

第3節 電子実験ノートを活用した研究、実験データの効率的な蓄積と限界

(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。

本節では、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。

1. 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であり、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。

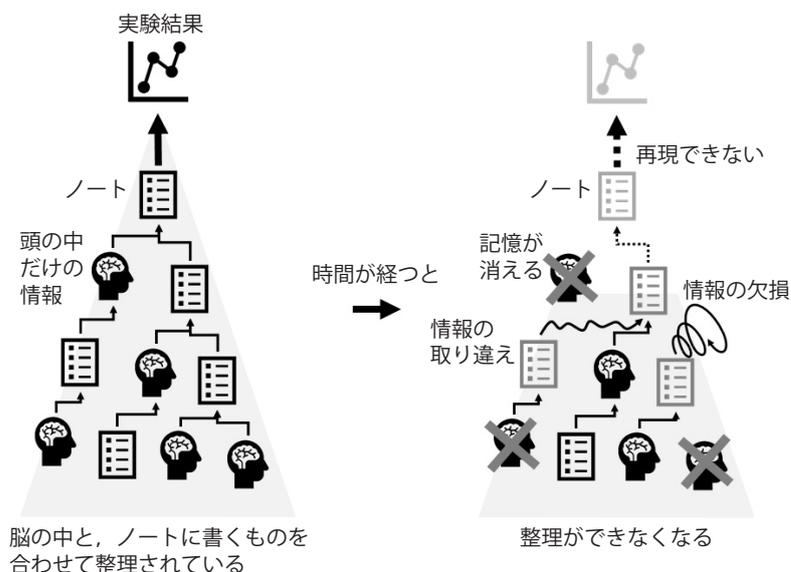


図1 属人的なデータ管理とその特性

「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、間違っただけの情報を基に実験や解析を進めることが発生し、間違っただけの結論が導かれたり、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。

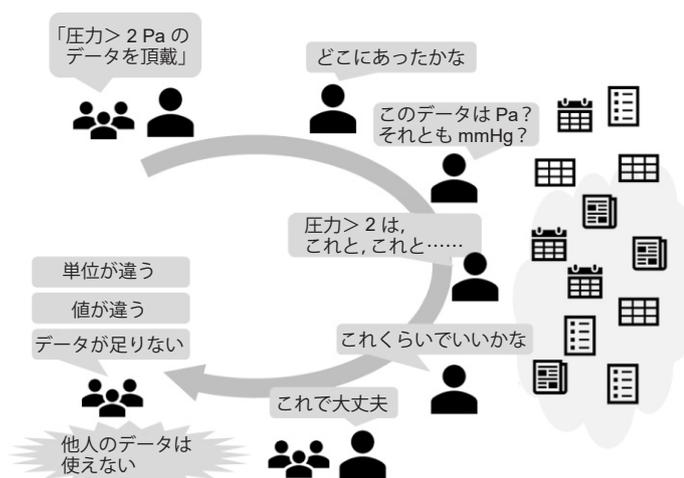


図2 属人的に管理されたデータが引き起こす問題

「属人的なデータ管理」状況が生み出される直接的原因は、そもそも第三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、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。

2. 報告書の共有で期待して良いこと、良くないこと

前項では、「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。本項では、「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しておく」とことと説明した。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っているからである。そのような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、そういうことは通常しない。もちろん、Appendix 等に報告書で参照した実験の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にしておけば、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。

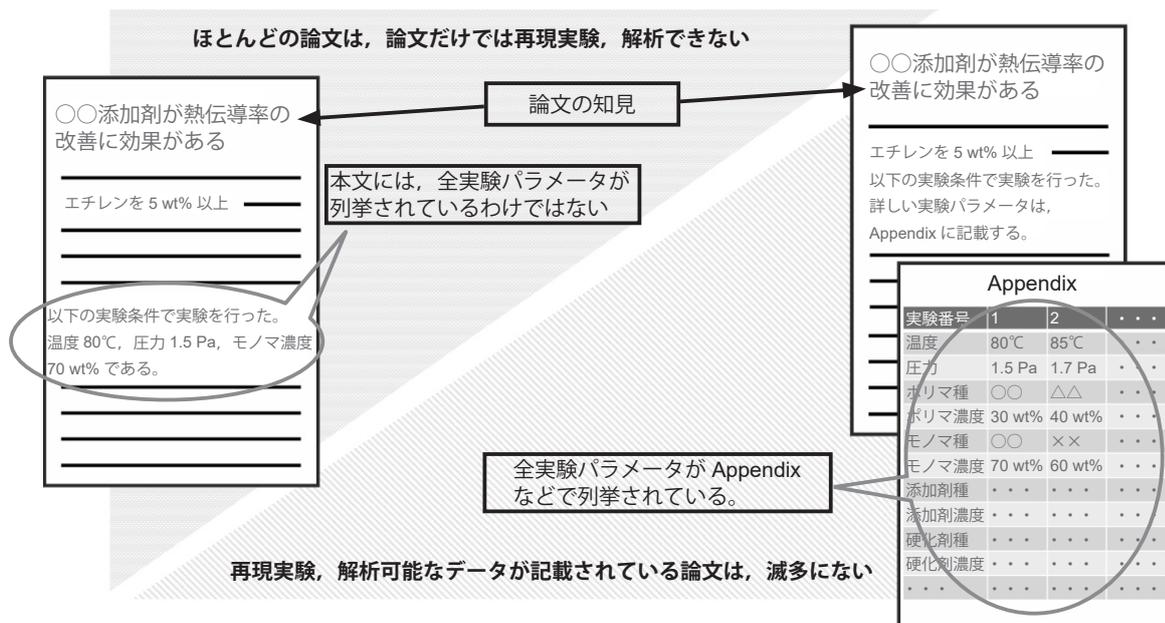
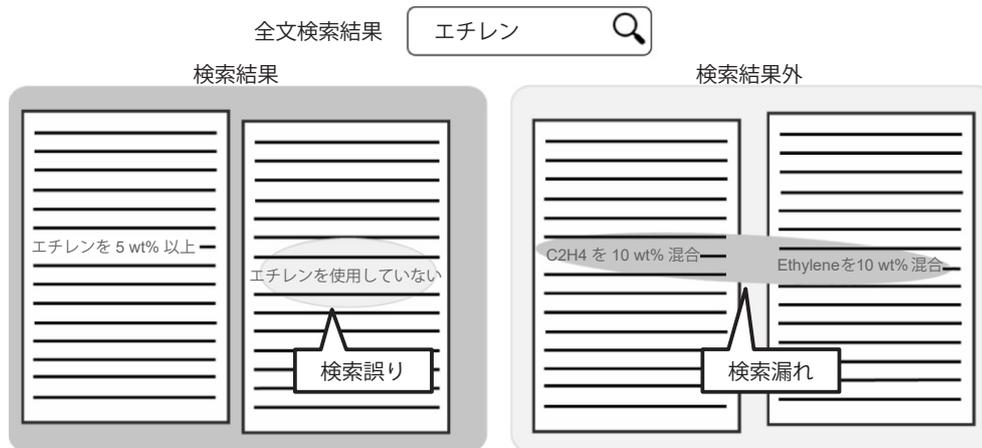


図3 報告書には実験や解析を再現するために必要な情報が書かれていない

百歩譲って、報告書で参照した実験の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しておく」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は hit してしまうし、「ethylene を 10 wt% 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 5 wt% 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でしょうか？



正しい検索結果以外に、検索誤りや、検索漏れが起きる

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 10 wt% 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えば、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たと思う。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みに基づき、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。実際、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、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。

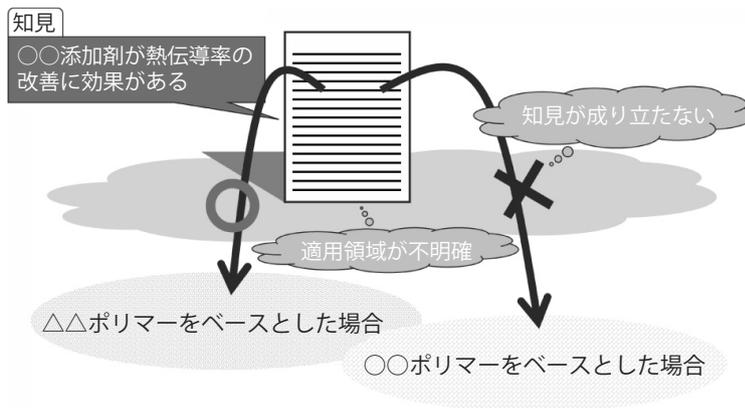


図5 報告書に書かれた知見や法則性は適用領域が不明確なものが多い

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はなく、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、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えば報告書を読む人は稀で、どちらかという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。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。

3. データ共有で研究の何が改善できるのか？

前項では、「報告書の共有で期待して良いこと、良くないこと」と題し、報告書共有の特徴に関して、説明を行った。本項では、「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「どんなことが改善できるのか？」、逆に言う「どん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、客観的なものだけということである。

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？どんな優秀な航海士もでたらめの海図と羅針盤技法では、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、

分析方法がでたらめでは、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（海図と羅針盤技法）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のである。「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。

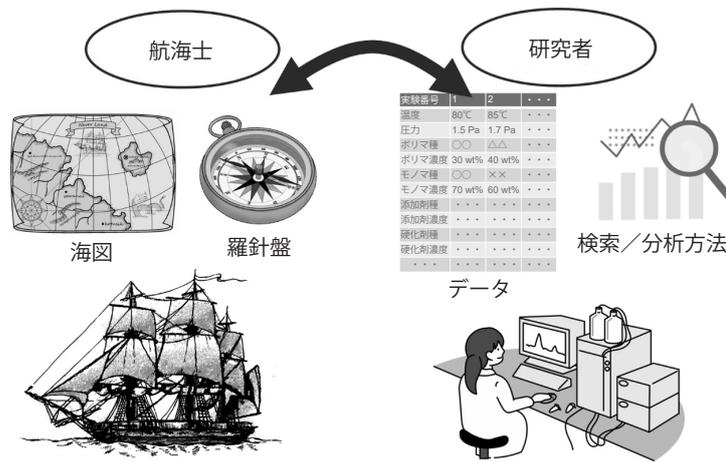


図6 研究者のデータと検索/分析手法は、航海士の海図と羅針盤と同じ

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だから、「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ふ。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、改善されたかどうかは判定しやすいはずである。

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だから、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、データ共有による改善で「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。

4. 実験データ蓄積基盤としての電子実験ノートの選び方

前項では、「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「どんなことが改善できるのか？」、逆に言うと「どんなことの改善は期待してはいけないのか？」に関して、説明を行った。本項では、「実験データ蓄積基盤としての電子実験ノートの選び方」と題し、電子実験ノートの種類と特徴及び実験データ蓄積基盤として気を付けるべきことを論じる。

電子実験ノートには、大きく2つの系列があり、まず、それぞれの特徴を把握しておく必要がある。

a) 報告書などの自由記述ベースの報告書型実験ノート

例) BIOVIA, E-WorkBook, E-Notebook 等

特徴: Wordなどに記録するのと同じ感覚で記録していける。

基本的には全文検索で、項目/数値条件絞り込みはできない。

b) 項目名-項目値ベースの項目型実験ノート

例) i-Reporter, R&D FormPat, RCM-SimpleLabReport 等

特徴: 項目名指定で、項目値の上下限值指定で検索が可能である。

WordやPowerPointに比べ記述、描画自由度が小さい。

(例えば、絵を貼る場所は予め決めておく必要がある)

報告書型実験ノートは様々な拡張機能がある。以下主要なものを列挙しておく。

あ) 化学構造式の描画支援

五員環、六員環など化学構造式を描画支援する機能

い) 化学物質の部分構造検索機能

化学構造の部分構造での化学物質の検索が可能となる機能

う) 自動処理機能

実験ノートに張り付けた実験生データなどのデータ処理を自動的に行わせる機能

え) データベース連携機能

実験ノートに張り付けたデータを外部データベースに登録し、データベースを使った項目名-項目値検索を可能にする機能

扱うデータのフォーマットやデータ項目の追加、変更がない場合は、え) のデータベース連携機能は、非常に便利で、簡単に項目名-項目値データをデータベースに登録し、項目名-項目値で検索可能にできる。しかしながら、R&D 部門の研究では、項目の追加、変更が頻繁であり、その変化に追従させながら、データベース連携機能で持続的にデータベースを拡張していくことは非常に難しい。

一般的に電子実験ノートなどの基盤システムにデータベース系の拡張機能があったとしても、業務システムとは違い、R&D 部門のシステムの場合、項目名追加や変更など R&D の変化に追従させるために様々な運用機構が必要となり、単純に拡張機能を使えば運用できるというものではない。例えば、実験パラメータを項目名-項目値で実験ノート内に入力フォーマットを作り込んだり、値を入力した Excel を報告書型実験ノートに貼り込んだりする形で、実験データを蓄積していくときに、「データベース連携機能」を使い、項目名-項目値検索を可能にしようとする場合を考えてみる。そもそも R&D 部門では、実験テーマ毎に共通の入力フォーマットを作るのは困難なので、実験ノート内に入力フォーマットを作り込む方法では、様々なフォーマットが乱立することになる。さらに、項目の追加、変更が頻繁にある場合、追加、変更毎にフォーマットがさらに増えていってしまうので、実質的に項目名-項目値のデータ共有システムとしては運用することは難しい。また、この方法では、異なるフォーマットの実験ノート間でデータ共有（横串の検索など）することも非常に困難になる。一方、値を入力した実験 Excel を実験ノートに貼り込む方法の場合、電子実験ノートのフォーマット自体は共通化できるので、データ共有範囲は広がる。ただ、Excel 内の項目名-項目値をデータベースに登録し、それを検索条件に指定して検索する機能以外に様々な機構を作りこむ必要がある。どのような機構が必要で、R&D の変化（項目の追加、変更していく）に追従するためには、どのようなことに注意しなければならないか分かるだろうか？

項目名-項目値のデータをデータベースに登録し、それを検索条件に指定して検索するだけなら「データベース連携機能」で比較的簡単に実現できるが、実際の実験環境を考える場合、それで十分なのだろうか？実はそれだけでは、全く十分ではない！今まで使っている実験ノートや実験データを記録している Excel は何度も値を追記や修正をするはずである。ここで言う、実験データの値の追記というのは、新しい実験のことではなく、ある実験に対して、後から追加計測をした結果を追記したり、記入漏れをしていた項目に追記をしたり、分析を進めていくことで分かってきたことをコメント項目などに追記することを意味している。「そういう追記がないようなところまで実験が落ち着いてから実験データは電子実験ノートに貼りつけてください」という運用をしているところも見かけるが、それは追記すべきことを追記しないままになったデータが蓄積されるだけになっているはずである。実験ノートというのは、リアルタイムに書き込めるから便利なのであって、紙の実験ノートがいまだに重宝されるのは、そのためである。電子実験ノートになり、いつでもどこにでも追記できるという利便性がなくなれば、紙の実験ノートの代替にはならない。

修正に関しても同じことが言える。「十分間違いがないことを確認してから入力すれば、修正は不要ではないか？」という人もいるようだが、修正が発生しないことを想定するデータベースというのは、恐らく存在しない。R&D より間違い入力が許されない経理、会計などのシステムでさえ、修正機能は提供されているのである。つまり、データ入力というのは間違いがあり得るので、修正機能を提供しないと正しいデータが入力されず、大きな問題が発生するということを意味している。まだ、「いや、実験データでは経理、会計などと違い、修正が不要だと言っていた研究所があった」と聞いたことがある人もいるかもしれない。それは、実験装置からの出力データを直接、データベースに取り込むところを自動化している場合であり、その場合は確かに修正不要である。しかし、人が入力する場合は、やはり、修正機能は必須なのである。修正機能を提供しないのであれば、すべての実験パラメータ、計測パラメータ、計測値を実験装置から出力させ、データベースに自動的に登録する仕組みを作り上げる必要がある。これは、加速器

や天文台のような大規模実験もしくは、生産ラインなどでは可能だろうが、それら以外では、現実的でなく、小規模な R&D 部門では実現できないことはわかっていただけと思う。

ちなみに、「データベース連携機能」以外の基本的な電子実験ノート部分（つまり、全文検索で実験ノートを探す機能）は、追記や修正がいつでも可能ではあるので、誤解がないようお願いしたい。ここで、指摘しているのは、実験データを項目名-項目値検索を可能にするために「データベース連携機能」を使って、電子実験ノートを拡張する場合の話である。

電子実験ノート自体の基本機能部の値（つまり、文章）を追記、修正するのは、該当部分を編集して、保存すればいいだけである。しかしながら、データベース連携機能を使って、項目名ごとの項目値を登録できるようにする場合、該当部分を編集して、保存すればいいだけにするのは非常に難しいのである。例えば、下表のような Excel を電子実験ノートに貼りつけると 1 行目の各列を項目名と考え、2 行目以降を対応する項目名の項目値として登録することを実現するのは、簡単である。

表 1

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

しかし、いったん登録してしまったデータに間違いが見つかり、引張強度の 3 行目の実験結果を 12 から 12.5 に変更する必要がある場合、「既にデータベースに格納されている多くの値の内どの値のものを変更すればいいか」をどう確定するかという問題にぶち当たる。ちなみに、1 回の登録だけでは、こういうことは考える必要がないのはお分かりいただけると思う。

「〇〇実験ノートの××ページの 3 行目、5 列目セル」という情報をデータベースに入れておけば、データベースのどの部分を変更すべきなのかが確定できそうだが、本当にそれでよいだろうか？実は、実験は 3 個でなく 4 個で、最初の実験を記録し忘れていたとする。実験順序通りに記録するルールにしているので、2 行目に書き洩らした実験を追加することがあるかもしれない。また、配合材料が 2 個から 3 個に増え、列が挿入される可能性もあるかもしれない。そうすると、初期登録時に各値に対して「〇〇実験ノート、××ページ、△△行目、□□列目」という情報をデータベースに入れていたとしても行も列も変わってしまうと、修正したいところと異なる部分を修正してしまうことになる。つまり、行数、列数が変動しうる場合、セルを行と列で指定してはいけないのである。R&D 部門で「行数、列数が変動させてはいけない」という運用が現実的でないことは、皆さんはよくお分かりだと思ふ。また、ページ番号も文書記述が増えるとずれるので、ページ番号指定もという運用も現実的でない。

そうすると、どうすればいいのか？記録用 Excel の最初の一行に実験 ID のようにユニークな番号をつけておくようにし、初期登録時に各値に対して「実験 ID」という情報をデータベースに入れておくようにすると良い。

表 2

実験 ID	原材料名 1	原材料 1 濃度	原材料名 2	原材料 2 濃度	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

こうしておくと「実験 ID が“EXP2”の引張強度項目のセル」ということで、実験ノート名、ページ番号、△△行目、□□列目が変わったとしても、データベース上のどの値を変更すべきかを特定できる。しかし、この方法は実験 ID が全電子実験ノートでユニークであることを要求するので、それをユニークであることを保証する仕組みが別途必要

である。また、付け加えると、当然だが他の項目名（引張強度等）自体もユニークでないといけな。このユニーク性は電子実験ノート側でどうにかなる問題ではないので、Excel 側でそれを担保する機構が必要ということになる。

電子実験ノートを運用していると実験データの項目値だけでなく、項目名を変更しなければならないこともよく起こる。例えば、上記「引張強度」という項目名だと様々な単位の値が入っていたり、項目値側に単位を書いてしまい、数値扱いできないものが混入してきたので、項目名を「引張強度 (N/m)」に変更すべきという状況は、よく起こる。データベース上の項目名は、「引張強度」を「引張強度 (N/m)」に変更すればよい。

表 3

実験 ID	原材料名 1	原材料 1 濃度	原材料名 2	原材料 2 濃度	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

ただ、過去に登録した実験ノートは、「引張強度」のままになっているはずである（電子実験ノートに添付したファイル内の文字列に関して、各電子実験ノートを開かず一括変換する機能は提供されていないし、提供されていたとしても Excel の項目名部分だけを変換することなんてできない）。この状態で、実験ノートを開き、「引張強度」を修正し忘れたまま、文章を追記、編集し、再登録しようとする、データベースにまた、「引張強度」項目ができてしまう。また、新しい実験ノートに貼り込む Excel においても、項目名を「引張強度」ではなく、「引張強度 (N/m)」に徹底する必要がある。項目名変更に対応するのが、如何に難しいかわかっていただけたらだろうか？

本項の最初の方で「データベースに登録するだけなら簡単」と言ったが、実際に皆さんが実験データ記録に使っている Excel のことを考えると、実は、そんなに簡単ではない。実際に皆さんが実験データ記録に使っている Excel は、下表のようにセル結合や項目名を大項目、中項目のように分けて書いているものもあるのではないだろうか？

表 4

乾燥工程				焼結工程		性能評価
1 段目		2 段目		温度	風速	引張強度
温度	風速	温度	風速	温度	風速	
10	25	42	7	14	29	8
15	27	47	18	19	31	10
20	29	52	31	24	33	15

注) 温度は℃単位、風速は m/ 秒単位の数値で記入すること。

大項目＝乾燥工程、中項目＝1 段目、小項目＝温度となっている場合、項目名、項目値で考える項目名は、大項目＋中項目＋小項目＝項目名＝乾燥工程 1 段目温度とするとルールを決めておけば、この Excel の書き方でもデータベースにデータを登録することは可能かもしれない。ただ、「乾燥工程」と「1 段目」を 1 行にして、「1 段目乾燥工程」と書く人がいたりすると、勝手にルールを考え出すプログラムというのはできな。データベースにデータを登録するときに項目名が正しく構成されなくなる。そういうルールを研究者が守れたとしても、そもそもあらゆる項目で、大項目＋中項目＋小項目＝項目名のようなルール設定が必要で、そのルール設定や新しい項目や中項目が増えるとそのたびに大きな手間が発生する。また、「注) 温度は・・・」という文字列部分も具体的に排除指定しないとこの文字列も 1 列目なので、温度だと判断し、登録されてしまいかねない。先ほどまでの例だと、項目名、項目値が整然と並んでおり、何を項目名とし、何を項目値とするか、簡単に指定できたので、「変更、更新でなく、新規登録なら簡単」と書いた。もちろんそれでも、Excel の何行目が項目名になるのか等の指定は、人がプログラム上で指定する必要がある。指定するにしても Excel のフォーマットが揃っていなければ、指定するプログラムは書けない。しかし、現実

の実験記録に使われている Excel は、構造が複雑というだけでなく、人やテーマによって、目視確認しやすいように書かれているので、これを項目名-項目値をプログラムで自動判断することは、ほぼ不可能である。つまり、実験記録に使われている Excel に関して、フォーマットを決めてしまわないといけないのである。

R&D では新しい材料やプロセスが日々発生するので、それに対応する項目名も日々増えていく点で、経理や総務などの事務系の固定フォーマットではダメで、そういう項目名の増大にプログラムが自動追従（項目名が増えるたびにプログラムを修正するのは、R&D の変化に追従できない）できるフォーマットでなければならないのである。それは、見やすさを優先する構造的な（大項目、中項目など）ものではなく、単純な項目名-項目値として、項目名を決めるべきということである。例えば、先ほどの例だと、最初から「乾燥工程 1 段目温度」、「乾燥工程 2 段目温度」、「乾燥工程 1 段目風速」・・・のような項目名にしておくということである。

材料名を項目値として入力するような Excel を使っていると上表のように C₂H₄, Ethylene などの表記揺れすることは、よく発生する。特に、他人の実験ノートまで考えると表記揺れしないほうが稀である。

表 5

実験 ID	原材料名 1	原材料 1 濃度	原材料名 2	原材料 2 濃度	引張強度
EXP 1	Ethylene	80	プロピレン	20	10
EXP 2	ブタン	75	C ₂ H ₄	25	12
EXP 3	C ₃ H ₆	60	ブタン	40	8

しかし、表記揺れしてしまうと検索やデータ分析が正しく行えなくなる。後述するが、実は材料名は項目値とするのは分析上、よろしくなく、本来は項目名化しておくべきである。表記揺れを許容できない項目値は、実験ノートに貼り込む Excel 自体に表記揺れを抑制するように項目値をドロップダウンで指定するなどの工夫が必要である。また、項目値の追加やスペルミスなどの修正が発生することもあるので、表記揺れを許容できない項目値の追加、変更機構も項目名と同じように備えておく必要がある。こういう話をすると、「表記揺れぐらいは、AI でうまく処理してほしい」という要望をよく聞く。こういう欲求を持つ方は、AI の特徴を誤解しているので、以下で少し説明をする。

最近、AI が強力になってきたので、「表記揺れぐらい AI でうまく処理できるのでは？」と思っている人が結構いる。結論から言うと、AI は、たくさんのデータで学習すれば、それなりの精度で表記揺れを統一語に変換してくれるが、100%間違いなく変換してくれることはまずない。100%近い精度が要求されるもの、また、手続きを厳格化することで100%近くにできるものは、AI に頼るべきではないのである。AI は、精度がそこまで要求されない対象や、そもそも手続きが厳格化できない対象で、威力が発揮されるのである。そもそも、実験データの表記揺れに関して、AI でそれなりの精度を出すためにはデータ数が足りなさすぎるので、絶対に踏み出してはいけない領域である。

ここまでの話を聞いていただいて、結局、扱うデータのフォーマットやデータ項目の追加、変更が頻繁な場合には、多くの考慮が必要であることはわかっていただけたと思う。考慮が必要な主要なものを以下でまとめておく。

- 1) データ入力や検索時の項目名指定を定義した項目名通りにできるようにすることが必要（例えば、全項目名からのドロップダウン）
- 2) 項目に付随する付属情報や項目の定義を一般アカウントではなく、項目管理者的なアカウントで登録できることが必要
- 3) 既登録の項目に付随する付属情報や項目の定義を登録できる仕組みと項目管理者的なアカウントのみで、それを修正できることが必要
- 4) 項目名自体を項目管理者的なアカウントで修正できることや、その修正に追従して自動的に GUI だけでなく、既登録データの項目名も自動的に変更されることが必要

5) 一般アカウントで、扱える項目に付随する付属情報や項目の定義で検索し、結果を表形式で表示できることが必要
ここで挙げたことは、考慮すべき事項のごく一部で、様々なケースを考えるともっとたくさんの考慮すべきことが出てくる。つまり、ある拡張機能を追加しようとする時は、単に提供された機能を組み込めばいいというわけではなく、上記のようにその機能を使うにあたって、他にどのような機構が必要かを、在りうる全てのケースを想定し、それらケースに対応でき、かつ維持可能な運用仕様に落とし込む必要があるということである。「拡張機能利用する場合、在りうる利用全ケースを考慮した維持可能な運用仕様に落とし込む」を実施するには、「在りうる全てのケースを洗い出し、どのような機構があれば問題なく運用維持が可能か」を明確化する能力が必要で、それはプログラム（例えば、Python プログラムが書ける）の能力ではなく、相当ハイレベルなシステムエンジニアの能力である。システムエンジニアというとソフトウェア分野の能力と思いがちだが、上記能力はソフトウェアとは関係なく、何か（例えば、旅費申請と旅費立替金の振り込み）の手続きをどうすれば漏れや間違いを起こさないように実施可能かの手順を構想、明確化しきる力なのである。これができる人が社内にとだけいられるだろうか？残念ながら、R&D 部門のデータ取り扱いの実態（項目が頻繁に追加、変動する）を知らない IT ベンダも、実はこれはできないのである。IT ベンダができるのは、項目がほとんど追加、変動されない事務系や生産ライン系だけなのである。

項目名（材料名）の追加変更が頻繁に発生するような R&D 領域では、報告書型実験ノートは、これまで説明したように拡張機能で項目名（材料名）の追加変更に対応させるのは、持続可能な利用、運用の読み切りが必要で、非常に困難である。したがって、報告書型実験ノートは、実験データはあくまで補足的な情報で、実験報告書の共有のためのものであるべきである。つまり、項目名-項目値で絞り込むのではなく、報告書の全文検索で絞り込み、実験データ自身は、その報告書を開いて確認するということである。項目名があいまいな部分は実験者に問い合わせるしかない。ただし、決まった計測条件での定点計測などの項目名の追加変更がないものばかりの実験や特定のテーマで項目名の追加変更がないテーマに関しては、その限りではなく、データベース連携機能で項目名-項目値をデータベース化し、項目名-項目値で絞り込むことも実現可能である。

項目名-項目値ベースの実験ノートは、項目名（値）をデータベース登録、利用することを想定しており、項目名（値）及びその追加変更に関して、「運用に必要な作業やそれを持続するための仕組み」は考慮されている。しかし、項目ごとに入力セルに値を入力する必要があり、入力項目数が多い場合には、非常に手間がかかる。Excel で Copy&Paste で複数項目を一気に入力したり、他のセルを参照し、複雑な数式で値を自動計算させたりすることに慣れている人には、入力自体が相当大変である。したがって、項目名-項目値ベースの実験ノートは、項目名追加変更が想定されているので、実験データが中心的な情報の実験データ共有は可能だが、項目数が多いと入力が煩雑になるため、数十項目を超える実験の記録をするのは適切とは言えない。

このように報告書型実験ノート、項目型実験ノートは、それぞれ一長一短なので、目的に応じて、どちらを導入するか判断すべきである。

5. 電子実験ノートでデータ蓄積、共有が難しい実験

前項では、「実験データ蓄積基盤としての電子実験ノートの選び方」と題し、報告書型実験ノート、項目型実験ノートの特徴および提供されている拡張機能を使う場合の注意点に関して、説明を行った。本項では、「電子実験ノートでデータ蓄積共有が難しい実験」と題し、報告書型実験ノート、項目型実験ノートで、データ蓄積共有が難しい実験に関して、どのように蓄積、共有していくべきかを論じる。

前項で説明したが、決まった計測条件での定点計測などの項目名の追加、変更がないものばかりの実験や特定のテーマで項目名の追加、変更がないテーマに関する実験は、報告書型実験ノートでデータ蓄積、共有が可能である。一方、実験データの項目名の追加、変更はあるが、項目総数が数十項目内で収められる実験は、項目型実験ノートでデータ蓄積、共有が可能である。逆に言うと項目名の追加、変更はあるが、項目総数が数十項目を超える実験は、電子実験ノートでのデータ蓄積、共有は難しいということになる。

実は、それ以外にも電子実験ノートでのデータ蓄積、共有が難しい状況がある。それは、複数段の工程による実験

である。例えば、合成工程でエチレン、プロピレンなどを混ぜてポリマーを合成し、配合工程でそれらポリマーの短鎖、長鎖の2種配合し、配合材料を作成するような実験があったとする。そして、それは確かに下表のように合成工程の Excel と配合工程の Excel で記録できる。

表 6

合成 ID	エチレン濃度	プロピレン濃度	ブタン濃度	引張強度
Poly 1	10	90	0	100
Poly 2	0	30	70	150
Poly 3	20	30	50	175

表 7

配合 ID	短鎖ポリマー	短鎖ポリマー濃度	長鎖ポリマー	長鎖ポリマー濃度	粘度
Sample 1	Poly 1	10	Poly 2	90	102
Sample 2	Poly 1	30	Poly 2	70	203
Sample 3	Poly 1	20	Poly 3	80	304
Sample 4	Poly 2	20	Poly 3	80	403
Sample 5	Poly 2	60	Poly 3	40	505

この時、ある配合材料の短鎖ポリマーのエチレン濃度を知りたいときにどうするのだろうか？配合工程の Excel において、当該配合材料で使われている短鎖ポリマーがどの合成工程の実験で作られたものか、つまり、合成 ID を確認し、その後、合成工程の Excel において、その合成 ID の合成樹脂でどの程度の濃度のエチレンを合成に使ったのかを確認すれば、情報は得られる。しかし、知りたい「ある配合材料」というのが 100 個あれば、上記 2 つの Excel を行き来する作業を 100 回行わなければならないわけである。そして、当然そこで得た情報は覚えていられないので、別 Excel に短鎖ポリマーのエチレン濃度などの情報を一時記録しておくに違いない。そして、100 回も調べるときと何回かは間違っ一時記録してしまうものもあるはずだ。短鎖ポリマーのエチレン濃度は一時記録しておくといったが、後で他の量との比較やグラフを描いた分析をすることを考えると配合工程の Excel を複製したものに短鎖ポリマーのエチレン濃度の項目を追加するのが、便利なはずである。

実は、同じように短鎖ポリマーのプロピレン濃度やブタン濃度、調査ポリマーのエチレン濃度やプロピレン濃度やブタン濃度も同じようにして、配合工程の Excel を複製したものに項目追加しておけば、その Excel 一つで他の量との比較やグラフを描いたりし、分析をすることは可能である。それなら最初から工程を分けずに 1 つの Excel で記録すればいいのではないかとと思うかもしれない。

しかし、1 つの Excel で記録しようとする合成工程で大量に作ったポリマーをたくさんの配合工程で少しずつ配合比率を変えて行う実験を行う場合、固定であるはずの当該ポリマーのエチレン濃度、プロピレン濃度、ブタン濃度を配合実験毎に同じ値を記入していく必要がある。また、配合は、短鎖ポリマーと長鎖ポリマーを配合するが、エチレン濃度、プロピレン濃度、ブタン濃度と書いてしまうとどちらのポリマーの組成のことを示しているのかわからなくなるため、短鎖ポリマー内エチレン濃度、短鎖ポリマー内プロピレン濃度、短鎖ポリマー内ブタン濃度、長鎖ポリマー内エチレン濃度、長鎖ポリマー内プロピレン濃度、長鎖ポリマー内ブタン濃度のような項目にして、記録するようにしなければならない。実際、エチレン濃度、プロピレン濃度、ブタン濃度という項目名だけでは不十分で、何のポリマーの組成の濃度かの情報がなくてデータ分析などはできないのである。

実際、「実験の独立変数（実験パラメータ）は何ですか？」と聞かれた場合、合成工程と配合工程の実験パラメータを列挙して、エチレン濃度、プロピレン濃度、ブタン濃度、短鎖ポリマー濃度、長鎖ポリマー濃度の 5 変数と答えてしまいがちだが、それは間違いである。「実験の独立変数（実験パラメータ）は何ですか？」の正確な答えは、短鎖ポリマー濃度、短鎖ポリマー内エチレン濃度、短鎖ポリマー内プロピレン濃度、短鎖ポリマー内ブタン濃度、長鎖ポリマー濃度、長鎖ポリマー内エチレン濃度、長鎖ポリマー内プロピレン濃度、長鎖ポリマー内ブタン濃度の 8 変数

である。つまり、記録の煩雑さや誤記を減らそうとすると合成工程と配合工程を別々の Excel で記録していく方がいいが、データ絞り込みやデータ分析を行うために独立変数というものを意識しなければならない場合は、配合工程に合成工程を繰り返した 1 つの Excel で記録する方が良いのである。

表 8

配合 ID	短鎖ポリマー濃度	短鎖ポリマー内エチレン濃度	短鎖ポリマー内プロピレン濃度	短鎖ポリマー内ブタン濃度
Sample 1	10	10	90	0
Sample 2	30	10	90	0
Sample 3	20	10	90	0
Sample 4	20	0	30	70
Sample 5	60	0	30	70

表 9

長鎖ポリマー濃度	長鎖ポリマー内エチレン濃度	長鎖ポリマー内プロピレン濃度	長鎖ポリマー内ブタン濃度	粘度
90	0	30	70	102
70	0	30	70	203
80	20	30	50	304
80	20	30	50	403
40	20	30	50	505

工程ごとの項目を分けて記録をすると見かけ上、独立変数が減ってしまうので、実際データ分析をするときは、この 8 項目に展開しないと X-Y Plot さえ描けないのである。当たり前だが、X-Y Plot を描くということは、データが単純な 2 次元表形式になっている必要がある、2 つの表のようなものは、そのままでは X-Y Plot さえ、描けないのである。複数工程実験を複数の表で記録している人は、X-Y Plot を描くたびに複数の表を一つの表にまとめる作業をしているはずである。ただし、全ての項目を一つの表にするのは大変なので、X-Y Plot で使う、X 軸、Y 軸、凡例だけを取り出して、一つの表にまとめる作業をしていると思う。この作業が面倒な人は、良く使う項目だけをマクロを使って、自動的に一つの表にまとめる機構を作っている人もいると思う。しかし、このようにすることで、それ以外のデータ絞り込み、データ分析はより億劫になってしまい、分析範囲が無意識に狭まってしまうので、実は、研究としては望ましい状況ではない。

もちろん、最初から短鎖ポリマー濃度、短鎖ポリマー内エチレン濃度、短鎖ポリマー内プロピレン濃度、短鎖ポリマー内ブタン濃度、長鎖ポリマー濃度、長鎖ポリマー内エチレン濃度、長鎖ポリマー内プロピレン濃度、長鎖ポリマー内ブタン濃度という項目名の 1 工程実験として、データを記録すれば問題ないのだが、通常、そういう記録は煩雑で好まれないはずである。また、2 工程程度ならまだいいのだが、5、6 工程にもなれば、項目名が幾何級数的に増えることになってしまい、恐らく正しく記録していくこと自体が困難になる。また、多工程実験のデータでは、項目名の揺れを抑制するだけでなく、工程間の紐づけ（短鎖ポリマー濃度のポリマーは合成工程のどのポリマーに対応するのか？）揺れ抑制も重要となってくる。つまり、下表のように工程間の紐づけに揺れがある記録では、データ絞り込みや分析が正確にできないので、工程間の紐づけ揺れの抑止機構も必須である。

表 10

合成 ID	エチレン濃度	プロピレン濃度	ブタン濃度	引張強度
Poly 1	10	90	0	100
Poly 2	0	30	70	150
Poly 3	20	30	50	175

表 11

配合 ID	短鎖ポリマー	短鎖ポリマー濃度	長鎖ポリマー	長鎖ポリマー濃度	粘度
Sample 1	Poly 1	10	Poly 2	90	102
Sample 2	Poly 1	30	Poly 2	70	203
Sample 3	ポリマー 1	20	ポリマー 3	80	304
Sample 4	Poly 2	20	Poly 3	80	403
Sample 5	Poly 2	60	Poly 3	40	505

つまり、データを共有し、利活用するためには、記録は工程ごとの Excel で、データ絞り込みや分析はそれらを 1 工程に繰り込んだ Excel で行える必要があるのである。そして、複数工程を 1 工程に繰り込むときには、工程間の紐づけ情報を分析し、項目名を内部で自動生成する機構が必要であり、そういう機構がないと複数段工程のデータ検索、分析はできないのである。データベースはそういうことを内部で行えるようになっており、もうひとつのデータベースの大きな利点である。つまり、データ共有、利活用を考えると「複数の人が同時編集及び誤入力、誤編集リスク抑止」、**「多工程データの紐づけと 1 工程への繰り込み」**が最初から考慮されているデータベースを基盤に据えることが適切だということである。ちなみに、Excel でそういうことをできる仕組みを作り込むことも不可能ではないが、これら機構を Excel に作り込むというのは、ある意味、Excel をベースとしてデータベースを開発するのと同じことになり、車輪の再発明になるため費用対効果やソフトウェア品質の観点から止めておいた方が良い。

先ほど話をした、多工程データを 2 次元表に射影するには、どのような機構が必要なのかを少し説明しておく。例えば、モノマー調整工程、樹脂合成工程、配合工程、測定工程の 4 工程からなる実験を考える。例えば、測定工程の配合材料において、内部で使われているモノマーの分子量を表記したり、条件指定したい場合、同じモノマーが異なる樹脂に使われていたり、同じ樹脂が異なる配合材に使われていたりすると、計測に使った配合材内のどのような樹脂に使われたモノマーの分子量ということを指定できる方法（指定できる項目名）が必要になる。したがって、『計測で使われた計測使用配合材 ID = “F01” の中で使われた配合内樹脂 ID = “P02” の中で使われた樹脂内モノマー ID = “M03” の分子量』と指定することで初めて、項目として意味のあるものになるのである。したがって、以下のような項目名を自動生成し、表現できる必要があるということになる。

計測使用配合材 |F01| 配合内樹脂 |P02| 樹脂内モノマー |M03| 分子量

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, 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