「株式会社技術情報協会 月刊 研究開発リーダー 2024年7月号」掲載記事
URL:https://www.gijutu.co.jp/index.htm

◆新連載◆

《隔月連載 全5回》

第1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

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年間行っていた。その中でR & Dデータが属人的に処理され，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18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探し出せ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は，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が，それ以外の部分は，上記状況と変わらない。これらの状況は，「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」ができていない状態ということになる。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さえ，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わざるを得なくなることもある。共有した結果に間違いがあったり，再実験や再解析を行う羽目になった結果，データ提供者の信頼が失われ，データ共有の意欲がなくなっていき，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。

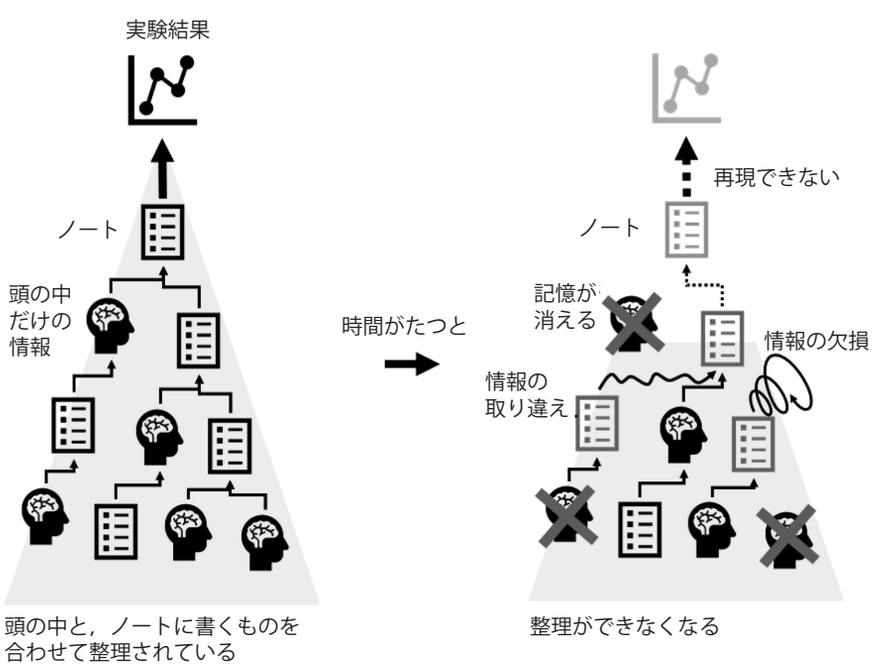

図 1　属人的なデータ管理とその特性

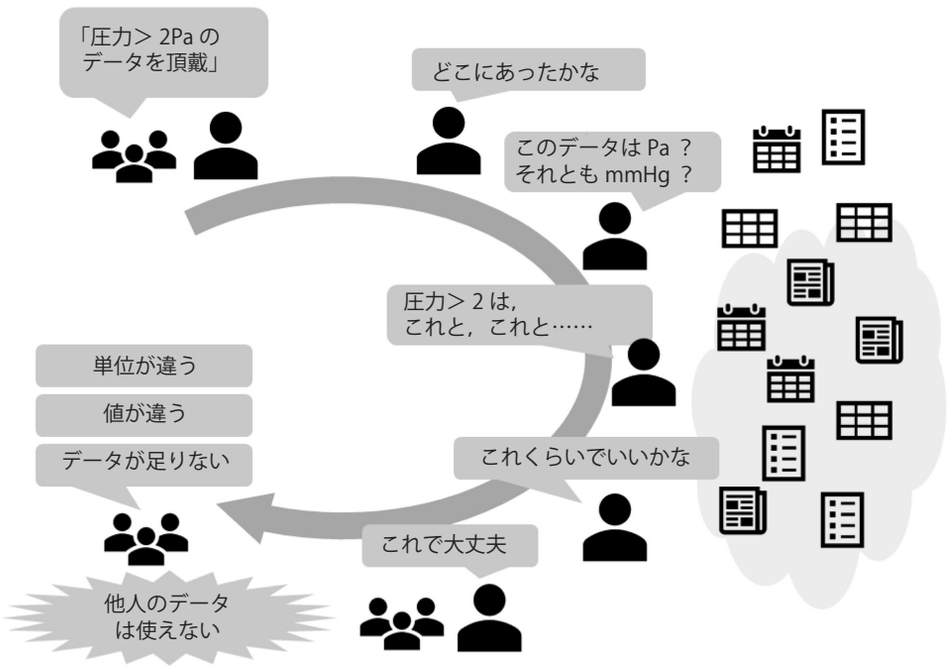

図 2　属人的に管理されたデータが引き起こす問題

「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。

3　報告書の共有で期待して良いこと，良くないこと

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，「報告書の共有で期待して良いこと，良くないこと」と題し，データ共有と報告書共有の違いに関して論じる。データ共有は「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」ことと説明した。それでは，報告書に「実験及び解析を第 3 者が再現するために必要な情報」が書かれていれば，報告書を共有するだけで，データ共有にもなるのではないか？と思うかもしれない。その考えは，いくつかの点で間違っている。

第一に，論文や報告書に「実験及び解析を第 3 者が再現するために必要な情報」が書かれていることは非常に稀であり，実際，論文や報告書だけでは再現実験，解析ができない場合がほとんどである。これは，論文や報告書は，どのような実験を行えばどうなるというミクロな観点ではなく，複数の実験から何かしらの一般的な知見や法則性を導き出し，それを提示することが目的であり，その目的に適った論述形式になるのが原因である。説明的論述形式の中に大量になるであろう「実験及び解析を第 3 者が再現するために必要な情報」を埋め込むと整った文章にならないので，通常そういうことは行われない。もちろん，Appendix 等に報告書で参照した実験パラメータと計測値一覧のようなものを列挙するようしておけば，そういう文章のまとまりの悪さもなくなるが，そういう報告書はあまりお目にかからない。

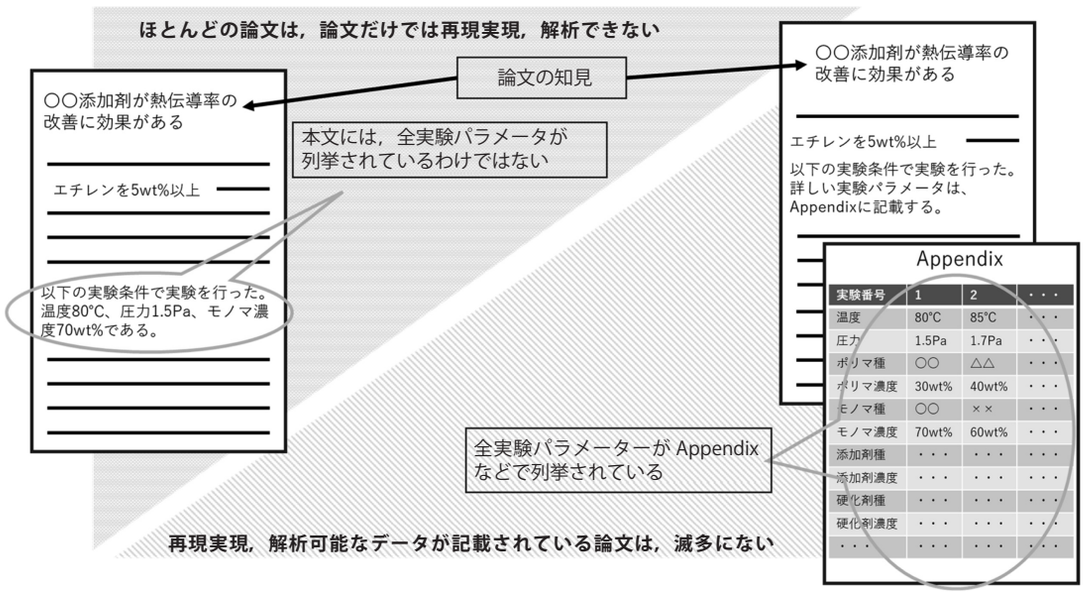

図 3　報告書には実験や解析を再現するために必要な情報が書かれていない

百歩譲って，報告書で参照した実験パラメータと計測値一覧がしっかり書き込まれた報告書があったとする。しかし，それでも「必要な時に迅速かつ確実に探し出せる状態に保っておく」ということは，実現できないのである。「報告書を全文検索で検索すればできるじゃないか？」と思われる方もいるかもしれないが，それは全文検索というものを過大評価し過ぎである。例えば，「エチレン」を使った実験のことが書かれている報告書を探したいとする。全文検索で，「“エチレン”という文字列を含む」という条件で，検索すればいいのではないかと思うかもしれない。しかし，その検索条件では「本実験では，エチレンは使用していない」という文章の入っている報告書も hit してしまうし，「ethylene を 10wt％混合した」という文章の入っている報告書は hit しないのである。もちろん，「エチレンを 5wt％以上使った実験」を探すことは，当然のことながらできない。報告書を共有しただけでは，「必要な時に迅速かつ確実に探し出せる状態に保っておく」とは言えないということはわかっていただけただろうか？

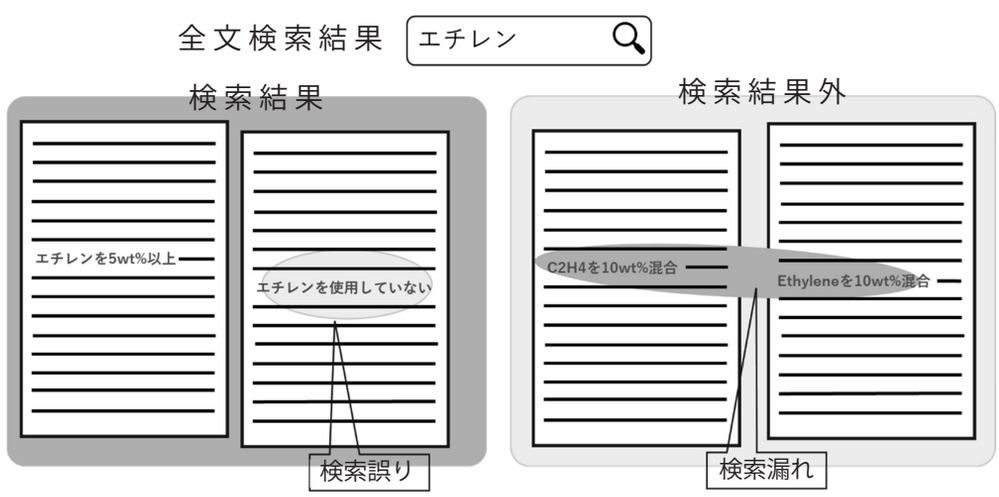

図 4　全文検索では検索誤りや検索漏れを防ぎきれない

少し余談だが，「“エチレン”という文字列を含む」という条件で，「ethylene を 10wt％混合した」という文章を含んだ報告書を hit させることはできるようになってきている。それは，去年ぐらいから世間を賑わせている ChatGPT などの生成 AI である。生成 AI では学習用の多量の文章から“エチレン”と“ethylene”が同じ物質であることを判定することが可能である。しかし，これは“エチレン”と“ethylene”を大量に含む学習用文章があるからこそのなせる業であり，R & D 部門で使われる多くの単語，例えば，購入品の材料商品名や商品型番のようなものは，学習用文章自体がインターネットに転がっているものではない。社内文章だけを学習用文章として考えると，単語数に比して十分な文章量はないから R & D 部門で使われる多くの単語の等価関係は ChatGPT などの生成 AI であっても無力である。R & D 部門という特殊かつ閉鎖的な環境では，実は，ChatGPT などの生成 AI だけでなく，機械学習のような多量の学習データを必要とする技術が有効な手立てにならないことが多いのである。

報告書の共有がデータ共有の代わりにはなりえないことは，ここまでの話で理解できたと思う。それであれば，報告書の共有に全く意味がないのかというと，当然そんなことはない。前述した通り，論文や報告書は，複数の実験から何かしらの一般的な知見や法則性を導き出し，それを提示することが目的であり，一般的な知見や法則性という強力な実験指針を得られるので，非常に価値の高いものである。例えば，ニュートンの運動方程式やシュレディンガー方程式を使うと，世の中の多くの現象を説明，予測できることを考えれば，一般的な知見や法則性がいかに強力で価値が高いかはわかっていただけると思う。ただ，論文や報告書を参考にする場合には，注意すべき点もある。例えば，「○○添加剤は，熱伝導率の改善に効果がある」と報告書に書かれていても，実際，あらゆるケースに対して，「○○添加剤が熱伝導率の改善に効果がある」ことを確認したわけではないはずで，特定の実験パラメータ領域での知見のはずであり，今，自分が適用しようとしている領域でその知見，法則が確認されているとは限らない。報告書のタチが悪いのは，どの領域ならその知見が成立するかを証拠データとともに明示的に書かれていなかったり，拡大解釈（実際に確認をしていないがそう信じている）して書かれていることが多かったりする点である。これは，社内報告書だけでなく，学術論文でもその傾向がある。また，これは論文の査読レフリーでも十分に指摘しきれていない部分であり，学術論文でもその点を十分注意して，参考にする必要がある。

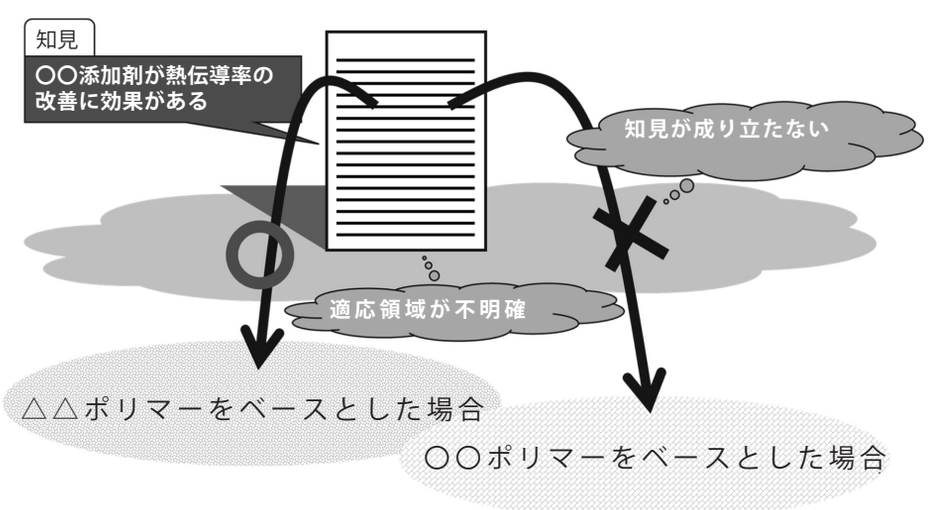

図 5　報告書に書かれた知見や法則性は適用領域が不明確なものが多い

実は，ニュートンの運動方程式やシュレディンガー方程式も，どの領域ならその法則が有効かをデータとともに明示的に記して提案されたわけでなく，結構広い領域で成り立っているであろうという信念のもとに提示されている。どちらかというと，それを読んだ人がその法則に懐疑的で，反例を探そうと様々な実験をし，反例が見つからないことで，その法則が成立する領域が明らかになっていったという流れになっている。そして，長い年月と持続的な反例探しに耐えることで，蓋然性の高い法則になっていっているのである。こういうスタンスで，論文や報告書を参照するのであればいいのだが，社内共有の観点では，報告書に懐疑的で反例を探そうと思い報告書を読む人は稀で，どちらかというと書かれていることは，すでに蓋然性の高い法則，知見だと思って参照してしまう人が多い。それはかえって研究を遅延させてしまいかねないものなので，そういう点をよく理解した上で，報告書は活用すべきであることをここで指摘しておく。

最後に少しだけ個人的認識を書いておく。一番安心な報告書の活用方法は，報告書内に書かれた知見，法則を確定的なものと考えず，研究の新たな観点，閃き，アイデアを得る源泉として利用することではないだろうか。

4 データ共有で研究の何が改善できるのか？

前章では，「報告書の共有で期待して良いこと，良くないこと」と題し，データ共有と報告書共有の質的な違いに関して，説明を行った。本章では，「データ共有で研究の何が改善できるのか？」と題し，そもそもデータ共有で，研究の「どのようなことが改善できるのか？」，逆に言うと「どのようなことの改善は期待してはいけないのか？」を論じる。

「データ共有で研究の何が改善できるのか？」を論じる前に，まず，「研究」というもの自身がどういう要素から成っているかを考えてみる。研究は，閃きや発想という独創的な力とデータを分析し，傾向を把握する力という 2 つの力から成っている。閃きや発想は，属人的，主観的なもので，それが研究者の個性的能力ともいえる。一方，データを分析し，傾向を把握する力は，一見，閃きや発想と同じように属人的，主観的なもののように感じるが，よく考えてみると数学や統計などの客観的方法で確立されるべきものである。実際，同じデータを使って，分析し，傾向を把握したとすると分析結果の解釈以外は，同じ結果になるべきものである。また，分析結果の解釈も単なる閃きや発想と比較すると属人的，主観的というよりも結果解釈時に何に重きを置くのか？何を無視するのか？という取捨選択のパターンという客観的なものとして考えることも可能である。つまり，データを分析し，傾向を把握する力は，主観的でなく，客観的なものだということである。

前述のことを念頭に置きながら，「研究とは何か」を少し考えてみる。大きく想像を膨らませてほしいのだが，大航海時代に冒険家が世界の海を渡り，次々に新発見をしていったことと，研究者が研究で新しい発見をしていく過程は近いものがあるのではないだろうか？そう考え，研究を航海に例えてみると，航海士の経験，勘は，研究者の閃きや発想に対応し，航海での海図と羅針盤とその技法が研究でのデータとデータ抽出，分析手法に対応するのではなかろうか？どれだけ優秀な航海士もでたらめの海図と羅針盤とその技法では，まともな航海はできないように，研究もデータとデータ抽出，分析手法がでたらめでは，まともな研究はできない。実際，海図と羅針盤とその技法が航海士で共通のものであるべきであるように，データとデータ抽出，分析手法も研究者で共通であるべきで，それによって研究の客観性を担保するのである。優秀な航海士は，その客観的情報（海図とと羅針盤とその技法）に，自身の主観的な経験，勘を組み合わせることで，他の航海士より，危険な海をより安全に航海できる。当然研究者も客観的情報（データとデータ抽出，分析手法）に，自身の主観的な閃きや発想を組み合わせることで，他の研究者ではできない発見が可能なわけである。「研究と航海」，非常によく似た対応関係になっているように見えないだろうか。

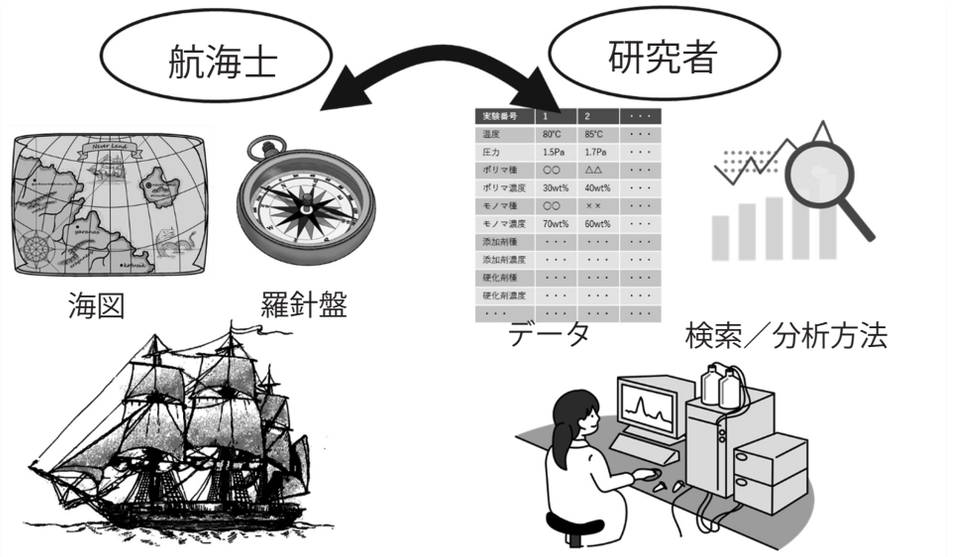

図 6　研究者のデータと検索／分析手法は，航海士の海図と羅針盤と同じ

本題に戻って，「データ共有で研究の何が改善できるのか？」について，考察を進めよう。研究は，閃きや発想という独創的な力とデータを分析し，傾向を把握する力という 2 つの力から成っているので，「データ共有で研究の何が改善できるのか？」は，「データ共有で閃きや発想が改善できるのか？」と「データ共有でデータ分析や傾向把握することが改善できるのか？」に言い換えることができる。「閃きや発想」は，属人的，主観的なものなので，改善できるとしても研究者ごとにその改善内容や改善度合いは異なってしまう。また，「閃きや発想」というものは，定量的に評価しにくいものであり，改善されたかどうかを判定すること自体が非常に難しい。一方，「データ分析，傾向把握」は，根本的には数学的，統計的データ処理やグラフ化などの客観的方法で実施されるべきもので，主観ではなく対象データ自体によって決定づけられるものである。つまり，数学的，統計的データ処理やグラフ化の結果解釈以外，すなわち，数学的，統計的データ処理やグラフ化自体は主観的でなく，客観的なものである。

原理的には，データ共有で「閃きや発想」も「データ分析や傾向把握」も改善は可能ではあるが，主観が絡み合った部分は，改善指針や改善評価自体が研究者個人個人に合わせこむ必要があるので，負担が大きく，取り組みにかかる負担が改善による効用を上回ってしまう可能性が高く，最初に取り組むべき課題ではない。したがって，改善指針や改善評価が人によらない数学的，統計的データ処理やグラフ化部分をデータ共有での最初の改善対象と考え，その改善に道筋ができたあとで，数学的，統計的データ処理やグラフ化の結果解釈部分をデータ共有での改善対象として，取り組むべきである。

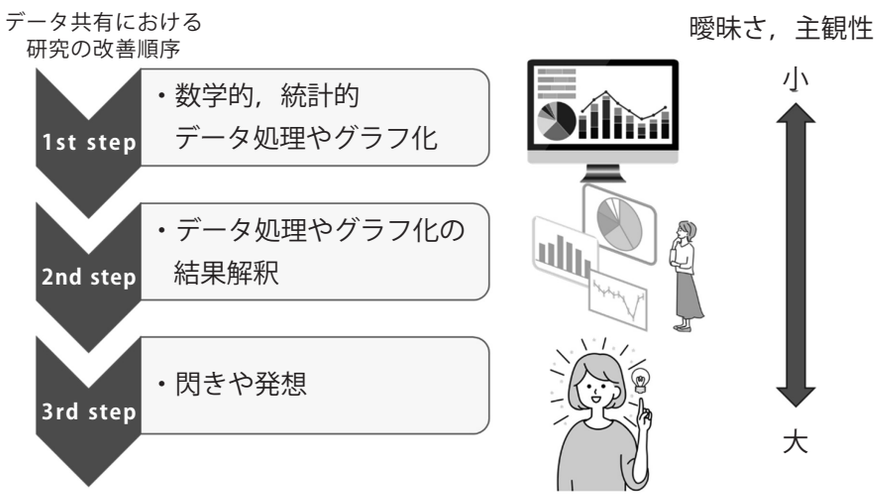

図 7　データ共有における研究の改善順序

「閃きや発想」は，直接的な改善を期待するのではなく，上記の改善で「閃きや発想」にかけられる時間的余裕ができ，新しい「閃きや発想」が生まれる機会が増えると考えておくのが現実的である。もちろん，直接的な改善は絶対ないということではない。もしそういうことがあれば，共有したデータをどのように扱えば「閃きや発想」が改善されるのかを客観的，統計的に検証し，他の人でも「閃きや発想」の改善が確実にできるような方法を確立していけばよい。実際，こういう客観的，統計的検証を経るまでは，「閃きや発想が改善している」ということ自体が主観的な思い込みの可能性が高いので，信用してはいけない。

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