「株式会社技術情報協会 月刊 研究開発リーダー 2024年9月号」掲載記事
URL:https://www.gijutu.co.jp/index.htm

◆連　載◆

《隔月連載 全 5 回》

第 2 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

5　機械学習などの MI の特性と注意すべき点

前章では,「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。本章では，データ共有自体の話から少し外れて，「機械学習などの MI の特性と注意すべき点」と題し，皆さんが興味あるであろう機械学習などの MI の特性とそれらを利用する上での注意すべき点に関して，解説をする。

機械学習，AI という言葉で万能の打ち出の小槌が手に入るように思われているかもしれないが，一般的に非常に多くの過去データと AI 最適化のための設定値調整が必要であるし，必要とされる精度，推定力によって現実的に使い物になるかどうかも変わってしまう。

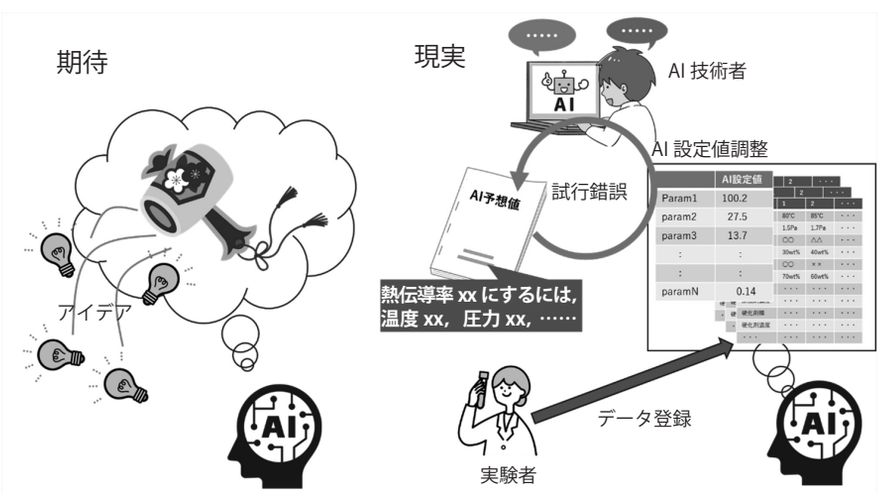

図 8　機械学習，AI は打ち出の小槌ではない

例えば，顔認証の機械学習を現在のような使いものになるレベルにするためには，数十名規模のプロジェクトで数百万以上のデータと数年以上に及ぶ機械学習モデル及び設定値の調整が必要だったのである。飛行場でのバードストライクを判定する画像認識は，羽田空港で使えるレベルのチューニングができても，それをドゴール空港で使用すると，誤判定が多すぎて使い物にならなく，結局，各飛行場でのチューニングが必要なのが現状である。「顔認証」も「バードストライク判定」も，どのレベルの精度が必要か？という問題が実用性に大きく絡んでおり，人を雇って「判定」をしたほうが安く，精度が高いようであれば，機械学習導入は見送られる。

材料開発のような分野では，恐らく，実験パラメータ数（材料種数，配合，焼結等のプロセスに関する設定値）は数百～数千に上るので，ある領域では期待する予測精度が出るかもしれないが，顔認証のようにあらゆる対象の判定は難しいはずである。また，空港の場合は，どの空港というのは簡単に指定できるが，材料開発の場合，そもそもこの適用領域というものをしっかり指定（定義）すること自体が難しい問題なのである。

この領域定義がしっかりしていないと，「利用者にこういう領域はこの機械学習は高い精度を有するので使っていいですよ」とアナウンスさえできないのである。この適用領域定義を明確にしないまま利用開放をすると，機械学習を使った方が使わない時より，材料開発が遅延してしまいかねない。また，利用者がアレルギー反応を起こしてしまい，機械学習を避けるようになってしまう可能性もあるので，適用領域の明確化は非常に重要である。ただ，R & D 部門のように広大な対象が広大なパラメータ領域で，常に未知の領域への挑戦を続けるような部門では，概して，必要とされる精度，推定力を確保できる適用領域が確定したころには，その領域は研究対象でなくなっている可能性も高い。逆の言い方をすると，精度，推定力を確保できる適用領域を確定するためには，当該領域で多くの実験が必要で，これは研究において，「その狭い領域において，機械学習の性能確認のためにそんなに多くに実験を行うことが，研究加速につながるのか？」という，悩ましい問題でもある。

適用領域の話は一旦置いておいて，ここでは機械学習と従来研究方法の関係に関して，考察する。機械学習は予測結果を返すが，なぜそのような結果になるのかの理由を提示しない。一方，研究というものは，その理由（仮説，理論）を考え，検証をしていくものである。

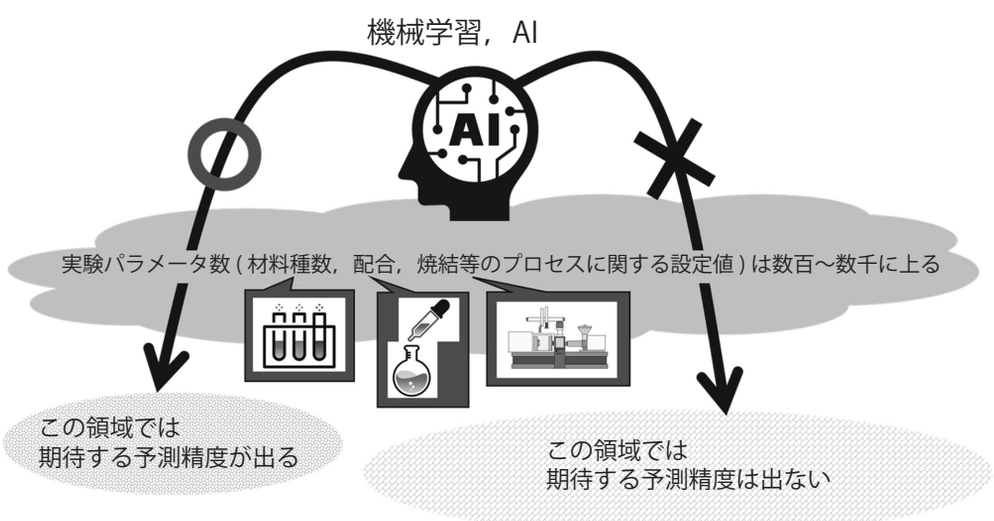

図 9　機械学習，AI は期待する精度の出る領域の明確化が難しい

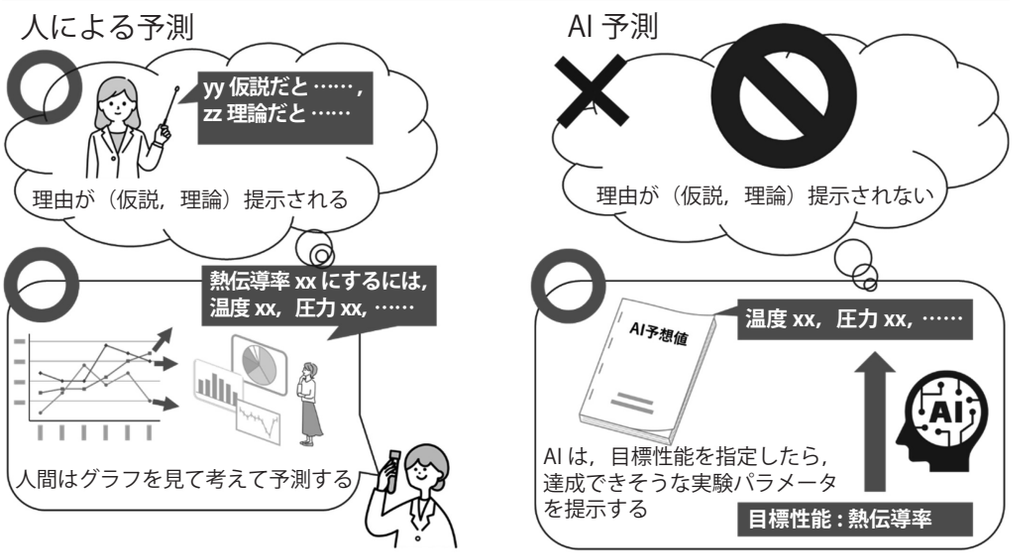

図 10　機械学習，AI は予測値の理由を提示するのが苦手

また，前述した通り，機械学習の結果が常に正しいわけではないので，それがどれぐらい正しそうかを研究者の独自の方法でデータ分析をし，その理由（仮説，理論）検討を行うことはやめてはいけない。それらの検討を踏まえ，最終的に機械学習が示す実験を行うかどうかの判断は，研究者自身が行うべきなのである。

機械学習の精度が十分に高い領域が確立すれば，その領域の材料開発では，仮説，検証のためのデータ分析は不要になるかもしれないが，研究者はデータ分析を常に行っておくべきである。そうしないと，研究者のデータ分析能力が劣化し，機械学習の精度の低い領域の材料開発ができなくなってしまう。実際，機械学習の精度が十分に高い領域が確立されたというのであれば，そこは，データ分析ができる研究者に材料開発をさせるのではなく，実験作業に特化した作業員を請負や派遣で実施するのが良いはずである。そして，そのような領域は将来的には機械学習と連動した自動実験に移行することも可能ははずである。

図 11　機械学習，AI の精度の高い領域は，自動実験へ移行していくべき

ここまでで説明したように，機械学習は，蓄積データがあれば大きな力を発揮しうるが，評価，運用には注意が必要である。実際に，機械学習に関して気を付けるべきことを最後にまとめておく。

1）機械学習は万能の打ち出の小槌ではなく，非常に多くの過去データと設定値調整が必要
2）機械学習が適用できる領域を広げようとすると実現が遠のく
3）どのような材料開発で機械学習が有効か明確にしないと反って開発を遅延させてしまう
4）機械学習を利用しても機械学習の結果のデータ分析をし，その理由（仮説，理論）検討を行うことはやめるべきではない
5）機械学習の精度が十分に高い領域は，実験作業に特化した作業員を請負や派遣で実施するのが良い

6 機械学習などの MI の研究への組み込み方法

前章では,「機械学習などのMIの特性と注意すべき点」と題し，機械学習などの MI の特性と注意すべき点に関して，説明を行った。本章では,「機械学習などの MI の研究への組み込み方法」と題し，機械学習等を研究で活用するときに発生する様々な問題や研究者側のアレルギー反応などをできるだけ抑えられるような研究スタイルに関して，考察をする。

前章で，材料開発の場合，機械学習等の MI が必要とされる精度，推定力を確保できる適用領域を確認すること自体が難しい問題であること，そして，この適用領域定義を明確にしないまま利用開放をすると，機械学習を使った方が使わない時より，材料開発が遅延してしまいかねないだけでなく，利用者がアレルギー反応をしてしまい，機械学習を避けるようになってしまうかもしれない点を指摘した。

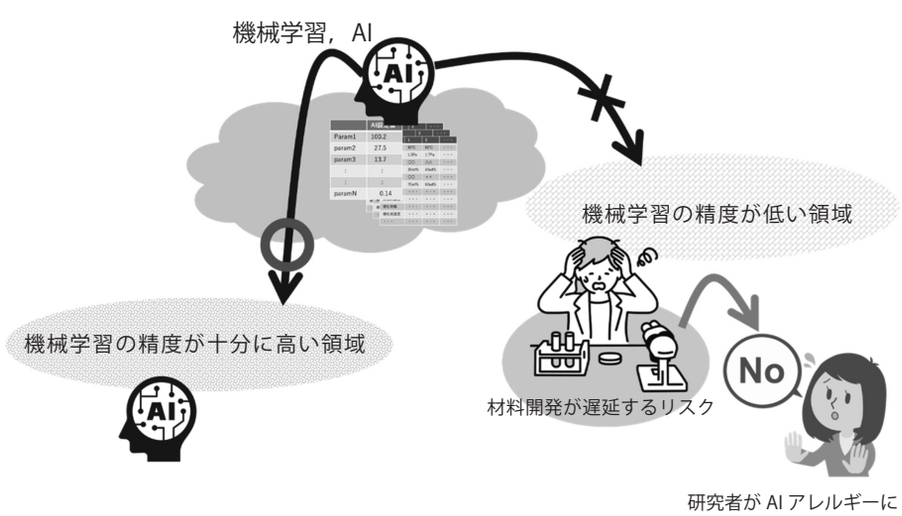

図 12　機械学習，AI の精度の低い領域を明確化しておかないと研究者が AI アレルギーに

しかし，難しい，難しいと言っていただけでは，いつまでたっても研究に機械学習や MI を導入できない，もしくは，研究を阻害する形での導入になってしまうことになる。

機械学習などのMIを導入しようとしている部門では，目的生成物の性能を向上させるための実験パラメータを提示させる形で機械学習などの MI を使おうとしていることが多い。しかし，この形での導入では，必要とされる精度，推定力を確保できる適用領域が明確でない限り，前述で指摘したように様々な問題が生じる。また，機械学習などの MI から提示された実験パラメータに単に従って実験してはダメで，提示された実験パラメータがどれぐらい適切かを研究者の独自の方法でデータ分析をし，その理由（仮説，理論）検討を行う必要があるが，実はそれも結構難しい。それでは，実際の研究現場には，どのようにして機械学習などの MI を導入するべきだろうか？

実際の研究現場に機械学習などの MI を導入するためには，大きく二つのハードルがある。まず，一つ目のハードルは「必要とされる精度，推定力を確保できる適用領域が明確でない」ことを前提にしなければならない点である。そもそも「必要とされる精度，推定力」＝「従来の方法で研究者が生成物の性能改善に要した試行錯誤回数」を明確化すること自体，難しいはずである。実際，性能改善のケースごとのばらつきが多く，平均値に意味があるわけでないので，性能改善難易度毎の分類が必要だが，その分類自体相当困難であり，「必要とされる精度，推定力」を明確化することは諦めたほうが良い。もう一つのハードルは，「機械学習を使うべきでない領域に使ってしまうと，材料開発が遅延して，利用者がアレルギー反応をしてしまう」ことである。これは，「適用領域を明確化する」ことと等価だが，そもそもその前提の「必要とされる精度，推定力」を明確化すること自体が困難で，もし，それができたとしても，「適用領域」というものを厳密に定義することも同じぐらいに困難で，利用者に悪影響をもたらすという問題である。したがって，機械学習を使った方が使わない時より，材料開発が早くなるようにできればいいのだが，機械学習の開発者，メーカーとしては，それは保証できないはずである。「必要とされる精度，推定力」とその「適用領域」を明確にしないまま機械学習などの MI を利用しても，材料開発が現状よりも遅延しなければ，アレルギー反応は起こらないはずである。例えば，以下のようにすればアレルギー反応は抑えられる。

a）まず，最初は，従来の方法で，つまり，過去データと属人的経験則も生かしながら自力で試行錯誤する。

　この方法の試行錯誤が煮詰まってきたと感じたら

b）次に，実験計画法などの統計的手法に則った実験及び解析を行なう。

　この方法でも要求される性能改善に到達しなければ，

c）最後に，機械学習などの MI に頼ってみる。

機械学習，AI を効果的に使った新しい研究の進め方

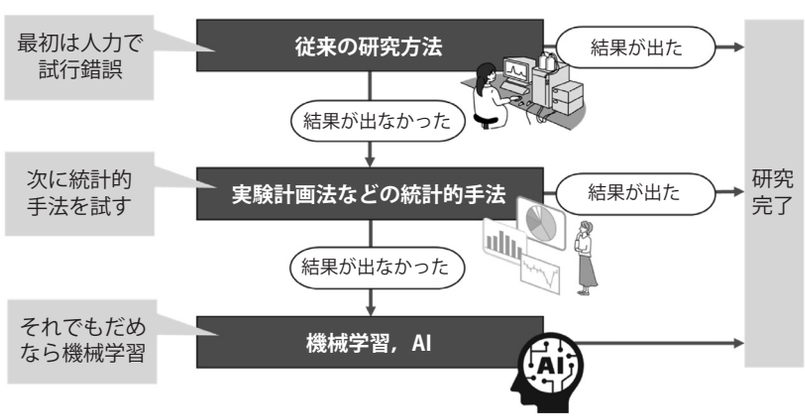

図 13　AI アレルギーを起こさない機械学習，AI を使った研究の進め方

この手順であれば，すでに自力での試行錯誤を十分行った上で，それも自力では解決できなかったわけなので，機械学習などの MI で解決できなくともアレルギー反応は起きないはずである。しかし，この方法だと，今までの 3 倍も試行錯誤が増えてしまう可能性がある。それを抑制するために試行錯誤を開始する前に，開発期限などを参考にして，試行錯誤に掛けられる最大試行回数を最初に見積もっておき，自力，実験計画法，機械学習などの MI に 1/3 ずつ割り当てるようにするとよい。ある意味，今までは自力で何とかするしかなかったので，試行錯誤が煮詰まってきたと感じても，無駄と感じながら試行錯誤をやめられなかったのだが，この方法ではあっさりギブアップが許されるのである。そして，後半は，ある意味自分の責任ではなく，実験計画法や機械学習などの MI の責任なので，少しは気持ちは楽になれるはずであり，実験計画法や機械学習などの MI に感謝の念を抱けるかもしれない。

実は，機械学習などの MI は，もう少し，活用範囲を広げることができる。a,b）を実施するときに，実験前に機械学習などの MI にどんな結果になりそうか聞いてみるのである。

a）まず，最初は，従来の方法で，つまり，過去データと属人的経験則も生かしながら自力で試行錯誤する。

b）次に，実験計画法などの統計的手法に則った実験及び解析を行なう。

a）ではある程度，良い結果が得られそうな算段がある実験パラメータの実験を計画しているはずである。その自分の判断と機械学習などの MI の判断を突き合わせるのである。自分の判断と機械学習などの MI の判断が近ければ，自分の判断にも自信がもてるだろうし，機械学習などの MI に対しても，「なかなかやるな」という感情を持てるのではないだろうか？実際の実験の結果が予想通りであれば，自分と MI で称えあい，予想とは全く違う結果でも「お互いまだまだだな」と慰めあえる。自分の判断と機械学習などの MI の判断が全く違うようであっても，a）の段階では MI の判断に従う必要はない。実際の実験の結果が自分の予想通りであれば，MI に対して，「やっぱりお前はまだ半人前だ」と言い放てばよく，MI 予想の方が近ければ，真摯に自分の負けを認め，MI の優秀さを認めることができると思う。どうだろうこの方法であれば，MI が愚かでも，賢くでも MI アレルギーは起こらないし，それどころか同じ研究に対峙する同士として，親近感さえ覚えられるようになるのではないだろうか？

b）の時では，自分が計画した実験パラメータではないので，どんな結果になりそうかは，a）に比べると予想できないことが多いはずである。ただ，自分なりの考えで，どのような結果になるか推測することは，推測力を鍛えるためにもよいことなので，是非，推測を試みてほしい。そして，MI にもどんな結果になりそうかは聞いてみて，a）の時と同じように一喜一憂してみればよい。こうやって，自分の推測と MI の推測を常日頃突き合わせることで，お互いの強い部分と弱い部分がわかるようになってきて，お互い頼れる相棒になっていけるのである。実際，相手がいることで，自分の弱い部分が意識できるようになり，単に弱いということがわかるだけでなく，改善もされていく。MI も自分の推定が当たったり，外れたりした結果をどんどん取り込んでいくので，改善されていく。

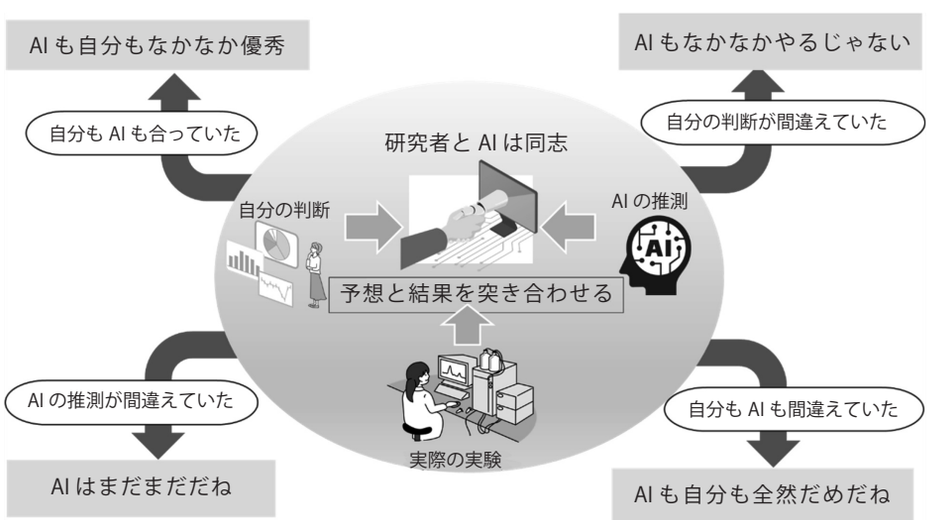

図 14　機械学習，AI との理想的な付き合い方

この話は，単なる物語として，書いているのではない，将棋の世界では，まさにこれと同じようなことをして，AI と対話をしながら，自分の弱いところ，AI の弱いところを意識し，改善していくことで，双方，強くなっているのである。研究の世界では，将棋のようにまだ AI が圧倒する状況ではないが，AI との付き合い方には学ぶところが多いはずである。

7 R & D 部門におけるデータ蓄積を Excel で行えるか？

前章では，「機械学習などの MI の研究への組み込み方法」と題し，機械学習などの MI をどのように研究に組み込んでいくべきか？について，説明を行った。本章では，話をデータ共有に戻し，「R & D 部門におけるデータ蓄積を Excel で行えるか？」と題し，R & D 部門においてデータを Excel で蓄積していくことが可能か？，課題があるのであればどんな課題があるのか？を解説する。

R & D 部門におけるデータ蓄積が他の部門のデータ蓄積と大きく異なるのは，項目（名）の多さと項目（名）の追加，変更頻度の高さである。十数名レベルの R & D でさえ，全実験パラメータ，計測パラメータ及び計測値すべての項目を数え上げると数千項目に上ってしまう。また，数週間毎に，頻度が高い場合は毎日のように項目（名）の追加，変更が必要となる。R & D 部門では，常に新しい材料を評価し，新しいプロセスを考案し，日々試作物の性能向上を目指すという業務の特性上，致しかたなしのことである。通常の事務系や営業系や生産ライン系の業務では，こういうことは起こらない。そういう意味で，データ蓄積，共有が一番難しい部門と考えて，間違いではない。

R & D 部門におけるデータ共有，利活用がうまくいっていない第一の理由は，本連載の最初に話をした「第三者が実験及び解析を再現するために必要な情報を記録」がなされていないことである。ただ，データ共有に挑戦しようとした組織では，この問題はクリアされおり，問題はこれだけではないことは明らかである。データ共有，利活用がうまくいっていない第二の理由は，項目名が統一されていないことである。項目名の統一には，利用者皆での合意形成と項目名の定義辞書のようなものが必要ではあるが，それができるのであれば Excel 程度でも項目名が統一の仕組みを作ることは可能である。

研究者皆が参照可能な共有ファイルサーバに項目名を列挙したテキストファイルを置いておき，実験データを記録する Excel には，この項目名を列挙したテキストファイルを読み込み，項目名のドロップダウンリストを自動生成する仕組みをマクロで組み込めばいいのである。そして，ファイル保存時に項目名として認められない（項目名を列挙したテキストファイルに存在しない = ドロップダウンリストに存在しない）項目名を書き込んでいたら Excel を保存できない仕組みにしてしまえばよい。項目名が追加，変更された場合は，項目名を列挙したテキストファイルを変更すればそれを参照する全 Excel も項目名が追加，変更される。そうであるならば，「Excel だけでデータ蓄積は問題ないのではないか？」と思われるかもしれない。それは，半分正解で，半分は間違いである。

研究者が各自の Excel に実験データを書き込む運用の場合，研究者ごと別ファイルになるためデータを探すときに結局すべてのファイルを開いて確認するしかない。また，ファイルごとの項目並びも一定でないので，1 ファイルへまとめることも困難で，データの蓄積はできてもデータを探すのは難しい。結局，自分のデータだけの利活用は進むが，他者のデータを含むデータ共有，利活用は進まない。

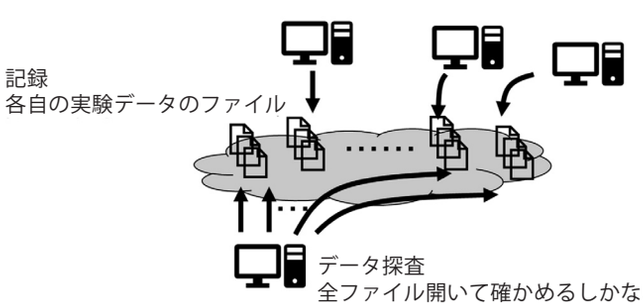

図 15　研究者が各自の Excel に実験データを書き込む運用のイメージ

共有ファイルサーバに 1 個のマスター Excel を置き，研究者がマスター Excel に実験データを直接書き込む運用のケースもある。その場合，マスター Excel には同時に書き込めないので，書き込み待ちが頻繁に発生する。研究者は，実験をしながら入力をしたいはずであるが，それをすると長時間マスター Excel を占有することになるので，行えず，一時的に個人の Excel に実験データを入力し，後からマスター Excel へ書き込むようにすべきだが，結局，面倒でマスター Excel への書き込みをしないままになってしまうことが多い。また，その使いにくさを避けるために，マスター Excel を自 PC にコピーし，自 PC で入力後，共有ファイルサーバのファイルを上書きする人も出てくるのだが，多くの人がそのようなことをしてしまうと，一部データが欠損しまったり，複数のマスターファイルができて，収拾がつかなくなってしまう。

Sharepoint を使えば，マスター Excel には同時に書き込みは可能になり，上記状況は幾分改善するが，それでもマスター Excel に自分の実験データを入力したり，転記するのは，思った以上に大変である。R & D だと 1 実験当たりの項目数が百近くになり，そういう実験が数百集まると実は数千もの項目になる。マスター Excel を作り，そこに入力するということは，列数が数千もの Excel から入力する百程度の項目を探しだし，入力しなければならないのである。また，変更する必要のない項目が数千もあり，常に表示されているので，誤入力，誤編集リスクも大きく，他者の実験データを間違って上書きしてしまうことも頻繁に発生する。さらに，データ蓄積が進んでくると Excel を開いたり，編集する処理がどんどん重くなり，使用に耐えられない状況になってしまうのである。

「変更や確認する必要のない実験（行）や項目（列）を表示せず，変更，確認する実験や項目のみ表示し，複数の人が同時に値を変更できるようにすることで誤入力，誤編集リスクを抑制する」は，事務系業務でも必要なことである。実は，事務系業務で，これらを実現するためにデータベースというものが作られたのである。

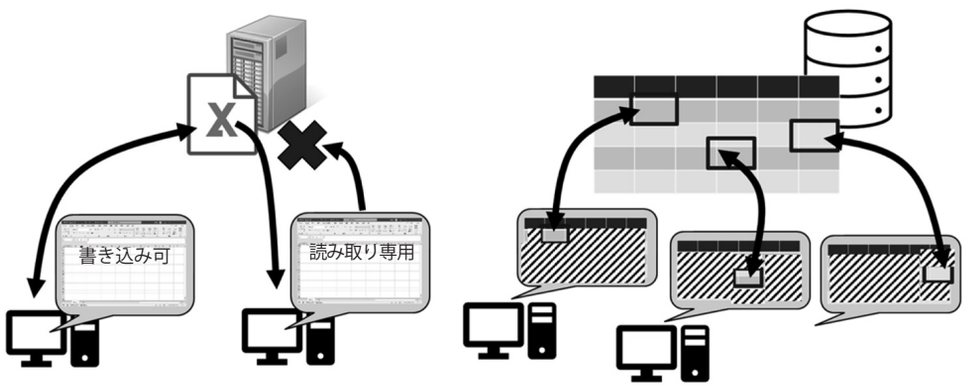

図 16　研究者がマスター Excel に実験データを直接書き込む場合のイメージ

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