第1節 R&D 部門における自動・自律実験運用のデータプラットフォーム構築と その運用・推進体制

(株)キャトルアイ・サイエンス 上島豊

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、偶発的なデータ取り違え、主観的データ操作など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 年間行なっていた。その中で R&D データが属人的に処理され、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。

本節では、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R&D 部門におけるデータ共有、利活用の改善方法、さらに一歩踏み込んだ「自動・自律実験運用のデータプラットフォーム構築とその運用・推進体制」に関して、簡単に解説する。

1. R&D 部門におけるデータ共有・利活用の実情

自動・自律実験ができたとしても、「データ共有・利活用」ができていない状況であれば、本末転倒である。したがって、まずは、自動・自律実験以前の「データ共有・利活用」から説明をして行く。

皆さんは、「データ共有」ができているとは、どういう状態を意味しているのか正確に説明できるだろうか?「データ共有」とは、読んで字のごとくデータを共有することなのだが、より具体的に言うと、データを生み出した実験及び分析を第三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の共有は「共有」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 部門を見てきた経験をもとに、R&D の現場では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分析では、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されているだけの場合が多い。実験や分析結果の比較評価がある程度難しい課題に対しては、実験や分析の情報が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分析の場合も上記状況と変わらない部分もあるが、他の人への実験や分析の引き渡し(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。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、Excel 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たりしているところもある。これらは、「データを生み出した実験及び分析を第三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、「データ共有」ができていないということになる。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る。この状態を「属人的なデータ共有」と呼ぶことにする。

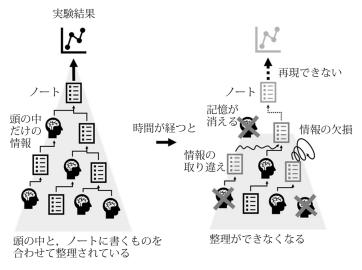


図1 属人的なデータ管理とその特性

データ共有の意識が高い組織では、しっかりしたデータベースを導入し、実験データを蓄積、検索できるようになっているところや自動・自律実験ができるようになっているところも稀ではあるが、存在はする。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。データベースがあり、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「データ共有」ができていないということになる。この状態を「形骸化したデータ共有」と呼ぶことにする。

また、自動・自律実験ができる状況になれば、「属人的なデータ共有」状況は大きく改善されることが多いが、「形骸化したデータ共有」状況は、ほとんど改善されないことが多いことを指摘しておく。

2. 属人的なデータ共有,形骸化したデータ共有状況から生まれる問題点

前項で、説明した通り、ほとんどの R&D 部門では、「データを生み出した実験及び分析を第三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共有」状況になっている。本項では、「属人的な、もしくは形骸化したデータ共有」状況がどのような問題を生み出すかについて論じる。

「属人的なデータ共有」状況では、実験及び分析の詳細なことは、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分析を第三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だが、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、「〇の実験・分析情報、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分析を進めることが発生し、間違った結論が導かれ、検討を進めたのちに始まりに立ち返って、再実験や再分析を行なわざるを得なくなることもある。

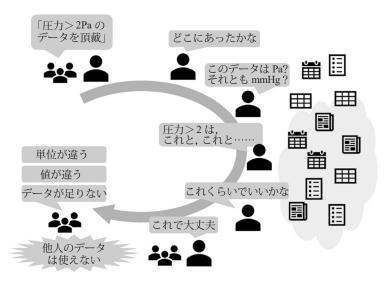


図2 属人的に管理されたデータが引き起こす問題

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、他の研究者のデータを参照せず、自分が実施した実験や分析のみしか参照しないような状態が進行していく。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、継承困難化を進めてしまう。

「形骸化したデータ共有」状況は、利用、活用が形骸化するという問題だけでなく、利用、活用されないこと自体も問題である。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る。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る。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だろうか?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だが、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ある。実際に、このような「実験及び分析を第三者が再現するために必要な情報が記録されていない」システムが多いのである。システムが利用、活用されていないので、蓄積されたデータに問題があることなどの発覚が遅れ、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ある。したがって、「実験及び分析を第三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、速やかに利用、活用を促す体制を見直すべきである。

3. 属人的なデータ共有、形骸化したデータ共有状況が生み出される原因

前項では、「属人的な、もしくは形骸化したデータ共有」状況が、どのような問題を生み出すかについて説明した。 本項では、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共有」状況が生み出 されるのか?、その原因に関して考察する。

「属人的なデータ共有」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ある。R&D部門の業務は、新しい実験及び分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る。したがって、「属人的なデータ共有」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る。「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るが、「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る。

実験及び分析を行なう場合、実験及び分析の全てのパラメータを変化させることは、まずない。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、実験及び分析を行なうはずである。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ある。当然、実験及び分析の最中や直後は、記録がなくとも実施者自身

であれば実験及び分析の再現は可能である。目の前の実験及び分析の結果に意識を集中している場合、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、記録することを省略してしまうのである。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、「実験及び分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うのである。また、そもそも実験及び分析の最中の研究者は、実験及び分析パラメータを忘れるわけがないという意識があり、「実験及び分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る。

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る。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、水泳を覚えたりす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ない。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ない。そして、研究においても当の本人は、共有することのメリットを感じにくいのが普通である。結局、この属人的でパターンマッチング的な認識能力が機能していることと、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、「実験及び分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ある。手動実験から自動・自律実験に切り替えていくことで、「実験及び分析を再現できるだけの情報記録」はまさしく、自動的になされるようになるので、大きく改善が見込まれる。ただし、後述するが、どんな情報を記録すれば「実験及び分析を再現できるだけの情報記録」になるのかは、自動では判断できなく人が判断する必要がある。

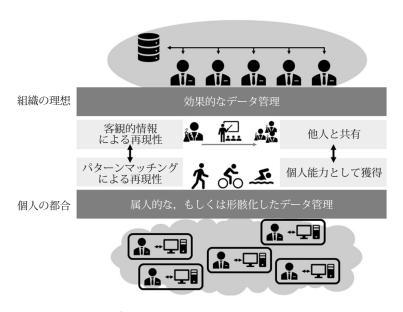


図3 データ管理に関する組織と個人の乖離の原因

「形骸化したデータ共有」状況が生まれる原因は、「属人的なデータ共有」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」ということと根本的な部分では同じではあるが、ここではもう少し掘り下げて、考察しておく。「形骸化」=利用、活用されていない直接的原因は、大きく2つある。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」、「探したデータを分析したいと思ったときに、どのようにデータ分析をすればいいのかわからない」といったもので、そもそもデータを探せるように、分析できるようにシステムが作られていないということである。これは、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ある。しかし、「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ない。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ある。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。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う。自動・自律実験ができる状況になっても「形骸化したデータ共有」状況は、ほとんど改善されないことが多いというのは、まさしくこれが理由である。自動・自律実験ができることで、データ探査や分析も自動・自律化できるのであれば、もちろんこういう問題は起こらない。そもそもデータ探査や分析が自動化されるのであれば、研究者は不要ということになってしまう。もちろん、遠い未来には研究者が一人もいない自動・自律研究所も生まれるかもしれないが、現時点のAIの能力ではある程度決められた範囲でのデータ探査や分析は自動・自律化される可能性はあるが、それは研究の深部になればなるほど、自動・自律化されないことは自明である。

もう一つは、蓄積されたデータが不完全で、「実験及び分析を再現できるだけの情報記録」となっていないことである。このような状態では、前項でも述べた「間違った結論が導かれ、検討を進めたのちに始まりに立ち返って、再実験及び再分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。研究者同士の信頼関係が崩れないだけまだましだが、データ共有という側面では実質的には失敗である。なぜ、「実験及び分析を再現できるだけの情報記録」とならないかは、「属人的なデータ共有」状況が生み出される原因と同じ理由である。また、それを未然に検知できなかったのは「実験及び分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ある。前述しているが、どんな情報を記録すれば実験及び分析を再現できるだけの情報記録」になるのかは、自動では判断できなく人が判断する必要があることをここでもう一度、指摘しておく。

4. 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

前項では、「属人的な、もしくは形骸化したデータ共有」状況が生み出される原因について説明した。本項では、自動・ 自律実験データプラットフォームを導入し、属人的共有状況を脱するために必要な事前準備の前に行うべきことにど のようなものがあるのかを説明する。

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、改ざんされたりする。
- -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捉えないなら、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ったりすることが発生しうる。

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。ちなみに、自動・自律実験ではデータは自動的に記録されるので、この部分は問題ないと思うかもしれないが、それは違う。まず、第一に実験、分析のすべての工程を自動・自律実験化できることは少ない、つまり手動部が残ってしまうという点がある。次に自動・自律実験でも意識的に記録をするようにしないと記録されない情報がたくさんあり、実験や結果分析の再現に十分な情報の特定、合意は、自動・自律実験でも意識的に

行わなければならないものなのである。

- 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。また、グラフ作成時のX軸,Y軸,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。お分かりだと思うが、この点は、自動・自律実験をデータ分析のところまで拡張しようと、しまいと検討、合意が必要な部分であり、見逃しがちな部分なので注意が必要である。

これらは、データベース化、システム化や自動・自律実験という IT 観点で必要な事柄ではなく、もっと一般的な意味での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、自動・自律実験の取り組みを行うことが肝要である。

5. 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!

自動・自律実験のためには、実験のデータベース化、システム化が必須となる。以下でデータベース化、システム化のため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、業務系システムは、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。

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む込むなら、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ことになる。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。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、ツールかシステム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) Excel、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、実験実施、データ記録、データ処理の自動化とデータ探査及び、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、どのようなデータ記録、処理、探査をするのかはオルゴールのドラムである。どのような音(記録、処理と探査)を奏でるかは、オルゴールを作る前に詳細に決め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、システムがない状態で、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、手順書をシステムに組み込む作業に対応するのである。

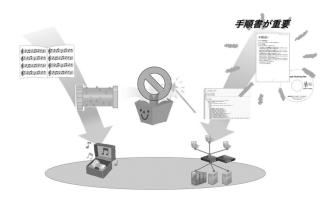


図4 データベースやシステムは魔法の箱ではなくオルゴール!

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、自動・自律実験の最初の一歩ということである。

6. 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・利活用や自動・自律実験をするためには、データベース化、システム化すべ きなのかを考えておく。データ共有・利活用や自動・自律実験をするためには、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記録、処理し、探査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ばその通りである。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、システム化の最大の目的は、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 とが起こるかを思い起こしてほしい。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○○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。

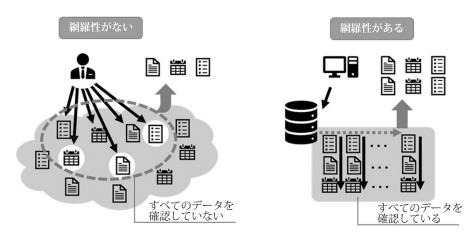


図5 データベースの利点の網羅性とは

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。

例)耐電圧が500V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。

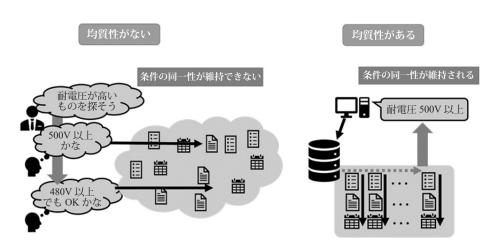


図6 データベースの利点の均質性とは

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、システムであれば、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。

先ほど、データベース化、システム化において、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ると説明したが、システム化を自動・自律実験まで進めると作業者負担を大きく低減した状態で24時間、休日も実験可能になるため、大きなメリットがあると言える。また、実験装置、計測装置の台数を増やせば、作業者負担を増やさなくとも実験実施数を増やすことが可能になる。自動・自律実験は作業者が実験をするのに比べ、実験自体の再現性は高くなり、データ記録の誤記や未記録もほぼ完全に防ぐことができることも注目すべき特徴である。

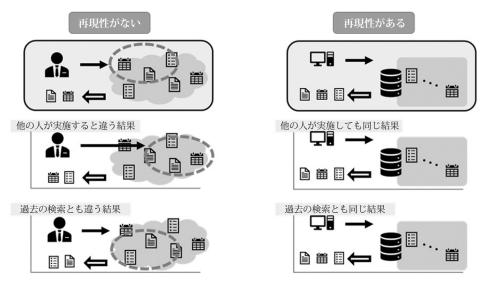


図7 データベースの利点の再現性とは

7. 自動・自律実験に先立つ属人的データ共有状況を脱するための事前準備(手順書作成編)

話が脱線しすぎたので、話を戻す。データ共有のために必要な準備(実験や分析を第三者が再現するのに十分な情報及び利用、活用観点を考慮した項目のリストアップと合意形成)が終わった後、データベース化、システム化、自動・自律実験のために直接的に必要な準備を行う。「システムはオルゴールのようなものだ」という話を思い出してもらうと、何をしないといけないかは自ずとわかってくる。そう!楽譜が必要なのである。つまり、「実験自体及びデータ記録とデータ探査とデータ処理を誰が読んでも実施できるレベルの手順書」を作ることが最初の課題である。

「誰が読んでも実施できるレベルの手順書」を作ることは、簡単なようで、なかなか難しい。実験自体やデータ探査とデータ処理自体は、データベースやシステムがない時からすでに行っているはずである。ただ、それを手順書としてまとめている人は、研究者の中では非常に稀な存在で、ほとんどの研究者は経験という記憶に頼って行っているはずである。経験を的確な文書に起こすことが難しいことは、どのようにして自転車の補助輪を外せるようになったか?、どのようにして平泳ぎができるようになったか?、どのようにして日本語を話せるようになったか?を考えてみればわかるはずである。ただし、実験自体やデータ探査とデータ処理に関しては、これができない限り、システム化はできない。実際、補助輪外しや水泳や日本語習得に比べると格段に簡単な課題なので、覚悟を決めて取り組み、慣れてしまえば、この程度の経験を手順に落とすのは簡単である。研究の中核的な思い付き(この材料を配合してみれば良さそうだ、ここの攪拌は泡立てないように注意しながら)は、どのようにして思いつくかを手順に落とすことは、非常に難しい(とはいえ、本当はグループリーダからすればそれをしたいのだと思うが…)はずで、それを要求されているわけではないので、実験自体やデータ探査とデータ処理の手順化ぐらいは頑張ってほしい。

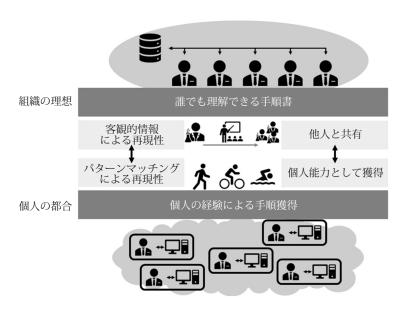


図8 他の人でも実施できる手順書作成の難しさ

手順書を書いていくといくつかの課題にぶち当たるはずである。一つ目は、手順が確定していないことが発覚するということである。「実験や分析を第三者が再現するのに十分な情報及び利用、活用観点を考慮した項目のリストアップと合意形成」は終わっているが、実際にまだデータ共有・利活用を行っていない場合、実験を記録したファイルの保存場所がまだ、決められていない(属人的になっていて人によってバラバラ)などの問題が浮上していない場合がある。当然、このような課題が出てくるたびに合意形成をし、改善していくことが必要である。想像だけで大丈夫と思っても、案外抜けがたくさんあるもので、合意形成が終わった後で、実際にそれに沿って、データ共有・利活用を実践しておくことが重要ということである。「これをシステムで改善したいんだ」という話をよく聞くが、実験データファイルを場当たり的な場所に置かれるとそれを見つけてシステムに登録することはできない。つまり、システムがない段階で、一定のルールに従って、実験データファイルを置けないのであれば、システム化を行ってもデータは登録さ

れないだけである。「これをシステムで改善したいんだ」と思った場合、「システムは魔法の箱ではない」ということを思い出して、手順書作成を進めてほしい。

二つ目は、「多種多様なすべてのデータ探査と処理の手順書なんて、書けるわけがない」というものである。これ 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。R&D では、経理や財務のように完全に決まりきっ たデータ探査と処理はほとんどなく、それを逐一挙げていくのは、現実的ではない。また、今日、新しいデータ探査 と処理が発生するかもしれない。しかし、システム化をするときには、それでは困るのである。 どうデータ探査をして、 どうデータ処理をするのか?詳細な情報なしでそれができるような魔法の箱はないのである。「だって、Excel やグラ フソフトや実験計測装置についているソフトは、詳細な情報なしでデータ処理ができるように作られているじゃない か?」、「google も我々の詳細な情報なしで世界中の情報を探査できるじゃないか?」と言った感じで、そんなことは ないはずだという人もいるかもしれない。しかし、それはどんなことでもできるようコストをかけて、様々な機能を 組み込んであるか、深く洞察して見ると、実は単純なことしかできなかったことが見えてくる。Excel が長い期間を かけて現行の機能を実現するために費やした開発費は数百億円はくだらないだろうし、google の検索では、項目名を 指定した検索(圧力が○○以上など)は、そもそも目指していないのである。また、なんでもできるソフトウェアを使っ て、自分たちが利用しようとしていることを実現しようとすると結構難しいもので、高機能な Excel などは、よく知っ ている人やネットで、使い方を調べないといけないことがよくあるはずである。実はもっと汎用的なソフトウェアが 存在する。それは、プログラミング言語である。Excel もワードもパワーポイントもプログラミング言語で作られて いる。巷を賑やかせている機械学習の多くも python というプログラミング言語を使っている。プログラミング言語 も歴としたソフトウェアで,ある意味汎用ソフトウェアの極みである。何を言いたいかというと,様々なことができて, 汎用的になればなるほど、使いこなすためのハードルがあがり、実際使う力があったとしても多くの手数がかかると いうことである。そういう意味でも魔法の箱は存在しない(あったとしたら使いこなすためのハードルが高く、使う ときの呪文がやたら長くなるので、便利ではない)のである。世の中で便利だなというものは、ある一定の場面に特 化していて、その場面で発生するあらゆることをできるだけ簡単に実現できるように設計されているのである。つま り、「すべてのデータ探査と処理」の場面で発生するあらゆることがわからないとそれをオルゴールのドラムに刻み 込めないのである。しかし、「その場面で発生するあらゆること」を本当に手順化しているかというとそれは少し違う。 スマートフォンの音楽アプリで音楽を聴く時に、「アプリを起動して、○○という音楽を選択して…」というときに、 ○○の楽曲名ごとに手順書を書いていたら大変であるし、実際、アプリの開発者はそんなことはしていない。楽曲部 分は、変数化して手順を書くのである。

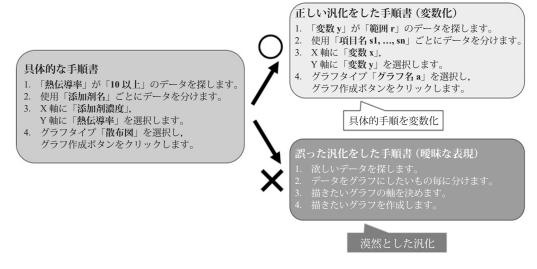


図9 手順書の正しい汎化方法

ただし、一度も具体的に楽曲を探さないで手順書を書くと失敗する。楽曲名がわからないことも多いし、同じよう な楽曲名や全く同じ楽曲名があるとき,「○○という音楽を選択して」の前に,「最近の」とか「△△が歌っている」 という絞り込みを入れないといけないはずである。この辺りはいろんなパターンがあると思うので、そのパターンが 出尽くすように色々具体例で手順書を書き、その手順通り実施をしてみて、その後、具体的な部分を○○、△△とい う変数に置き換える(汎化する)のである。また、汎化の時に正確性を犠牲にして、曖昧になってしまうことがよく あるが、実際には数学の変数式のように一般化されているが厳密性は一切劣化していないものでないといけない。簡 単に「具体的な部分を変数に置き換える」と書いているが、この変数名が具体的なものをすべて的確に言い表す単語 でないと、結局自分が行いたいことの手順を知ろうとして手順書を見ても変数名がいい加減だと、自分が今しようと していることの何をどこに置き換えればいいのかわからない状態になる。変数名は重要なので、頭を悩ませるところ であるが、しっかりと考え抜く必要がある。実際には、「様々あるすべての実験自体及びデータ探査と処理の手順書」 も本当にすべてを列挙するのではなく、できるだけ多くのパターンに関して具体的な値で手順書を作成し、手順書通 りで意図通りのことができるかを確認してから、具体的な部分を変数に置き換えて汎化した手順書を作る必要がある。 そうすると手順書はそんなに多くの数にはならないし、稀な作業に関しては、今回のシステムでの利用対象外として、 手順書を書くのをやめてもよい。どちらにしても「様々あるすべての実験自体及びデータ探査と処理」のパターンと 頻度は、把握しておく必要はある。そして、これはソフトウェア会社にはできなく、実際の研究者しかできないこと である。

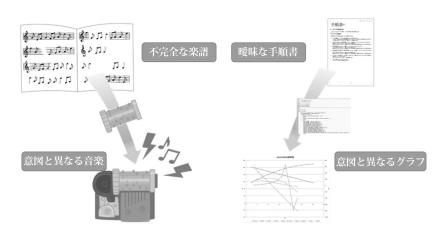


図 10 曖昧な手順書は意図とは異なる結果をもたらす

手順書づくりの中で、取り組んだほうが良いことがもう一つある。実験自体及びデータ探査やデータ処理の手順が明確になったら、「なぜ、そのデータ探査をするのか?」、「なぜ、そのデータ処理をするのか?」を複数の研究者で意見交換をし、合意を形成してみることである。たぶん、漠然とした答えはすぐ出るだろうが、新人や素人にその理由がはっきりとわかるというレベルで、明確な答えが出ることは少ないはずである。もちろん、専門的な背景の有無により、わかるわからないということはあろうと思うが、それ以上に曖昧になってしまっているはずである。手順の話と同じで、人間の頭の中で分かっていても他人にはそれを伝えられないし、わかっていると思っている人同士でも思っていることが同じかどうかは確かめられないのである。そして、「なぜ」が明確になっていないということは、目的もはっきりしていないということで、そのように目的が曖昧なままの状態で作業をすると、結論もおかしくなるのは皆様よくお分かりの通りである。また、「なぜ、そのような方法でデータ探査をするのか?」、「なぜ、そのような方法でデータ処理をするのか?」も同じように複数の研究者で意見交換をし、合意を形成することは、非常に重要ということである。これらの「なぜ」の真の目的を明確化するためには、障害の原因分析によく使われている「なぜなぜ分析」という方法が便利である。5回程度「なぜ」を繰り返すことで、深堀ができ、真因に近づける。「なぜなぜ分析」自身の方法は、ネットや書籍でもいろいろな解説や注意点が書かれているので、漠然とではなく、必ずそちらを参照して実施してほしい。真の目的が明確になれば、真の目的から実験自体及びデータ探査やデータ処理方法を

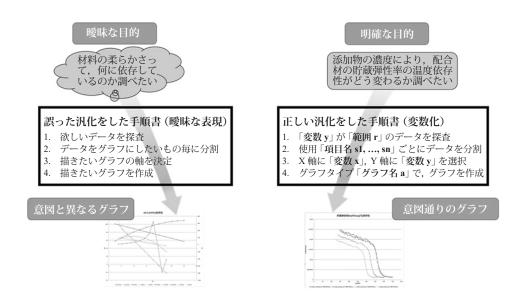


図 11 目的を明確にしないと手順も結果も意図通りにはならない

手順書ができれば、実際にその手順書が汎用的に利用可能かを検証する意味を含め、その手順が発生すると手順書の各工程にかかる時間を典型的な数パターンで測定し、記録する。時間を記録するために手順書通りに実施するのではなく、それこそ新人などにお願いして、日常業務を遂行するために手順書を使い、ついでに時間を記録してもらうのである。それなりの日数をかけることで、手順書の問題も明らかになってくるはずである。頻度に関しては、ある程度は実測できるだろうが、限られた時間ではどうしても偏ってしまうので、全員で意見を出し合い、合意形成をするしかない。ここで、手順の問題を明確化しておかないとシステム化してからでは、原因がどこになるのかを特定する作業や手戻りが大きすぎる。急がば回れである。

8. 自動・自律実験に先立つ属人的データ共有状況を脱するための事前準備(システム化対象特定編)

データ共有・利活用,自動・自律実験のシステム化及びシステム運用には必ず負担がかかる。したがって、闇雲にいるいろな作業をシステム化の対象と考えるのではなく、作業に品質(網羅性、均質性、再現性)が重要なもの及び作業の発生頻度と作業にかかるネット時間(研究者が実質的に作業をしている時間で、待ち時間や他の作業をしている時間を除く)の大きいものを序列化し、その上位の作業からシステム化対象にあてるべきである。実際、この序列は現業務の問題の深刻度とも対応するはずで、もし、現状深刻な問題を感じている業務が上記序列で上位にあがってきていない場合は、頻度や作業時間の算定が間違っているか?問題の深刻度に主観的な思い込みのバイアスがかかっているか?のいずれかである。複数の人と意見交換することや再計測をすることで、その認識ずれは解消しておく必要がある。

システム化対象が決まったら、次はどのようにシステム化を行うかを検討していく。この段階になって、初めてソフトウェア的観点を加味して考えることになり、システム開発関係者も参加して意見交換を交わすことになる。この時、注意すべき点がいくつかあるので、ここで触れておく。あくまでシステム化の主目的は、データ共有・利活用に対しては、網羅性、均質性、再現性という品質を担保すること、自動・自律実験に対しては、実験の省力化及び再現性向上、データ記録の誤記や未記録を防止することである。ここで重要な点は、システム化に便乗して「より楽に」、「より便利に」ということを考えてはいけないということである。業務系システムでは、手順が明確になっており、手順の実行頻度も安定的に固定化されており、また、変更は簡単に実施されないので、1回のシステム導入や更新でできるだけ使用者が「より楽に」作業遂行できるように配慮する。しかしながら、R&D 部門でのシステムでは、R&D の

宿命上、多くの手順変更や新たな手順が発生し続ける。つまり、できるだけ迅速に手順追加や手順変更に追従できることが最優先課題で、「より楽に」、「より便利に」ということは次点以下の課題なのである。実際、「より楽に」、「より便利に」を考慮しだすと結局、運用、メンテナンスが追い付かなくなり、本末転倒な状態になる。手動手順と同等のことが同等時間以下の作業で達成できれば、もれなく、網羅性、均質性、再現性という品質がついてくるので、これを大きな価値とする考えを持つべきである。実際、R&D部門で、データ共有・利活用がシステムでうまくいっているところはすべて、このような方針が徹底された上で、開発、運用を行っている。当該課題は、自動・自律実験に対しても言えることで、自動・自律化を「より楽に」、「より便利に」に設定できるようにすると、実験手順追加や手順変更性が低いシステムになってしまうことが多いのである。システムの利用を「より楽に」、「より便利に」することと実験手順追加や手順変更性を高くするということは、一般的には相反する条件なのである。

実際、「より楽に」、「より便利に」は、客観的でない(主観)欲求であることがほとんどで、客観的負担対効果を 蔑ろにする考えであり、システム開発、利用に行き詰まりを生む元凶である。「より楽に」、「より便利に」という要 求があった場合、あくまで、手動で共有・利活用をした時と比べて、どれだけ省力化できるかどうかという客観的指標で考えるようにする。そして、この客観的省力化量がシステム化の優先序列であり、省力化量の小さいものは、客観的負担対効果は小さく、取り扱うべきでないと判断をすべきである。もちろん、「より楽に」、「より便利に」の部分が、客観的省力化量が大きい部分があるのであれば、それは採用されてしかりである。自動・自律実験に関しては、単純なデータ共有・利活用システムより、判断基準が複雑になる。自動・自律実験は対象の作業部分が自動化されるので、手動実験と比べ必ず「より楽に」、「より便利に」になるのは明らかである。しかし、単純な作業自動化だけでなく、設定や結果確認など様々な部分に対しても「より楽に」、「より便利に」を求めてしまいがちである。R&D部門であっても業務系のように変化しない部分の実験は、「より楽に」、「より便利に」を追求してもよいが、そうでない R&D の真骨頂の部分は、実験の自動化以上の「より楽に」、「より便利に」はあきらめて、実験手順追加や手順変更性の確保を優先する必要がある。

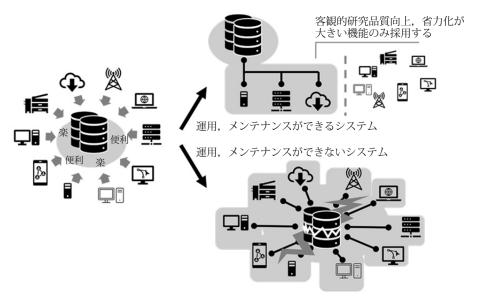


図 12 便利と楽を追及すると運用、メンテナンスできないシステムになってしまう

システム化に便乗して、今まで行っていないデータ探査、データ処理をこの機にシステム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

システムは、オルゴールの外箱であり、楽譜があって、それをドラムに刻み込まなければシステムは完成しない

ことは上述している。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奇跡である。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、悪戦苦闘している状況を考えると、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) その作業が他の作業よりシステム化すべきか客観的に判断できない

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、判断がつかないということである。つまり、そもそもシステム化対象を決める序列の「作業に品質(網羅性、均質性、再現性)の重要性」、「作業の発生頻度×ネット時間」の正確な値が算出できず、他のものと客観的に比較ができないのである。

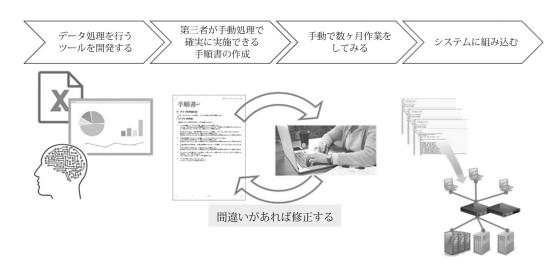


図 13 手動で行えていないことをシステムに組み込むまでの手順

「今まで行っていないデータ探査,データ処理をこの機にシステム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システム化は、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、システム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) ツールとシステム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、システム化を難しくしている原因である。このような R&D 部門でもシステム化を可能にするためには、ツールとシステムの分離が重要で、システムからツールを呼び出す形のシステムでなければならない。そのようにすることで、システム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、ツールとシステム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システム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システム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、ツールやシステムのバージョンアップ時にシステム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、システム部及び手順の問題の切り分けができない

ツールとシステム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?、ツールに問題があったのか?、システム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、システムを使わない状態で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。システム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、システム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、ツール部やシステム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システム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。

(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確認するのは、相当大変だが、システム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、ツールをリリースした1年後などに問題が見つかった場合は、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。

(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、システムには意味判別ができない記述)ものは、適切な数値に修正したり、新たな項目(下限値、上限値、信頼幅など)を作ったりという作業をする。この作業は一般的にデータクレンジングと呼ぶ。

9. 自動・自律実験データプラットフォーム導入に必要なソフトウェア要件

前項では、「自動・自律実験に先立つ属人的データ共有状況を脱するための事前準備」に関して、説明した。本項では、 R&D 部門の自動・自律実験に先立つデータプラットフォームのソフトウェア要件を説明する。

まず、自動・自律実験データプラットフォームがあれば、データ登録も分析も自動化され、お任せで所望の性能の 試作品が完成すると思っているのであれば、考えは改めなければならない。生産ラインのようなところでは、そこまで変化が激しくないので、商品ごとによく検討された自動・自律データプラットフォームを構築、導入することができるが、R&D のように変化が激しい部門では、自動・自律のためのアルゴリズム自体が完璧なものとして作り上げられるわけではなく、常に改善させていく必要があるようなものになる。また、R&D 部門では、実験で作成する試作品への性能要求も刻々と複雑化していくため、過去の自動・自律アルゴリズムが全く使い物にならないようなものも出てくる。R&D 部門における自動・自律実験データプラットフォームでは、そういうことを十分に加味し、様々な蓄積されたデータを使って、自動・自律アルゴリズムを改善していくことが必須なのである。また、あらゆる実験のあらゆる情報を自動でデータ登録するようなデータプラットフォームができればよいのだが、試行錯誤で刻々と複雑化していく実験全てを対象にすることは現実的ではない。自動・自律実験データプラットフォームを完全自動実験だけを対象にしたものとしてしまうことは、データプラットフォームを使わない実験の比率を上げてしまう。逆に研究者が手動部のある実験を避けてしまうことで研究自体の硬直化を生んでしまう場合もある。したがって、自動・自律実験データプラットフォームには、手動の実験や手動のデータ登録部も包含して取り込めるような仕組みを持たせる必要がある。それらを加味し、自動・自律実験データプラットフォームに必要なソフトウェア要件を以下で列挙する。

(1) 手動入力時の項目名の単語揺れや他データとの紐づけ時のミッシングリンクが発生しないように、データ登録や検索部分では項目名や紐づけはデータベースと連動したドロップダウンリストから選択できる機能を有するこ

と。また、データベースと不整合があったときには警告メッセージなどが出て、登録前に抑止できる機能を有すること。

- (2) 外注をしなくても研究者自身で項目名の変更, 追加や他データとの紐づけ変更, 追加の対応が可能な機能を有すること。
- (3) データ登録の手動入力部分は、外注をしなくても研究者自身でカスタマイズができる機能を有すること。具体的には、データ登録の入力部分は研究者が日頃から使用している Excel シートであることが望ましく、実際に登録するデータはその入力シートの登録したい部分を参照する形になっていることが望ましい。こうすることで、調液表などのように複雑な入力、確認が必要な場合でも研究者自身で、Excel シート内で演算をしたり、項目値に関しても選択肢を設定できたり、必須項目を設定したり、入力時の確認欄(合計値が 100% になるなど)を設けたりすることが可能となる。つまり、R&D 部門の多様性、変化性に追従することが可能になる。
- (4) ネットワークに接続された実験装置や計測器のデータを自動的にデータベースに回収、格納する機能及び検索結果のデーを他のサーバ、PCの既存アプリケーションに処理を受け渡すことができる機能を有すること。具体的には、Linux や Windows にそれぞれの利用者アカウントでログインし、既存アプリケーションをコマンド実行し、結果ファイル等をシステムに自動回収し、同時にデータベース化を行う機能を作りこむのではなく、システム基盤機能として有していること。

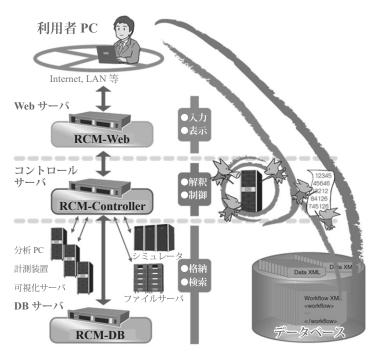


図 14 R&D の変化に追従するために必要なシステム

- (5) 機能変更・追加が頻繁に必要となるので、機能拡張が容易となる機構を有していること。具体的には、スキーマレスなデータベースを持ち、システムからデータベースやツールを起動し、様々な処理をする部分を追加変更するにあたり、システムのリビルドが不要な機構を有すること。
- (6) 一つのシステムハードウェア内に複数の独立したサブシステムをくみ上げることができ、アクセス制御を行うことで、サブシステム間の情報のやり取りも可能な仕組みを非仮想環境上で実現する機構を有すること。サブシステムごとに仮想環境にすると、独立性はより高くなるが、メモリや CPU や HDD のリソースまで完全独立になり、

また、OS部がサブシステム分だけ重複するので、非常に非効率なシステムとなってしまい、要求ハードウェアが数倍以上ハイスペックで、一桁以上高価格になる。

- (7) R&D 部門用システムでは、機能変更・追加が頻繁なので、システム無停止で、サブシステムの追加や変更が可能な機構を有すること。他グループのシステム追加や機能拡張のために他のグループのサブシステムが運用停止することは、受け入れられないことが多い。
- (8) アクセス制御がプログラムレベルではなく、DBMS レベルで制御されいること。R&D 部門用システムでは、様々なカスタマイズ開発が行われていくはずで、プログラムレベルで、アクセス制御を行う仕組み(データベースにアクセス制御のための情報を登録しておき、それをプログラムで判断して、アクセス制御に合致する結果だけを返すようにする)になっていると、「アクセス制御に合致する結果だけを返す」というプログラム部に不具合があると簡単に、情報漏洩が起きてしまう。カスタマイズ開発で変更できないシステムの基盤部であるデータベース部でアクセス制御は判断されるような機構を有している必要がある。

10. 自動・自律実験データプラットフォームにおける項目名の決定の方法と注意点

前項では、「自動・自律実験データプラットフォーム導入に必要な要件」に関して、説明した。本項では、自動・ 自律実験データプラットフォームにおける項目名の決定の方法と注意点について説明する。

自動・自律実験データプラットフォームのようにデータ共有を行うためのシステムを運用する為には、7、8項だけ の事前準備では十分ではない。R&D 部門におけるデータ共有では、他の業務系のデータ共有と違い項目名の決定方 法にいくつかの注意点がある。R&D 部門と他の業務系の違いは、前項までに様々な違いを説明しているが、項目決 定における一番重要な違いは、データ分析を前提にし、データ検索、分析を行いやすい項目にしないといけないと いう点である。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、データ検索、分析時に必要な項目は、システム内部で自動生成する形にし、データ検索、分析 をし易くしてしまうのである。R&D 部門では、データ検索、分析が複雑なので、システム内部で自動生成が困難と いうだけでなく、そもそも項目名が多く、追加や変更も頻繁にあるので、このようなことでシステムの運用管理負担 を増やしてしまうと、運用が破綻してしまうのである。つまり、R&D 部門のシステムでは、他の業務系のシステム のように運用が静的でなく、動的(項目名の追加、変更やデータ処理方法の追加変更)なので、如何に運用負担を軽 減するかが、運用可能なシステムにするための大きな鍵なのである。世の中のシステムは、殆どが静的な運用を前提 に作られており、システム技術者や情報システムの専門家でも動的運用を意識し、経験している組織、人は少ないの で,十分に注意する必要がある。システム技術者や情報システムの専門家の常識では,UX やデザインのように利用 者が利用しやすくすることを「正義」として、「より使いやすく」を追求するのが正しいとされているが、その観点 には持続可能な運用、運用負担の最小化という観点が全く欠落している。そこでは、運用負担を大きくしても十分メ リットのあるものは、システム化する価値があるという観点が底流に流れているからである。この観点で、R&D部 門のシステムを見ると、「運用負担を大きくしても十分メリットのあるもの」には該当しないことになり、それが現 状のような、R&D 部門のシステムが存在しない、若しくは存在しても実質的に使われない状況に陥っている原因な のである。以下, R&D 部門のシステムにおける項目名を決めるときに注意すべき点を列挙しておく。

(1) すべてのデータは、項目名 - 項目値という単純な形に落とし込む 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、以下のような表になっていることが多いと思う。

	乾燥	本本を士	士和		
1 É	2目	2 段目		焼結工程	
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、以下のようになる。

	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
Ì						

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。

(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。

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
EXP 1	エチレン	80	プロピレン	20	10
EXP 2	ブタン	75	エチレン	25	12
EXP 3	プロピレン	60	ブタン	40	8

上表は、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。

以下の表は、上表から「項目間の論理的関係」を排除した項目名である。

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
EXP 1	80	20	0	10
EXP 2	25	0	75	12
EXP 3	0	60	40	8

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。

(3) 他工程の ID 値を項目値とする項目は作ってはいけない

工程内に関しては、「項目名 - 項目値」に揃えるのは簡単なのだが、工程を跨いだ時に「項目名 - 項目値」を維持するのが難しい。実際、「項目名 - 項目値」の形が維持できなければ、工程を跨いだデータの分析が非常に困難になる。以下、複数工程実験のデータの項目名に関して、簡単な例を示しておく。

原材料をエチレン,プロピレン,ブタンとし,その原料を混ぜて,合成物質を作り,その合成物質を3種配合して,配合品を作るという複数工程実験を考える。合成工程では,エチレン濃度,プロピレン濃度,ブタン濃度という3つの項目名とする。この段階では,「項目名-項目値」の形になっている。2次元表形式で書き下すと,次のようになる。

合成 ID	エチレン濃度	プロピレン濃度	ブタン濃度
合成 1	10	0	50
合成 2	60	30	0
合成3	20	20	50
合成 4	40	0	50
合成 5	60	30	10
合成 6	10	20	50
:	:	:	:
合成 n	10	20	40

次の配合工程では、合成物質を 3 種配合するので、それぞれの濃度を合成 α 濃度、合成 β 濃度、合成 γ 濃度という 3 つの項目名とする。

配合 ID	合成 α 濃度	合成β濃度	合成γ濃度
配合1	50	0	40
配合 2	60	30	0
配合3	10	10	70
:	:	:	:
配合 m	10	20	0

この工程だけを見ると「項目名 - 項目値」の形にはなっている。ただ、このままでは、合成 α 濃度が、合成 1 の濃度なのか?合成 2 の濃度なのかはわからないので、困る。そこで、どの合成 1D を配合したのかの情報がわかるように α 合成 1D、 β 合成 1D という項目を追加する必要がある。

配合 ID	α合成 ID	合成 α 濃度	β合成 ID	合成β濃度	γ 合成 ID	合成γ濃度
配合1	合成 1	50	合成 2	0	合成 3	40
配合 2	合成 2	60	合成 3	30	合成 4	0
配合3	合成 3	10	合成 4	10	合成 5	70
:	:	:	:	:	:	:
配合 m	合成 n	10	合成 1	20	合成 2	0

この表の形は、どこかで見覚えがないだろうか?実はこれは、(2) で、「項目間に論理的関係があり、このままではデータを探すことが困難だったり、X-Y プロットグラフを描くことができない」と説明した形なのである。今回の複数工程データに関しても、同じように「項目間の論理的関係を排除」すると以下のような項目名の表になる。

配合 ID	合成 1 濃度	合成 2 濃度	合成 3 濃度	合成 4 濃度		合成 n 濃度
配合1	50	0	40	0	• • •	40
配合 2	0	60	30	30		0
配合3	0	0	10	10		70
:	:	:	:	:	:	:
配合 m	20	0	0	0		10

ある工程の実験を特定する値、つまり「行」に割り振られる ID は、それを参照する下流の工程の項目値として記録するのではなく、項目名の一部に組み込まれないといけないということである。今回の場合は、「合成 1」は、配合工程から参照される合成工程で作成された合成物の ID であり、それをどれだけ使ったかを意味する「濃度」という単語と組み合わされ、配合工程の項目名「合成 1 濃度」になっている。

(4) 論理的関係のある項目のままデータ分析をするとどうなるのか?

「論理的関係のある項目を作ってないけない」理由は、データの絞り込みやデータ分析が難しくなるという説明を してきた。ここでは、論理的関係のあるまま分析するとどんな問題が起こるのかを具体的な例をもとに説明をする。 以下表の原材料名と原材料濃度は、論理的関係のある項目である。

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	合成温度	引張強度
EXP 1	エチレン	10	プロピレン	30	100	350
EXP 2	プロピレン	30	ブタン	20	200	710
EXP 3	ブタン	50	エチレン	10	300	550
EXP 4	エチレン	20	プロピレン	40	400	400
EXP 5	プロピレン	60	ブタン	20	500	260

例えば、エチレンの濃度が引張強度にどのような影響を与えるかを調べたいと思ったときに、エチレンの濃度を X 軸, 引張強度を Y 軸として, グラフを描きたいと思っても上記表では, 非常に面倒だということはわかると思う。また、多変量解析の基礎である重回帰分析をしようと思っても、上記表のままではいろいろ不都合が生じる。以下, 重回帰分析を例にして、どのような点で困るのかを説明していく。

重回帰分析は、原則的には項目値が数値である必要がある。重回帰分析が、yを目的変数、xiを説明変数とした場合に、 $y=f(x1,x2,x3,x4,\cdot\cdot\cdot)$ 、例えばy=ax1+bx2+cx3+dx4+e(a,b,c,d,e は定数)の形の数式になることを考えると、文字列が入っていると演算ができないので、当然といえば当然である。ただし、世の中のデータ分析をしたいものには、文字列が入るもの(例えば、性別、出身都道府県、業種など)も多い。実はそういう時のために、ダミー変数化という処理を行うことで、項目値が文字列の項目でも重回帰分析が可能なように拡張できる。ダミー変数化処理とは、以下のように項目値が文字列の項目をその項目値を項目名とし、値が0もしくは1となるように変形することである。

実験 ID	エチレン1	プロピレン1	ブタン1	原材料濃度 1
EXP 1	1	0	0	10
EXP 2	0	1	0	30
EXP 3	0	0	1	50
EXP 4	1	0	0	20
EXP 5	0	1	0	60

エチレン 2	プロピレン 2	ブタン2	原材料濃度 2	合成温度	引張強度
0	1	0	30	100	350
0	0	1	20	200	710
1	0	0	10	300	550
0	1	0	40	400	400
0	0	1	20	500	260

ダミー変数化処理を行うとすべての項目は数値化され、重回帰分析が可能になる。上記表では、目的変数が引張強度で、説明変数がエチレン 1、プロピレン 1、・・・温度の 9 変数になる。機械的に行うと上記表のようになるが、原材料 1、2 が交換可能な対称性を持つという項目間の論理的関係性を考慮すると、以下のように表を変形することも可能である。こうすると説明変数の数は 6 個になる。

実験 ID	エチレン	プロピレン	ブタン	原材料濃度 1	原材料濃度 2	合成温度	引張強度
EXP 1	1	1	0	10	30	100	350
EXP 2	0	1	1	30	20	200	710
EXP 3	1	0	1	10	50	300	550
EXP 4	1	1	0	20	40	400	400
EXP 5	0	1	1	60	20	500	260

これは、(原材料名 1, 原材料濃度 1) と(原材料名 2, 原材料濃度 2) が交換可能な対称性をもった変数ペアだからできることであり、どんな場合でもこのように変形していいという訳ではない。つまり、データ表を純粋なデータアナリストやデータサイエンティストに渡しただけでは、このような変形さえできないことを理解しておくべきである。

実際には交換対称性を考慮した上表の6個の説明変数でも実は、説明変数としては過剰である。後で説明するが、過剰というのはまだ、考慮できていない項目間の論理的関係性があるということである。一般的に、説明変数の数は、その実験の独立変数の数=実験パラメータの種類数と一致している必要があり、それよりも多くても、少なくても正しい重回帰分析ができない。正しいデータを分析するためには、項目間の論理的関係性を完全に取り除いた以下のような項目にする必要があるのである。

実験 ID	エチレン濃度	プロピレン濃度	ブタン濃度	合成温度	引張強度
EXP 1	10	30	0	100	350
EXP 2	0	30	20	200	710
EXP 3	10	0	50	300	550
EXP 4	20	40	0	400	400
EXP 5	0	60	20	500	260

上表では、説明変数の数が 4 個になっており、これが真の実験パラメータであり、この実験の独立変数でもある。 上表では、変数が 4 個なので、1 次式(y=ax1+bx2+cx3+dx4+e)を想定すれば、5 個の未定係数を決めることができる。ちなみにこれを解くと、以下のようになる。

引張強度

 $=10 \times$ エチレン濃度 $-5 \times$ プロピレン濃度 $+3 \times$ ブタン濃度 $+4 \times$ 合成温度

本当の重回帰分析では、誤差も考慮する必要があるので、未定係数の数の十数倍の実験結果が必要であるが、原理的にはこのようにして、目的変数の説明変数依存性を明らかにすることができる。しかし、項目間に論理的関係があると、本当の実験パラメータではない、見せかけの変数が実験パラメータのように扱われてしまい、いくつかの問題により、上記、目的変数の説明変数依存性は導出できなくなってしまう。

一つ目の問題は、項目間に強い相関があると正しく未定係数が決定できないという重回帰分析の性質である。これは、論理的相関でなくとも、結果論として相関が強い場合も問題となるというものである。例えば、年齢、体重、身長などを説明変数とした分析をする場合、体重と身長には、論理的な相関はあり得ないが、実質的に身長の高い人の方が体重は重いという大きな相関があれば、データ分析が正しくできないというもので、多重共線性問題と呼ばれているデータ分析における基礎中の基礎の問題である。当然、論理的相関というものは、100%の相関であり、重回帰分析において、項目間の論理的相関は絶対に排除しないといけない問題である。

もう一つの問題は、項目間に論理的相関がある説明変数は、独立変数、つまり、真の実験パラメータではないため、 どの様な数式を対象数式(モデル式)として採用すべきかが、決めにくくなったり、別途制約条件を設定しなければ ならないという問題である。

例えば,

引張強度

 $=10 \times$ エチレン濃度 $-5 \times$ プロピレン濃度 $+3 \times$ ブタン濃度 $+4 \times$ 合成温度

を、説明変数がエチレン1、プロピレン1、・・・温度の9変数になる項目名で表し直してみよう。

引張強度 =10 × (エチレン1×濃度1+エチレン2×濃度2)

- 5 × (プロピレン 1 × 濃度 1 + プロピレン 2 × 濃度 2)

+3×(ブタン1×濃度1+ブタン2×濃度2)+4×合成濃度

見てわかる通り、エチレン $1 \times$ 濃度 1 のように項目間の掛け算が入ってくる。つまり、y = ax1 + bx2 + cx3 + dx4 + e のような 1 次式を仮定して、重回帰分析を行うのでは、この解にたどり着かないのである。2 次の重回帰分析であれば、エチレン $1 \times$ 濃度 1 のような項は、入ってくるが未定係数の数は 55 にも膨れ上がる。その 10 倍程度なければ、統計的に有意な未定係数が算出できないとすると 550 実験程度の実験が必要ということになる。また、その時に 1 次の項である単独のエチレン 1 や濃度 1 は、係数は厳密に 0 になるべきだが、重回帰分析で係数が偶然にも 0 になることはまずない。明確な形で、「1 次の項である単独のエチレン 1 や濃度 1 の項の係数は厳密に 0 になる」という制約条件を課す必要があるのである。

実際には、9変数から6変数にしたような論理的関係をしっかり考え、項目間に論理的関係が全くないような項目にしてから重回帰分析をすれば、見せかけの変数は消失し、正しい分析は可能になる。しかし、一旦、項目を決めてしまうとそれは結構難しい作業になり、データ分析の時にすぐに漏れなく思い出せるものではない。実際、9変数から6変数にするときは、(原材料名1、原材料濃度1)と(原材料名2、原材料濃度2)が交換可能な対称性をもった変数ペアということを考慮することで、見せかけの変数を3つ削減できたが、真の実験パラメータ自由度=独立変数の数

は4であり、さらに2つの変数削減が必要である。しかし、どのような項目間に論理的関係があり、どのように変数 削減を行うべきかを明確化するのは非常に難しいのではないだろうか?もし、漏れなく変数削減を明確化できたとし ても、データ分析毎に項目の変換が必要となり、データ分析が非常に煩雑になってしまう。そもそも、「データ分析 の時にどのような変数削減を行うべきかを明確化できる」ぐらいであれば、最初から項目間に論理的関係がないよう な項目で実験データ記録をしておくべきであろう。

実は、データを分析し易い項目とは、実験の独立変数を意識した項目のことであり、実験の独立変数に根差した項目はおのずとデータ分析がし易くなる。当然のことだが、項目間に論理的関係があるということは、その項目は独立変数にはなりえないので、何が独立変数なのかということを意識してデータ記録をすればいいわけである。実際、慣れてしまえば、実験の独立変数に根差した項目はデータ入力としても、違和感はないはずなので、データ共有を目指すのであれば、実験の独立変数とは何かをもう一度見つめ直して欲しい。

(5) 計測パラメータは独立項目にすべきではない

例えば、粘性係数を項目化する場合、測定パラメータとして、測定時温度が考えられる。データ分析の観点からは、 以下のように単純に、粘性係数、測定時温度の2つを独立の項目名としてもいい。

実験 ID	C2H2 濃度	C3H6 濃度	計測時温度	粘性係数
EXP 1	20	80	100	12
EXP 2	40	60	10	22

しかし、実はこの項目化は、データ分析の観点では問題はないが、蓄積の観点で、大きな問題がある。このように独立な項目にしてしまうと、測定時温度はその時々に測定者、実験者に都合がよい温度で計測をすることになり、一貫性が保てない。データを分析する場合、計測パラメータ値は同じもので比較をしないと意味がないのだが、このようにするとデータ全体で同じ温度のデータが少なくなり、結局、他人の実験結果を有効に活用できなくなるのである。上記例では、 10° C だが、 9° C だったり、 101° C だったり、計測が散在してしまう。このばらつきを抑えるために測定パラメータは、以下のように特性値項目名に組み込むべきなのである。

実験 ID	C2H2 濃度	C3H6 濃度	粘性係数 [10℃]	粘性係数 [100℃]
EXP 1	20	80	12	
EXP 2	40	60		22

もちろん、粘性の最大点、最小点、平均点などの温度は、測定者の任意性でなく、材料自体が持つ特性温度で計測パラメータに任意性が生まれないので、このような計測パラメータは項目化には問題がない。というよりもこのような計測パラメータは、粘性係数 [10℃] のように項目に埋めこむべきではなく、その粘性の特徴が分かるような項目名の最大粘性係数、最大粘性時温度として、項目化すべきである。

(6) 計測パラメータの連続値は項目にすべきではない

例えば、粘度の温度依存性を計測し、ガラス転移点を求める計測をした場合を考える。その時、ガラス転移点だけでなく、どのように粘度が変化していくのかの情報も記録として残しておきたいとする。しかし、このデータを(4)のルールに則ってすべて項目化し、データベースに格納しようとすると非常に多くの項目が必要になる。まず、そのようなデータはどのように検索し、どのようにデータ分析に使うのかを考察をしてみる。最終的に作成したいグラフは概ね次のようなもののはずである。

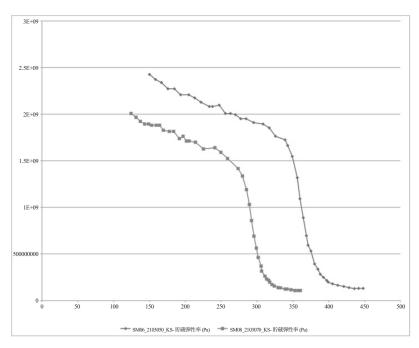


図 15 計測パラメータを X 軸にしたグラフの例

このようなグラフは、計測パラメータ依存性を確認するものであり、実験パラメータ依存性を確認するものではない。所望の性能を持つ材料を開発したい場合は、実験パラメータ依存性を見るべきであり、そういう意味で、このグラフは、試作品の性能向上で実験パラメータをどうすべきかの指標になるグラフではない。このグラフは、ある実験パラメータで作られた試作品の特性詳細(温度により特性がどう変わるか?)を知ることが主目的である。つまり、特定の比較したい試作品が決まっている場合に、このグラフを描くことになるはずである。通常は、何らかの方法で比較したい試作 ID を特定し、その試作品のデータで、このグラフを描くはずである。つまり、このグラフは試作 ID フォルダを作り、そこにデータを置いておけば簡単に描けるので、計測パラメータを項目化してデータベースに値を格納するのは、労多くして益が少ないということなのである。

ある実験パラメータで作られた試作品の特性詳細を知りたい場合は、確かに上図のようなグラフを書くことも確かにある。その場合は、下表のようなものを項目化し、データベースに値を1個、1個格納するのではなく、下表のようなものをファイル化して、ファイル自体をデータベースに登録しておけば事足りる。実際、上図のグラフを書くためには、検索で、比較したい試作品を特定でき、そのファイルを入手できればあとは簡単なはずである。実は、データベースに項目として登録すべきかどうかは、検索に使うかどうかが大きな分水嶺だと考えておけばよいのである。

粘度測定時温度	粘度
10	10
20	9
30	2
40	1

どんなものも項目化してデータベースに格納しておいて、自分の使いたいようにデータを引き出すというのは、一見便利に思えるが、「自分の使いたいように」の明確な定義が必要なのと、それを維持、運用する為の負担が非常に大きくなり、R&D では適切ではないということを改めて認識をしておいてほしい。

(7) 項目の単位は、データ分析をし易いものを考えるべきである

これだけでは、何を主張しているのかわからないと思うので、具体例で説明をする。例えば、エチレン、プロピレン、ブタン、触媒を混ぜて合成をするだけの単純な工程を考える。一番単純な項目名は、次のようなものになる。

a) エチレン重量(g), プロピレン重量(g), ブタン重量(g), 触媒重量(g)

この項目名で、合成物の粘度の触媒量依存性を調べる場合、どうすればいいか? X 軸に触媒重量 (g), Y 軸に粘度 $(Pa \cdot s)$ を設定して、X-Y プロットを描けばいいだけなのでは?と思うかもしれない。実は、データ分析をするためにはそれだけではダメで、X 軸以外の実験パラメータ、つまり、X-Y プロットを別グラフとして描く必要がある。そうしないと、X-Y プロットに触媒重量 (g) 以外の依存性が紛れ込んでしまい、正しい触媒重量 (g) 依存性をみることができないのである。

そうすると「エチレン重量(g)=プロピレン重量(g)=ブタン重量(g)=100g」の実験と「エチレン重量(g)=プロピレン重量(g)=ブタン重量(g)=300g」の実験は、別グラフのプロットになってしまう。しかし、すべてエチレン、プロピレン、ブタンの比率は同じなので、量効果(容器壁面、表面効果など)が無視できると仮定するなら、粘度など生成物の物理特性は同じになるはずであり、同じグラフに実験結果をプロットしたいはずである。もちろん、エチレン、プロピレン、ブタンの比率が同じものを何らかの方法で分類をして、それを1つのグラフにプロットすればいいのだが、データ数や項目数が多くなると非常に手間のかかる作業であり、その作業内で間違いを起こしてしまうとデータ分析が台無しになってしまう。実際に、実験データの記録は間違っていなくとも、その後のデータ処理時の変数値の四則演算などで間違いを起こし、データ分析を何度もやり直した経験のある方も多いのではないだろうか?現場の研究者はあまり意識していないことが多いが、データ分析時に煩雑な手動処理を避けることは、データ分析の再現性を確保し、データ分析の信頼性を高めるためには実は非常に重要なことなのである。

容器壁面,表面効果などの量効果はほとんど効かないことが多く,量効果を分離した分析をすることが多い場合は,総重量を分母にした重量百分率の単位の項目が便利であろう。

b) エチレン濃度 (wt%), プロピレン濃度 (wt%), ブタン濃度 (wt%), 触媒濃度 (wt%), 総重量 (g)

このような単位の項目にすると、X 軸に総重量 (g) を設定し、エチレン濃度 (wt%)、プロピレン濃度 (wt%)、ブタン濃度 (wt%),触媒濃度 (wt%)がそれぞれ一定値のデータで、グラフを描くと総重量依存性が簡単に確認でき、便利である。そして、エチレン濃度 (wt%)、プロピレン濃度 (wt%)、ブタン濃度 (wt%)のそれぞれが同じ値のデータごとにグラフを描くとすると「エチレン重量 (g) = プロピレン重量 (g) = ブタン重量 (g) = 100g」と「エチレン重量 (g) = プロピレン重量 (g) = ブロピレン重量 (g) = 300g」は、同一グラフにプロットすることができ、a) での問題は解消される。

しかし、よく考えるとエチレン濃度 (wt%) + プロピレン濃度 (wt%) + ブタン濃度 (wt%) + 触媒濃度 (wt%) =100 であるので、エチレン濃度 (wt%)、プロピレン濃度 (wt%)、ブタン濃度 (wt%)のそれぞれが同じ値 (例えば、それぞれ 15、40、42 wt%) だとすると、触媒濃度 (wt%)の値は 1 つの値 (3 wt%) しかとらなくなってしまい、X 軸に触媒濃度を設定し、触媒濃度 (wt%) 依存性を確認しようとしても、プロットは縦に並ぶだけ(触媒濃度 =X 値は、1 つの値だけ)になってしまう。ちなみに、重回帰分析を行う場合は、濃度のどれか一つ(例えば、エチレン濃度)を変数から外して、つまり、一つの変数を従属変数とし、独立変数の数に合わせて分析すれば、モデル式が適切かどうかということ以外は、問題は発生しない。あくまで、X-Yプロットを描いて、データ分析をしようとするときに問題が出るということである。

次に「触媒濃度 (wt%) 依存性を確認しようとしても、プロットは縦に並ぶだけになってしまう」問題の解消を試みてみよう。エチレン+プロピレン+ブタンの重量を分母にした重量百分率を 'wt% と書くこととすると、以下のような項目名が定義できる。

c) エチレン濃度 ('wt%), プロピレン濃度 ('wt%), ブタン濃度 ('wt%), 触媒濃度 ('wt%), 総重量 (g), 粘度 (Pa·s)

このような単位の項目にすると、エチレン濃度 ('wt%) + プロピレン濃度 ('wt%) + ブタン濃度 ('wt%) =100 であり、触媒濃度 ('wt%) はそれらと完全に独立になるので、X 軸に触媒濃度を設定し、触媒濃度 (wt%) 依存性を確認する場合でも、b) で問題となったようにプロットは縦に並ぶだけになってしまうことはない。もちろん、重回帰分析を行う場合は、エチレン、プロピレン、ブタン濃度のどれか一つ (例えば、エチレン濃度) を変数から外して、分析する必要はある。

実は、実験の進めやすさの観点から部数という単位で、材料の割合を記録しておくことが結構多い。エチレン、プロピレン、ブタン、触媒の例で言うと、エチレンを基準として、そこからの相対比でプロピレン、ブタン、触媒の量を規定するというものである。この単位でデータを蓄積していくためには、いくつかの注意が必要である。まず、基準のエチレンの部数が固定で、データベースに格納しているすべてのデータでそれが統一されていなければならない。よくあるのが、エチレンを貯めているタンクの容量をそのままエチレンの部数にしてしまう例で、複数種のタンクがある場合に統一されなくなってしまうので、a)と同じ問題が出てくる。次にもっと問題になるのが、基準物質自体が統一されない場合である。b)の wt% の場合は、全体量との比率なので、基準である分母が全体量となり、これが揺らぐことはないが、部数という単位は、基準を任意に変えることができるので、これを意識的に統一させなければならない。あと、どうしても避けられないのが、基準物質を一切含まない実験は、その単位でデータベースに格納できなくなるという問題で、必ず含む材料がない限り、部数という単位で、データ共有をすることは難しいと考えるべきである。もちろん、分析の仕方が部数の方がスムースということもあるその場合は、wt%と○○を基準部数 100 とした部数という 2 つの単位の項目を 2 重持ちさせることを考えるしかない。

本項で伝えたかったことを纏めると、記録蓄積するデータの項目およびその単位は、実験時の記録の取りやすさではなく、実験の独立変数を意識し、さらにデータをどのように分析するのかを具体的に考え、分析に沿った単位でなければならないということである。

11. 自動・自律実験データプラットフォームを使ったデータ分析の方法と注意点

前項では、「自動・自律実験データプラットフォームにおける項目名の決定の方法と注意点」に関して、説明した。 本項では、自動・自律実験データプラットフォームを使ったデータ分析の方法と注意点について説明する。

実験データがデータベースに蓄積されただけでは、データベース化したメリットはほとんどない。意図したデータを絞り込み、絞り込まれたデータを比較分析することができなければ、データベース化したメリットは小さいのである。一見当たり前のように思うが、多工程のデータを絞り込み、比較分析することは実はそんなに簡単なことではない。自動・自律実験データプラットフォームがあれば、そういう分析は不要で、自動、自律的に実験を行ってくれると思っているかもしれないが、自動、自律的が常に最適とは限らないし、自動、自律アルゴリズムは常に改善させていかなければ、陳腐化してしまい、役に立たなくなる。そういう意味で、自動・自律実験データプラットフォームがあっても人がデータを絞り込み、比較分析することを止めてはいけないのである。本項では、実験データの分析を行うためには、どのような手続きが必要かについて、初歩的な部分を説明する。

世の中では、MI、AIが叫ばれているが、MI、AIは、アイデアを提供してくれるという点では心強いが、それがなぜ、正しいのかは提示してくれないので、データ分析、理解という観点では、無力である。実際、MI、AIを活用するにしても、MI、AIが提示した案について、実データを分析することで、その背景理由の裏付けをとる必要がある。これは、自動・自律実験データプラットフォームでも同じことである。背景理由の裏付けをとるための初歩的なデータ分析方法が義務教育時代から親しんだ X-Y プロットである。たかが、X-Y プロットであるが、されど X-Y プロットである。結局、様々な学術論文でも賑やかしの 3D 可視化や画像などがあったとしても、論文で一番重要な部分はX-Y プロットになっているのは、偶然ではない。

X-Y プロットは,「Y 軸項目値の X 軸項目値依存性」を表すもので,データ分析,理解というものの基本的な部分

は、まさしくこれを使う。一般的に X 軸項目に実験パラメータを割り付け、Y 軸項目に実験結果項目を割り当てる。何らかの条件を指定して、データベースに蓄積されたデータから絞り込んだデータに関して、実は、注目している実験パラメータを X 軸項目に指定し、依存性を確認したい実験結果項目を Y 軸項目に指定しただけでは駄目なのである。実際のデータ分析では、いくつかの注意が必要である。まず、データベースに蓄積されたすべてのデータにおいて、実験パラメータが一つということは、あり得ないはずで、注目している実験パラメータ以外にも実験パラメータがたくさんあるはずである。その実験パラメータを全く無視して、X-Y プロットを描いてはいけない。実際、そのようなことをすると「Y 軸項目値の X 軸項目値依存性」を見ているようで、表には出てきていない X 軸項目以外の実験パラメータ依存性を X 軸項目値依存性と誤認してしまう恐れがある。

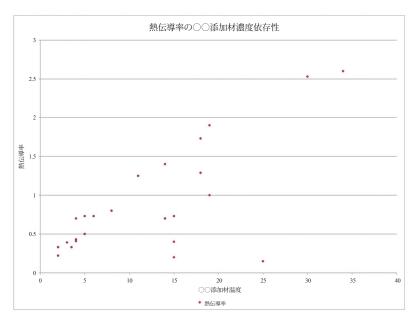


図 16 X 軸以外の実験パラメータ値で分類をせずに作成した X-Y プロットの例

実際、図 16 の X-Y プロットは、様々な実験結果に対して、注目している \bigcirc \bigcirc 添加剤温度を X 軸項目に指定し、依存性を確認したい熱伝導率を Y 軸項目に指定したもので、添加剤温度を上げていけば、添加剤温度 1° C あたり、熱伝導率は $0.07 \sim 0.08$ ずつ比例して上がっていく傾向にあることが読み取れてしまう。

実は、上記 X-Y プロットは問題なのである。「注目している実験パラメータを X 軸項目に指定し、依存性を確認したい実験結果項目を Y 軸項目に指定」する前に X 軸項目以外の実験パラメータが一定であるデータ毎にデータを分類しておく必要があるのである。多工程の実験では、実験パラメータは数百から千程度になるので、「X 軸項目以外の実験パラメータが一定であるデータ毎にデータを分類しておく」作業は、結構大変な作業になる。しかし、この作業を省略して、「Y 軸項目値の X 軸項目値依存性」をグラフ化してはいけない。

多工程の実験では、実験パラメータは数百から千程度になるので、「X 軸項目以外の実験パラメータが一定であるデータだけにデータを絞り込む」作業は、結構大変な作業になるが、この作業無くして、「Y 軸項目値の X 軸項目値 依存性」をグラフ化してはいけない。検討したい X 軸項目以外の実験パラメータが確定している場合は、データベースの検索条件にその値を指定すれば、「X 軸項目以外の実験パラメータが一定であるデータだけにデータを絞り込む」ことは、大変な作業にはなるができてしまう。しかし、そもそもその値がよくわかっていない場合は、そのような条件を指定しないで検索した検索結果を「X 軸項目以外の実験パラメータが一定であるデータ」毎に分類する必要がある。

数百から数千の条件指定を検索で指定することも大変だが、検索結果を「X 軸項目以外の実験パラメータが一定であるデータ」毎に分類するのは、手動処理では永遠に終わらないほどの作業量になってしまう。例えば、Excelで1000実験パラメータ項目を「実験パラメータが一定であるデータ毎に分類」しようとすると、1 個目のパラメータ項目の値を filter で 1 個指定し、次のパラメータ項目も値を filter で 1 個指定しという作業を 1000 項目すべてに行い、

やっと 1 つ実験パラメータが一定であるデータが分類できる。その後,最初のパラメータ項目の値を別の項目値に flter を変え,同じことを全項目の全項目値のパターンを抜けなく filter をかけていく必要があり,もし,項目値が 2 個ずつだとして,filter を 1 秒で行なえたとしても,全パターンの 2 の 1000 乗パターンを確認するには, $2^{1000}\times1$ 秒 = 2 の後に 0 が 300 個ぐらい付いた数(秒)= 宇宙の寿命よりはるかに長くかかってしまい,現実的ではない。実際には,何列か filter をかけると結果が 0 件(0 行)になることがほとんどなので,ここまで時間はかからないが,たぶんやりきる人はいないはずである。したがって,多工程の実験を行う R & D 部門では,こういうことが簡単に実施できるツールを整備しておくことが必須ということになる。

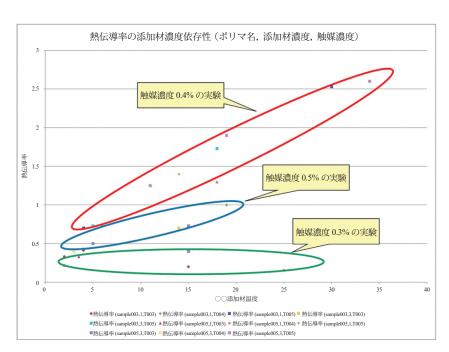


図 17 X 軸以外の実験パラメータ値で分類 (マーカー種分け) をして作成した X-Y プロットの例

先ほどのデータは、X 軸項目以外の実験パラメータ値で分類をしないで、X-Y プロットを描いたが、図 17 では追加しているポリマー種類や触媒濃度などの X 軸以外の実験パラメータの値が同じデータ毎にデータを分類(マーカー種を分け)し、データをプロットしたものである。そうすると「添加剤温度 1° C あたり、熱伝導率は $0.07\sim0.08$ ずつ比例して上がっていく傾向がある」のは、触媒濃度を 0.4% にした時だけで、0.3% の場合は、そもそも添加剤温度依存性がないことがわかる。X 軸以外の実験パラメータの値が同じデータ毎にデータを分類しないと、結論をミスリードしてしまいかねないのである。

実は、データ蓄積、共有を行うまでは、この作業は発生しないか、したとしても大きな負担ではない。データ蓄積、 共有を行うまでは、自分が依存性を調べたいと思う実験パラメータ数種だけを変動させて、それ以外の実験パラメータは固定した実験を行うはずであり、また、実験直後にデータ分析を行うはずである。その場合、ほとんどの実験パラメータは固定されており、また、自身も何を固定したという記憶が残っているため、ここで議論している「X 軸の項目以外の実験パラメータが一定であるデータ」を分類することは簡単である。X 軸の項目以外の実験パラメータを一定にして実験をするので、そもそも分類する必要がないことがほとんどである。つまり、こんなグラフを描きたいと思ってから、そのグラフが簡単に描けるような実験をしているということになる。

しかし、データ蓄積、共有をし、遥か過去の自分のデータや他研究者のデータを含めて分析をする場合、「どの実験パラメータが固定されているのか?」は、実データを確認するまで分からなく、また、殆どの場合は、実験パラメータの大多数が変動してしまっているはずである。実は、データ蓄積、共有をしない場合は、「実験前にグラフが描きやすいように実験パラメータを固定しているか」、少し前の実験であれば自身の記憶を活用することで「どの実験パラメータが固定されているのか」が分かるので、比較的楽にデータ分析ができるだけなのである。データ蓄積、共有を始めると自身の記憶を使わずにデータ分析をすることが強いられ、その分析の大きな変化を研究者が認識できてお

らず、当惑してしまうのである。データ蓄積、共有をする前に、自身の記憶を活用せず、実データをしっかり確認しながらデータ分析をする習慣をつけていくことが重要なのである。

10 項でも説明したが、X 軸項目に計測パラメータ(計測温度、計測周波数)を割り当て、Y 軸項目に実験結果項目(粘度、インピーダンス)を割り当てるグラフを書くこともあると思う。しかし、このグラフは、実験パラメータ依存性ではないので、試作物をどう作るべきかという指針を与えるものではなく、ある特定の試作物の特性を理解しようとするものである。もちろん、マーカー種別を変えることで、複数の試作物を並べてグラフ化することはあると思うが、あくまで、様々な試作物の計測パラメータ依存性を比較することができるだけで、試作物を作成するための実験パラメータに対する指針は、このグラフからは読み取れない(読み取るべきではない)はずである。試作物を作成するための実験パラメータに対する指針を得るためには、やはり、X 軸に実験パラメータを設定し、実験結果値をY 軸に設定する必要がある。

しかし、どうしても実験結果値の計測パラメータ依存性的なものを改善したいと思うこともあるのは確かである。そういう場合は、X 軸項目に計測パラメータ(計測温度、計測周波数)を割り当て、Y 軸項目に実験結果項目(粘度、インピーダンス)を割り当てたグラフを見るときに、グラフの中の何を見て、試作物間(マーカー種の異なるもの間)の違いを見ようとしているのかを自己分析的に考えてみるべきである。傾きであったり、変曲点であったり、最大値であったり、平均値であったりであり、連続的な量自身ではなくそういう特徴量を見ていることに気づくはずである。そういう特徴量を Y 軸項目に設定するとたぶん、自分が見たい「Y 軸項目値の X 軸項目値依存性」のグラフが書けるようになるはずである。

データ分析を行うときに、X 軸に実験パラメータでなく、特性値を設定することもあるはずである。例えば、試作物の粘度と熱伝導度の関係性を知りたい場合は、X 軸項目に粘度を設定し、Y 軸項目に熱伝導度を設定して、X-Y プロットを描くはずである。ただ、この時幾つかの気を付けなければならないことがある。特性値は実験パラメータのように項目値に一意性のある項目でない為、「粘度が○○の時に熱伝導度が××」というプロットがあっても、あらゆる実験でそれが成り立つかと言えば違う。実験パラメータが違っても「粘度が○○」になることは、十二分にありえ、その時、「熱伝導度が××」にならないことは十二分にありえるのである。だからと言って、全実験パラメータを固定して、X 軸項目に粘度を設定し、Y 軸項目に熱伝導度のグラフを描くと、1 点だけになってしまい、本来見たいものが見えなくなるはずである。もちろん、こういうことは、皆さんは承知で、いくつかの異なる実験パラメータの結果で、この手のグラフを描いているはずだが、どんな実験パラメータでもその傾向が同じかどうかはしっかりと確かめられていないことが多いはずである。つまり、この手のグラフを描く時は、たまたま、偏ったパラメータの実験での限定的な依存性(偽依存性)を一般的な依存性と勘違いしてしまう可能性が高いのである。したがって、この手のグラフを描く時には、事前に実験パラメータと当該 X 軸、Y 軸に設定する予定の特性値の相関を事前に分析しておき、どのパラメータ領域での粘度と熱伝導度の関係性を評価しようとしているかを明確化しておく必要があるのである。

12. 自動・自律実験データプラットフォーム導入時の落とし穴とそれを防ぐ方策

前項では、「自動・自律実験データプラットフォームを使ったデータ分析の方法と注意点」に関して、説明した。 本項では、R&D 部門における自動・自律実験データプラットフォーム導入時の落とし穴とそれを防ぐ方策に関して 説明する。

- 1) 項目名や項目間の関係性を明確化する段階で議論が紛糾し、纏まらない
 - =>まず、目的を再確認し、「項目名や項目間の関係性を明確にすること」は、目的のためになぜ必要なのかを皆で共有することが必要である。その上で、この議論が収束しない場合は、本課題を進めている人が適切かどうか再検討する必要がある。この手の合意形成を進めることができる人は、複雑な利害関係の絡んだプロジェクトを利害関係者間の落としどころを探りながら進めていける技量のある人である。単に、紛糾している状態を丸く収めるだけでは、問題を先送りしているだけであり、ガチンコ状態で、合意形成を取り付ける

技量が必要である。そもそも、皆がガチンコ状態になっていない状態であれば、ガチンコ状態に持っていく こと自身も大きな仕事である。たぶん、グループリーダなどが行うことが一番理想的であるが、リーダ以外 にもこういうことが得意な人物がいれば、リーダが背後に立ち、その人物に任せるのも一案である。社内の 適切な人員が見当たらない場合は、こういうことの経験のあるコンサルタントをファシリテータとして、採 用、外注することも考えるべきである。

- 2) 項目名などを一元化するための管理者が必要だが、それを決められない
 - =>「項目名や項目間の関係性を明確にすること」ができる人なら、この管理者も担うことができる。システムの管理者としての実務の管理者は、決められた作業をしておけばいいだけである。したがって、「項目名などを一元化することに尽力する人」は、1)と同じ人が担い、システムの管理者としての作業実務部分は、操作ミスを起こさない慎重な人に担ってもらえばよい。もちろん、どんな慎重な人でもヒューマンエラーは起こり得るのでダブルチェック体制は必要である。
- 3) システムが存在しない状態で、どのような機能を有するべきかを考えることは難しい
 - =>利用者が使いたい機能だけならまだしも、データ保全を意識した運用に必要な機能は、研究者にはどのような機能が必要なのか?また、どのような仕様であるべきなのかの判断は困難である。R&Dにおける運用績が豊富なパッケージ(機能を作らなくてよい)を採用するか、運用に必要な機能を研究者と合意形成を図ることができ、R&D部門のシステム開発、構築のコンサルティング経験のあるコンサルタントを雇うべきである。実際、こういうことをせずシステムを構築してしまうとシステムというより、ツール的(ツールは、その時のデータでその時に結果を返せばいいだけで、項目名が揃っているか?とか、同じことを他者ができるかという観点が極めて希薄)になってしまい。複数人の同時使用やバージョンアップ、データの意図しない削除など様々な問題を抱えたまま、その問題を認識しないまま運用してしまうことが非常に多い。また、そういうツール的システムは障害が認知された場合でも、何をして、どういう問題が起こったかなどの記録が残っていないことが多く、原因究明ができず、問題を放置した状態になることが多い。そうすると、信頼性を失ってしまい、次第に使われなくなってしまう。
- 4) 何をどのように探し、どのようにデータ処理しているのかの現状調査、分析が纏められない
 - =>これは、地道に具体的な作業を観察し、「何をどのように探し、データ処理しているのか」を調べ、記録する力とそれを汎化(パターン化)していく力が必要である。前者は、地道な作業ができる人、他の人に頼み事(実際に作業をしてもらうことを依頼し、横から観測して、何をしているかを記録する)するのが得意な人が適しており、後者はロジカルシンキング能力が高い人が適している。前者は、どんなグループにも一人ぐらいいるような気がするが、後者は理系研究者の中でも適する人は少なく、実は文系の企画担当や法律家やネゴシエータに適する人が多い。ただし、文系の企画担当や法律家やネゴシエータにとって、研究開発の基礎部分は単語一つをとっても今まで扱っていないものが多く、力が発揮できるまで、それなりの時間を要する。社内の適切な人員が見当たらない場合は、こういうことの経験のあるコンサルタントをファシリテータとして、採用、外注することも考えるべきである。

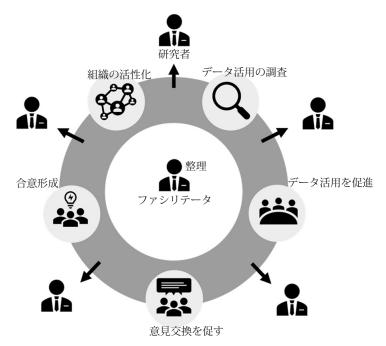


図 18 ファシリテータによる現状のデータ利用の実態や問題を把握、改善

ここまでの説明で、何度となくコンサルタントやファシリテータという言葉が出てきている。R&D部門の自動・自律実験データプラットフォームを導入、運用する際にコンサルタントやファシリテータに必要とされる力を以下にまとめておく。

ファシリテータに必要な能力は、研究者との深いコミュニケーション力とロジカルシンキング力であり、実際の研究者が行うしか仕方がない以下の課題の部分を支援し、完遂させる能力である。

- -1) 属人的データ共有状況からの脱却(情報記録,項目名定義)
- 0) 形骸化したデータ共有状況からの脱却(十分な情報,適切な項目)
- 1) 第三者が手動処理で確実に実施できる手順書の作成と汎化
- 2) 手順書作成時の問題点の列挙と改善案検討
- 3) 手順に沿った手続きの時間測定と発生頻度推定, 作業品質(網羅性,均質性,再現性)の重要性確定とそれらの序列化
- 4) 3) を基に作業のシステム化優先度(序列化)とどこまでシステム化するかの判断

上記を一般的な能力に置き換えると次のような能力ということになる。

- a) 研究者からの情報取得(項目列挙と定義, データ利用実態把握)
- b) a) の情報の曖昧さ、矛盾の指摘及び真偽を批判的に裏取り
- c) b) 後の情報を構造的にまとめて, 汎化
- d) 研究者間, グループリーダとの合意形成

上記 a~dの能力は、具体的にどのような能力か以下で具体的に説明をしていく。

- a) 研究者からの情報取得(項目列挙と定義, データ利用実態把握)
 - =>研究者と話ができ、実験データの項目及びその定義を聞き出し、また、そのデータをどのように利用しているかの実態を聞き出すことができる能力が必要である。もちろん、「実験データの項目及びその定義を教

えてください」、「そのデータをどのように利用しているか教えてください」といったところで、気前よく話してくれる研究者は少ない。本気で興味を持ち、相手が教えてあげようという気持ちになるように仕向ける必要がある。「こういう時困りますよね」とか、「こういう時は、こうすればうまくいったケースも知っていますが・・・」など相手が知らなかった情報を提示できると信頼関係が生まれ、一気に情報量が増えることが多い。過去の様々な経験を常に最前線に提示し、相手の出方でhit率の高い情報を提示できる訓練を日常的に行っておくとこれら能力は磨かれる。また、「そんなことないけど」と言われる失敗を恐れてはいけない。「そんなことない」場合は、「そんなことはない」ということが分かるということは、非常に重要な情報を得たということであり、自分が一段賢くなったと思うべきであり、全く失敗ではない。常に勉強をさせてもらうつもりで、情報取得にあたる心意気が必要である。

b) 情報の曖昧さ、矛盾の指摘及び真偽を批判的に裏取り

=>研究者は、悪意を持って嘘をつくことはまずないが、無意識的に嘘になってしまっていることも多いため、「情報が正しいかどうか」は必ず、裏取り(直接的に相手に確認するのではなく、他の情報等から論理的推定で確認する)する必要がある。取得した多くの情報を論理的に検証すれば、矛盾する情報も多々あるはずである。矛盾を解消する条件などを先に検討しておけば、真偽の裏取りにも使える。根本的には、ファシリテータは他人事というスタンスを捨て、明日から自分が代わりに研究を行う(単に決められたデータを探したり、決められたデータ処理をしたりするのではない。実際に実験はできなくとも、データを探して、データを処理、分析をして、他の研究者と議論ができるレベルでよい)つもりで、情報の曖昧さ、矛盾の指摘、真偽の確認を行えばよい。これら情報の真偽を直接聞くことなく真偽を確かめる力や矛盾を指摘する力は、探偵や検事に必須な能力でもあり、探偵や検事になったつもりで取り組むといいかもしれない。

c) 情報を構造的にまとめて、汎化

=>情報を構造的に分類し、まとめ上げる力、具体的な情報から仮説を立てて情報を汎化し、それを検証、改善していく力が必要である。また、それらを自然言語で正確に記述しきる文章化能力も必要である。

d) 研究者間, グループリーダとの合意形成

=>精査された情報を構造的、論理的に説明し、研究者間、グループリーダとの合意を取り付けるコミュニケーションおよび交渉能力が必要である。実際に合意を取り付けるためには、論理的に矛盾なくまとめ上げておく必要はあるが、論理一点張りでは難しく、相手が何に重きを置いているのか?などを察知し、それらを天秤にかける形で、交渉に持ち込む力が必要である。

13. 自動・自律実験プラットフォーム運用後の落とし穴とそれを防ぐ方策

前項では、「R&D部門における自動・自律実験プラットフォーム導入時の落とし穴とそれを防ぐ方策」に関して、説明した。本項では、R&D部門における自動・自律実験プラットフォーム運用後の落とし穴とそれを防ぐ方策に関して説明する。

1) 利用者がデータを登録してくれない

=>データ登録はそれ自体は利用者にとってメリットは全くなく、今までより手間がかかるだけの作業であり、 忍耐が必要な作業である。過去データの登録などの最初のデータ登録作業は、管理者等の特定の忍耐強い数 名が代表して行うことが好ましい。

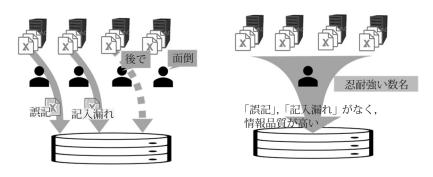


図 19 データ登録は初期段階は忍耐強い数名が行うべき

また、データは闇雲に登録をしてもそのあとの活用(データ探査、処理)に効果的ではない。そもそもデー タベース化しようという当初目的の具体的な動機部分に立ち返って、また、その動機が概念的であれば、こ れを機に具体化し、当初具体的目的を達成するために必要なデータを登録対象とすべきである。データの蓄 積が進みその効用が明確になってきたときに、その効用を具体例で説明し、徐々に一般利用者にもデータ登 録を行ってもらうようにする。ただ、効用を得るためには、研究スタイルを変える必要も出てくるので、研 究スタイルを頑なに変えない人たちには、効用で訴えるより、組織ルールとして、実施するように指示する 形が必要である。実際、データベースがなくて本当に困っている研究者はごく一部で、研究者は便利だった ら使おうという認識の人がほとんどである。業務系のシステムを見てもわかる通り、システムは便利にする ために運用しているわけではなく、属人性からくる揺らぎを排除することが主目的である。省力化の話にし ても業務系では、システム導入を理由に人員削減を実際に行うが、R&D はそういう評価の仕方はしないはず である。実際、システムが寄与する省力化は研究の数 % にしかならないはずで、それよりもシステムを導入 することで質的な変化(AIや機械学習による従来分析で発見できなかった材料開発)を求めているはずであ る。しかし、システムは手動でできないことはできないという宿命がある。これを乗り越えるには、従来手 動でできているデータ探査、データ分析に関して、システムを利用して行うことに慣れるようにし、次に、 そのままのシステムを使って、今まで行っていない少し高度なデータ探査、データ分析をする。その次に、 少し高度なデータ探査、データ分析の手順化をして、システムの拡張計画を立て、システムを拡張し、・・・。 これを続けていくことで、従来分析で発見できなかった材料開発手順が、AI や機械学習機能としてシステム に組み込まれてくるのである。こういう、step by step の変化は、その step ごとに研究スタイルを変えること になるので、使用方法やデータの解釈は慣れるまでは、全く「便利」ではない。したがって、「便利」を追及 している人たちを主対象にして、システム化は進めるべきではない。また、「使用方法やデータの解釈」を十 分理解しないまま、システムを使うと間違った他者のデータを使ってしまい、結果として誤った結論を導き 出す可能性があるので、「使用方法やデータの解釈」がしっかりできない人には、システムは使わせないほう がいいかもしれない。

2) 具体的にどう検索条件を設定すればいいかピンとこない

=>システムを使い始めた頃は、頭の中で「こんなデータを探したい」と思っても、具体的にどう検索条件を設定すればいいかピンとこないものである。手動作業時には、どう探していたかをよく思い出し、それを書き留め、等価なデータ抽出をするには、どう検索条件を設定すべきかを考え、それも書き留めておき、それらを皆で共有することから始めるとよい。どんなことでも文章に落として、それを自分以外の人にわかるように説明することで、自分が何をしようとしていたかが明確になってくるものである。実際には、この作業をしていると、「この方法はよくないかも」、「この方がいいかも」などの考えも浮かんでくる。自分の意識の中だけの考えは、しっかり固まっていると思いがちだが、結構いい加減な部分も多いものである。したがって、システムを使うというのを絶好の機会と思って、自分の頭の中を整理し、それを他者と共有しあうことを行ってほしい。



図 20 検索を使いこなすには慣れが必要

- 3)検索で指定したい、データ処理をしたい項目が存在しない
 - =>こういう事態がなぜ発生するかというとシステム開発前のデータ探査方法(データ処理方法も以下同じ)を 手順化するときに確認が不十分だからである。どういう不十分さがあるかという,二系統の不十分さがあ る。一つ目は、「配合比率で」とか、「評価特性値で」という個別の項目でなく総称名的な単語を使った記述 をする場合に、その明確な定義を書いていないからである。個別の項目名で手順書を作成すると数えきれな い手順書を作る必要があるので、同一パターンごとに汎化した手順書を作るという工程があるが、そこで汎 化したときに曖昧性を含む汎化を行ってしまっているのが原因である。これを防ぐためには、手順書を汎化 するときには、曖昧性を含まないように定義を明確にしておき、そして、具体的な項目名を列挙しておくこ とが重要である。二つ目は、項目名の合意形成の後にデータ探査手順書は作成することになっているが、手順 書作成時に合意形成されていない項目を使ってしまうことがあるからである。項目名の合意形成時には、実験 を再現することと評価をしたい値に対する項目を列挙していたので、データ探査のことは考えるようにはして いない。それは、その段階でデータ探査のことまで、考えると意識が発散し、「実験を再現することと評価を したい値に対する項目」に漏れが発生しやすくなるので、あえてそうしているのである。したがって、手順 書を作成するときは合意された項目名リストを見ながら行い,項目名リストにない項目で,データ探査をす る必要があるように思ったら、まずは、他の項目でその探査ができないかを吟味し、できるならそれを手順 書の TIPS(こういう検索条件を指定したい場合は、こういう条件設定をすれば等価である等)に書き加える。 他の項目でその探査ができない場合は、追加項目名の合意形成を行い、項目名リストにそれを追加すること が必要である。
- 4) 検索で指定したい(データ処理したい)項目は存在するが、検索条件として指定(データ処理)できない、しづらい =>これは、どういう状態の問題かというと、「サンプルIDがIDX0012a~IDX0115までのデータを抽出したい」とか「濃度が 1% 程度のデータを抽出したい」というデータ探査である。まず、「サンプルIDが ID2X0012a~ID2X0115までのデータを抽出したい」は、サンプルIDが数値になっていないので、範囲指定はできない。目視によりデータを絞り込む場合、人間の認識能力で特定の数字部分だけを無意識に分離をして、抽出している。そして、抽出したものが数値でなくとも範囲絞り込みができるように錯覚してしまうのである。手順書を書く時には、「人間の認識能力」を前提にする部分は危険で、単純な機械でもできるレベルに落とし込む必要がある。先の場合は、「特定の数字部分だけ」をどう取り出しているかのアルゴリズムを特定してそれを手順書に書き込むようにするか?、そのアルゴリズム特定が難しければ、特定の数字部分だけ別項目にすればよい。実は、研究でも日常会話でもこういう人間の便利な高度認識能力は、非常に便利でありながら、相手と意思の疎通がとれないもとになっていることも多々あるので、研究のような厳密な扱いが要求されるときは、意識をして、「人間の認識能力」に頼らない表現を心掛けることが必要である。「濃度が1%程度のデータを抽出したい」は、「1%程度」という部分で、これは本人も曖昧であることは認識している

とは思うが、手動で行うときはこれでうまくいくということも理解できる。「 $0.9\% \sim 1.1\%$ 」とするなら検索条件に指定することは容易である。しかし、そうはしたくないという気持ちもわかる。「 $0.9\% \sim 1.1\%$ 」で 2 点しかなかったり、1 万点もあったりしたら困るのである。だから、データを確認しながら分布状況を想定し、範囲を決めているのである。これもそういうところまで手順に書けば、手順書の改善が可能になる。「グラフを作成して傾向把握をしたいために 1% 付近でデータ点を 100 点ぐらいに絞り込みたいので、まずはデータの全体の分布状況を確認し、データ点を 100 点ぐらいになる区間のあたりをつけ、その区間を検索条件に設定する」という感じになる。少し脱線し本課題とは別の議論をするが、「グラフを作成して傾向把握をしたいためにデータ点を 100 点ぐらいに絞り込みたい」は、データ分析の観点からは、ある意味の意図的データ操作になり、本来はやってはいけないことである。100 点に分析的意味はなく、その人の作業労力や見やすいグラフを作成するためのもののはずだ。そういうことを理由にして、分析対象のデータを絞り込むのは行ってはいけないことである。分析結果には、検索条件も添え、なぜその検索条件を使ったかの理由を分析結果と論理づけて説明する必要があり、その時に説明がつかない検索条件は使うべきではないということになる。パラメータの多い実験では、検索条件を調整することで、自分の主張に近い結果を無理やり抽出することが可能になるが、それは事実を曲げてしまうことになるので、絶対に避けなければならないことであり、たくさんのデータを扱えるようになれば特に気を付けないといけない点である。

5)検索で抽出されたデータが少なすぎたり、多すぎたりし、適切な数にならない

=>ひとつ前の課題でも説明をしたが、検索条件指定し、データを絞り込むと2点しかなかったり、1万点もあったりすると困るという問題である。一つ目の原因は、すでに話したように「グラフを作成して傾向把握をしたいためにデータ点を100点ぐらいに絞り込みたい」という考えは、データ分析の観点からは、ある意味の意図的データ操作になり、本来はやってはいけないことである。そもそも、適切な数という概念自体がおかしい。今から抽出したい母集団の条件を何にするか?また、なぜ、その条件にするかを問われたときに答えられるように考えをまとめてから検索を行うべきで、できれば、その考えをメモしておき、検索結果の出力に書いておく習慣をつけておくべきである。このような研究の進め方を含め意見交換をする習慣をつけていくと、自分の意図を意識的に認識し、他者に伝える力が身についてきて、自ずとこういう問題は、発生しなくなる。

二つ目の原因は、データ数が多くなった時のデータ分析のノウハウや方針を持っていないことである。実際、他研究者のデータまで含んで検索すると思ったより(そもそも他研究者がどういう実験をしているか把握していないわけだからある意味当たり前である)もデータ数が大きくなることは十分あり得る。絞り込み条件が意図したものであれば、むやみにデータ点数を減らすのではなく、そのデータ点数のままで分析を継続すべきである。しかしながら、分析ソフトやグラフソフトで取り扱える上限点数があることも確かである。その場合は、そもそも何を分析したいのかを考えるべきで、そうすることで、一定区間内のデータ点に対して、平均、最大、最小点を算出し、それをプロットすれば済むことかもしれないし、重回帰分析を行ったほうがいいのかもしれない。そもそも、データ点数が増えると多次元の重回帰分析の精度が上がってくるので、データ点数が増えるということは、本来は悪いことではないはずで、それに適した分析ツール、手法を準備することがこの問題の正しい解決方法である。絞り込み条件が意図したもので、データ点数が分析をするのに対して少ない場合は、実験数が少ないということなので、実験を追加すべきということであり、データ点を増やすために検索条件を変えるのは本末転倒である。

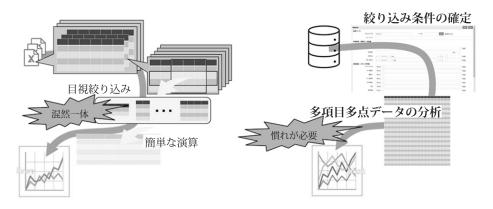


図 21 検索で得たデータで分析をするには多項目多点データの分析手法の慣れが必要

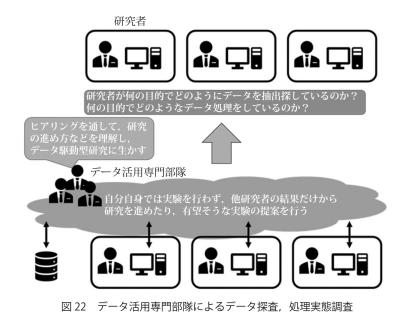
6) 項目数が多すぎて、データ分析が実施しにくい

=>今までは、実際にデータ分析をする時には、必要なデータがあるファイルや確認が必要な項目部分をファイル内から探しながら、データ分析に必要な項目だけデータ分析をしやすいように別ファイルに必要な項目だけデータを纏め直していたはずである。また、データを纏め直している過程で、追加で必要な項目に気づいて、その情報を再度、探すことも多いはずである。つまり、データ抽出とデータ分析の前準備が混然一体となっていたのである。それに対し、システムで行う場合は、検索一発で、結果が返ってくるのである。結果を返さない項目はもはや確認はできないし、確認したい項目をたくさん指定すると、膨大な項目が返ってくることもある。しかしながら、いろんなファイルを行き来しながらデータを確認、抽出することと比較するとそのを考えるその方が楽なはずであり、ヒューマンエラーも入り込む余地がない。ここで感じる「実施しにくい」は、頭がまだ手動手続きの思考のままなので、システム機能の使用に慣れていないからであり、手動で行っていた手順をシステム手順にしっかりマッピングして、意識をしながらシステム利用をしていくと慣れていく。つまり研究スタイルが変わっていくものである。実際、慣れる前にシステム利用をやめてしまったり、思い付きのシステム改善を提案してみたり、現行の研究スタイルにシステムを合わせるようにしてしまったりしがちであるが、これでは問題解決にはつながらない。

7) 手動の手順で問題ないので、システムは使わない

=>もし、手動手順で使うデータや項目がシステムの中に登録されていない場合は、まず、それらを登録できる ように項目追加などを行うことが最優先である。それと同時になぜ、手順書を作成した段階でそれを逃した のかの原因を調べる必要がある。手動手順調査後に、新たに始められた手順であれば致し方なしだが、過去 から高い頻度で使っている手順であれば、調査方法自体の再検討が必要である。そうではなく、システムで も実施可能なのに、あくまで手動手順で行っているのだとすると上述している「慣れ」の問題である。作業 品質と作業負担(時間)を客観的に評価し、負荷対品質の高い作業を選ぶ習慣をつける必要がある。システ ム利用に対する抵抗感が出た場合は、感覚、情緒的に判断を下すのではなく、定量的に判断するように心が けることが重要である。また、システム推進者が一度同じことをシステムと手動で行ってみて、手動より良 い点を実際の作業をしながら伝えるなどをすれば改善することがある。実際に、作業品質(網羅性、均質性、 再現性)は、システムの方が格段に良くなっているはずであるが、それに関して無頓着な人が多いのも確か である。しかし、グループリーダや事業部からすると作業品質(網羅性、均質性、再現性)は重要なはずで、 この辺りを理解してもらう教育は重要である。実際,手動処理をした結果に関して,システムを使って処理 をした結果、結論が変わったり、修正が加わった例を蓄積していったりすると、作業品質(網羅性、均質性、 再現性)の重要性が認知されてくるはずである。R&Dの良くない部分は、徹底的な議論が少なく、グループリー ダや事業部に半分あきらめてしまわれている点である。それ故、システム推進者の役割として、このような 活動を行うファシリテータの育成が必要となる。R&D 部門支援の経験のあるコンサルタントやファシリテー タを外注するのも一案である。

- 8) システムにデータが蓄積され、様々な統計、データ処理機能が提供されているが、結果考察がなく、少し考えれば おかしい結論を疑問に思わない人が増えた
 - =>統計処理, AI 処理は, ある程度データ数が大きく, 且つそれらの結果と物理的解釈の合意形成ができるまでは, ミスリードリスクが高い。信頼性が高くなってもそれだけから結論を導き出すことは避けるべきである。また, そもそも統計処理, AI 処理は高度な結果を示すので, 何か意味ありげに思えるが, 実際には, 根拠となるモデルや説明は提示されない。また, 実際に実データを直接見ていないので, 勝手な解釈が入り込む余地が大きくなる。したがって, 統計処理, AI 処理が進んでも, 実験データを実値のまま検討, 分析可能な環境は必ず必要で, 実値データの確認は必ず行い, 自分およびグループメンバ共有し, 仮説, 検証のループを回すことが重要である。
- 9) データ活用を専門とする MI (Material Informatics) 部隊は、データ分析結果はどんどん出してくるが、研究開発推進にはあまり貢献していない状況になっている
 - => AI がよく使われる画像処理などは、情報を捨象なしで処理をしているが、各ピクセル情報は正しく、入力サンプル数も大きくすることが可能である。もう一つの AI 成功例のマーケティングなどの社会分析は、個々の詳細な情報を捨象することで、情報次元に対しての入力サンプル数を大きくし、集団としての統計的特性を利用することで、AI を成功に導いている。一方、物理、化学を活用した材料、製品開発は、情報次元数に対して入力サンプル数は圧倒的に少ない。また、個々の詳細な手続きは意識的に行っているものなので、情報を捨象した統計処理をすべきではない(もちろん、分析の結果、その情報が不要だったということはありえるが・・結果としてそうなるだけで、最初から捨象してはいけない)。また、実験を再現するだけの十分な次元の情報が記録されているかどうか?記録したとしても正しく記録されているのか?など、情報品質も不十分なことが多く、情報品質の改善をかけながら MI を進めていく必要がある。つまり、MI 部隊は、実験者の実情を知って、実験者に実験の様々な面で改善提案をできるだけの信頼を得られる存在でなければならない。MI 部隊は、システム構築前の実験者ヒアリングにも緊密に関わることが重要で、そのためには単なるデータアナリスト集団ではだめで、同じ分野での研究現場経験者が7、8 割程度、また、それらを取りまとめるファシリテータも含まれる必要がある。



14. 効果的な自動・自律実験データプラットフォームと運用体制の作り方

前項では、「自動・自律実験データプラットフォーム運用後の落とし穴とそれを防ぐ方策」に関して、説明をした。 本項では、利用、活用される自動・自律実験データプラットフォームを実現する為の効果的な運用体制に関して、具体的な対処方法を提言する。

a) 「実験及び分析を再現するのに十分な項目」の調査,分析

研究者自身が「実験及び分析を再現するのに十分な項目」を意識でき、それをグループなどで話合い、合意形成を可能にする状態にできるのが一番好ましいのだが、研究者の特性上、難しい場合が多いのも事実である。そのような場合は、研究グループ内に支援者という名目でファシリテータを配置し、「実験及び分析を再現するのに十分な項目」の調査、分析を行なわせる体制を作ることをお勧めする。「実験及び分析を再現するのに十分な項目」がわからない、再現性の難しい「実験及び分析」も中にはあるはずだが、ここでいう「実験及び分析を再現するのに十分な項目」は、研究者自身が意識的に制御できていないものも含めて追及するという意味ではなく、意識的に制御しているものだけを意味する。もちろん、「実験及び分析」の理解が進むことで、意識的に制御できるものが増えてくるはずで、そういうものは随時追加できるような取り組みをすることが重要である。

b)「現状どのように過去データを探しているか?」の調査,分析

研究者自身が「どのように過去データを探しているか」を意識でき、それをグループなどで話合い、合意形成を可能にする状態にできることが一番好ましいが、研究者の特性上、それが難しい場合が多いのも事実である。その場合は、研究グループ内に支援者という名目でファシリテータを配置し、「どのように過去データを探しているか」の調査、分析を行なわせる体制を作ることをお勧めする。もちろん、研究の進め方が変わることで、「どのように過去データを探しているか」も変わってくるはずで、そういう変化に追従できる取り組みをすることが重要である。

c) 「蓄積されたデータをどのように利用するか?」の調査,分析

「蓄積されたデータをどのように利用するか?」は、自動・自律実験データプラットフォームを作る前に、入念に調査する必要があることであるが、システム構築をする関係者が研究者自身に直接ヒアリングをしても多くの場合、適切な情報が入手できない。研究者は、自分が現在どのようにデータを利用しているのか?、データが蓄積されたらどう利用したいのか?を意識的に考えていることは稀で、多くの場合漠然としたイメージでしか意識できていない。また、システム構築の関係者は、システムを構築することに前のめりになるため、研究者の言葉を直接的に鵜呑みにしてしまうことが多く、本当にその利用方法で研究者の課題が解決するのかまでの考察ができない。実際、研究者が行なっていることを、システム構築関係者が理解、把握することが専門領域的にも難しいのも事実である。したがって、前項で述べた研究グループ内に支援者という名目で活躍できるファシリテータ的な人材を調査、分析の任につけることも検討する必要があるのである。さらに、本調査、分析はデータ共有システムを作る前だけでなく、システム運用後も継続し、システムの改善、改良に役立てる必要がある。R&D部門は他の業務部門と異なり、多種多様、経時変化していくものなので、そういう変化に追従できる取り組みをすることが重要である。

d) 「実験及び分析を再現するのに十分な項目」を確実に記録させる体制

グループ内で「実験及び分析を再現するのに十分な項目」を記録する重要性について話し合いの場を持ち、自 律的に確実に記録できるようになるのが一番良いのだが、そうならないことの方が多いはずである。そのよう な場合は、研究グループ内に支援者という名目でファシリテータを配置し、「実験及び分析を再現するのに十分 な項目」が記録されているかどうかを確認し、足りなければ研究者に追記を依頼したり、ヒアリングをして自らが追記する体制を作ったりすることをお勧めする。MI 部隊がこのファシリテータ役を担うことも効果的な案である。実際にこのような体制を採ることは、誤記録や無意識の記録漏れ、データ捏造の抑制や実験及び分析の進捗状況の把握にも有効である。

海外では、徹底した分業体制を採り、分業間でデータ引き渡しをするポイントにおいて責任の所在を明確化することで、各自の責任部分の項目を記録する確実性を担保している。この方法は、分業部分毎の作業を明確化することで、業務の標準化と人材の配置転換の容易性を高くできる半面、各人の創意工夫の自由度や経験値を十分に生かしきれないというデメリットもある。日本の多くの企業では、良くも悪くも責任の所在の明確化や徹底した分業をせず、各人の創意工夫の自由度や経験値を十分に生かした業務の取り組み方法で、効果的なR&Dを進めてきた。しかし、「各人の創意工夫の自由度や経験値を十分に生かす」ために、データ共有が属人的になってしまっており、昨今のデータが大量に発生するR&Dでは属人化のデメリットがメリットを上回ってきている。

ファシリテータを配置しても改善できない,ファシリテータ自体を配置できない場合は,海外のような徹底した分業体制を取り、責任の所在の明確化と責任追及をする体制への移行も考える必要がある。

e) システム利用を推進する体制

単に「利用してください」とお願いをしたり、「どんな機能が必要ですか?」とヒアリングをしたりしても、効果的にシステム(蓄積されたデータ)の利用が推進されることはない。上述したが、研究者は研究対象について深く考えるが、自分が現在どのようにデータを利用しているのか?、データが蓄積されたらどう利用したいのか?を意識的に考えていることは稀で、多くの場合、漠然としたイメージでしか意識できていない。そのことを前提にして、システム利用を推進しようとするのであれば、研究者のどのような課題がシステムで解決するかを考え、課題がどのように解決できるかを提示し、利用へ誘導していく必要がある。また、その過程で、現在のシステムで解決できない課題が分かり、それを解決するためにはどのような機能があるべきかを考えることで、システムの拡張方針が決まることもある。

システムの利用推進や運用に関する体制には、もう一つ注意が必要な側面がある。運用者が研究者の行なっていることを理解、把握できなければならないことは、言わずもがなだが、理解、把握をしていたとしても、業務が完全に分離されている状況では、次第にそれぞれの立場に偏ってしまう。その結果、研究者は重要度が高くない機能を要求し、運用者はそれを漫然と受け容れるか、完全に否定的な態度をとってしまう状態に陥りがちになる。これらの状況を打開するためには、人事交流による役割のローテーションが有効である。研究者を推進運用側に異動したり、兼務させたりすることで、相手の立場、観点を意識できるようになり、特定の役割に偏り過ぎた認識が抑制され、効果的なシステム改善が可能になる。

また、システムの利用を推進するためには、運用体制だけでなく、利用者の体制に関しても工夫が必要である。 データ共有ができていない状況では、当然ながら研究者は自分以外の人が生成したデータを使うことを前提と して、研究を進めていない。しかし、データ共有できる状態になり、いくら便利になったとしても、急に研究 の進め方を変えられるわけではない。もちろん、一部の好奇心が旺盛な研究者は、これを機に研究の進め方を 変えるかもしれない。研究者は研究対象に対して好奇心は旺盛だが、それ以外の部分に関しては、人並み以上 に保守的な人が多いのも事実である。したがって、せっかく構築したデータ共有システムを最大限効果的に活 用するためには、従来の研究から離れて、他の研究者のデータのみで、研究を行なう専門部隊を立ち上げ、シ ステムを最大限効果的に活用し、成果を出すことを彼らの使命にするのである。彼らは、自身で実験を行なわ ないので、自分たちの成果をあげるために、有望そうな実験の提案やそのための説得資料作り、予算取りなど も自律的に行なうようになるはずである。

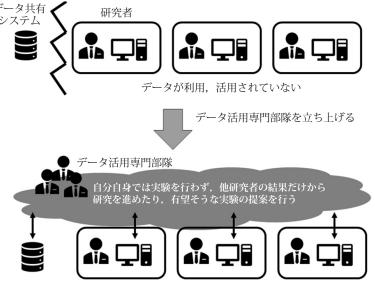


図 23 データ活用専門部隊によるデータ活用と既存研究者の活性化

おわりに

本節では、R&D 部門におけるデータ共有の実情が属人的、形骸化したデータ共有となっている実態を紹介し、その問題点及び原因を検討した。それらを基に効果的な自動・自律実験データプラットフォームと運用体制の作り方に関して提言をした。読んでいただいてお分かりだと思うが、R&Dのデータ共有、自動・自律実験を行なうということは、何かを買って、解決するという問題ではなく、企業風土及び R&D 従事者の意識変革が必要である。R&D 従事者は、業務内容が極めて専門的で、会社内でも対話する機会が少ない人たちだと思う。研究者は、とっつきにくいところはあるが、地頭は良く、非常にまじめなので、これを機に話をしてみれば、仲良くなれるかもしれない。そういう人とのつながりを作ることが、R&D 部門におけるデータ共有の第一歩である。

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 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