ず項目名－項目値の形になっていることは理解できると思う。しかし，これは結構守られていないのである。例えば，実験ノートや実験データを記録する Excel では，表 1 のようになっていることが多いと思う。

表 1

乾燥工程				焼結工程	
1 段目		2 段目			
温度	風速	温度	風速	温度	風速

実は，この状態は項目名－項目値の単純構造にはなっていないのである。「純粋な項目名は温度と風速だけで，乾燥工程，焼結工程や 1 段目，2 段目などは，分類的なもので項目名でない」と捉えていてはだめなのである。もし，そう考えてしまうと温度＝ 120 という項目名－項目値を見たときに，それが乾燥工程の温度なのか？焼結工程の温度なのか区別がつかない。項目名は重複がなくユニークな命名でないと，データ分析時にそれが何の値なのかが厳密に分からず，困ってしまうのである。つまり，乾燥工程，焼結工程や 1 段目，2 段目などの分類的な情報を含めて項目名にしなければならないのである。Excel でセル結合を使っているデータ記録は，この段階で失格ということになる。結局，先ほどのデータを項目名－項目値の単純構造にすると，表 2 のようになる。

つまり，Excel で言うと 1 行目の 1 セルずつに項目名が列挙され，2 列目以降は項目値だけが並び，それ以外の分類情報などは一切無い形が，項目名－項目値の単純構造の項目ということである。「Excel でセル結合を使っていても，ピボッドテーブルやマクロを使えば，データ分析はできるのでは？」という方もいらっしゃると思う。もちろん，それは間違っていない。ただし，R & D 部門で取り扱う様々な実験すべてに対応するようにできるかというと非常に悩ましいはずである。ピボッドテーブルやマクロは，ある一定の決まりきったことを何度も行う場合は，便利なのだが，そもそもそれを作るのが大変なのである。経理や営業などの業務系で，いつも同じ項目データで同じ処理をするのであれば，ピボッドテーブルやマクロは便利なのだが，扱う材料や処理プロセスがどんどん変わっていく R & D 部門では，ピボッドテーブルやマクロを作成し，メンテナンスすることの負担が大きく，現実的には運用できなくなってしまうのである。実際，使われなくなったマクロや GUI が施された Excel が乱立し，収拾がつかなくなってしまった経験のある人も多いのではないかと思う。R & D 部門のデータ記録は，単純で「それでいいのか？」と思ってしまうが，項目名－項目値の単純構造でなければならないのである。「Simple is best.」である。

3.2　項目名－項目値の項目間に論理的関係があってはいけない

「項目名－項目値の項目間に論理的関係があってはいけない」というタイトルにしているが，これでは何を言っているのかピンとこない人も多いと思う。以下で，例をあげて，説明をしていく。

表 2

1 段乾燥温度	1 段乾燥風速	2 段乾燥温度	2 段乾燥風速	焼結温度	焼結風速

表 3

実験 ID	原材料名 1	原材料濃度 1	原材料名 2	原材料濃度 2	引張強度
EXP1	エチレン	80	プロピレン	20	10
EXP2	ブタン	75	エチレン	25	12
EXP3	プロピレン	60	ブタン	40	8

表 3 は，1 行目の 1 セルごとに項目名が列挙され，それ以外の分類情報などは一切無い形なので，項目名－項目値の単純構造にはなっている。実は，原材料名 1 と原材料濃度 1，原材料名 2 と原材料濃度 2 は，2 つの項目がペアになっていることがわかると思う。これが「項目間に論理的関係がある」ということなのだ。焼結温度，焼結風速も項目間に関係はありそうだが，それは性能の高い目的物を作る場合の温度と風速の関係であったり，装置の設計上の制限からくる関係であったりで，論理的な関係性ではないのである。論理的関係性とは，原材料名 x と原材料濃度 x のときに「x が同じモノ同士がペアだ」というような項目名に対して決められた人為的なルールのことである。また，上記には 1，2 を入れ替えても同じ実験になるという対称性もこの項目の論理的関係性として埋め込まれている。ペアや対称性といった項目自体に内在させられた関係性は，物理（自然科学）とは関係のない人為的な，項目名命名による関係性である。そして，データを絞り込んだり，分析をする場合には，この項目の論理的関係からくる論理制約（ペアや対称性）を排除する必要があり，非常に面倒な作業が必要になってくる。例えば，エチレンを使っていない材料を探すにしても上記項目の論理的関係を考慮して，「原材料名 1 にエチレンがなく，かつ，原材料名 2 にエチレンがない，および原材料名 1 にエチレンがある場合は，原材料濃度 1 が 0，および原材料名 2 にエチレンがある場合は，原材料濃度 2 が 0」という条件で絞り込む必要がある。「項目間の論理的関係」が無いように項目名を定義しておけば，こういう面倒さはなくなる。また，「引張強度のプロパン濃度依存性」を確認しようとしても，この表形式のままでは X-Y プロットグラフを描くことはできない。

表 4 は，上表から「項目間の論理的関係」を排除した項目名である。

この場合は，エチレンを使っていない材料を探す場合，「エチレン濃度が 0」という条件で絞り込むだけでよくなり，「エチレンを使っていない」をそのままストレートに条件にすれば良くなる。また，プロパン濃度の列を X 軸に設定し，引張強度の列を Y 軸に設定するだけで，「引張強度のプロパン濃度依存性」の X-Y プロットグラフも簡単に描くことができる。実は，ペアや対称性といった項目名に内在させられた関係性を持つ項目というのは，本来 1 つの項目であるべきものを 2 つの項目に分けてしまったから発生したものなのである。

それでは，原材料名 1 と原材料濃度 1 という項目名は，データ分析が行い難いにもかかわらず，なぜ，よく使われているのだろうか？上記例では，データ記録のための列数は「原材料名 1 と原材料濃度 1」型の方が多くなるが，原材料種類が 100 種類になれば，「エチレン濃度」型は，100 列にもなってしまい，100 列の中で実験毎に使われるのは 2 列だけで，その 2 列を 100 列から探さなければならなく，非常に入力しにくい表形式になってしまう。つまり，研究者はそういうことを先取りして，入力のしやすい形式を選んでいるのである。しかし，この入力しやすい形式が分析をし難くしてしまっているのである。すべてのケースがとは言わないが，入力のしやすさとデータを探したり，分析をするしやすさは，相反関係になる。

1 つの Excel ファイルでデータを入力するシートとデータ絞り込み，分析するためのシートを分ければ入力と分析のしやすさの両立は可能である。しかしながら，入力シートと分析シートは参照式や結構面倒なマクロを書

表 4

実験 ID	エチレン濃度	プロピレン濃度	ブタン濃度	引張強度
EXP1	80	80	0	10
EXP2	25	0	75	12
EXP3	0	60	40	8

く必要があり，どの項目とどの項目がペアだったり，対称性があったりを自動判断させるためには，項目名の命名規則を相当厳密に決める必要もある。項目数が膨大で，かつ項目の追加，変更の多い R & D では，このような参照式やマクロ，項目名ルールを維持し続けることは，ほぼ不可能である。結局は，R & D ではデータ記録のための入力のしやすさは犠牲にするしかないという結論になる。実際は，入力し難いといったところで，数割程度入力にかかる時間が増えるだけである。その数割の手間を惜しんでしまうことで，データを絞り込み，分析をする毎に入力時よりはるかに多くの手間が発生することになる。当然，入力時の手間を許容できないようであれば，データを絞り込み，分析をする時の手間も許容できようはずがなく，結局，データは使われなくなるのである。

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 ー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，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