「株式会社技術情報協会　月刊　研究開発リーダー　2022年2月号」掲載記事
URL:https://www.gijutu.co.jp/index.htm

◆連　載◆

《連載 全 3 回》

R & D 部門における データ共有システム構築の事前準備の要所（中）

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

4　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，「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。

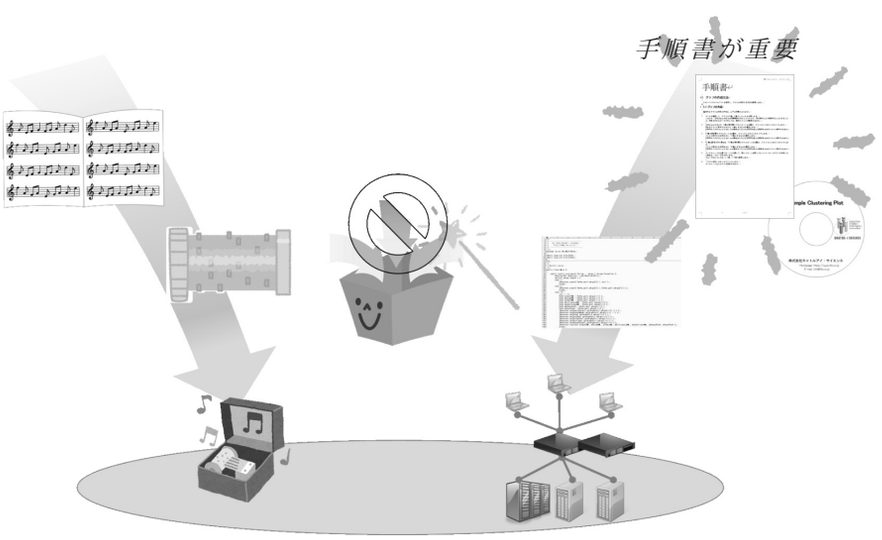

図 7　システムを作るには，データ探査とデータ処理の手順書が必要

「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」を作ることは，簡単なようで，なかなか難しい。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，自転車の補助輪をどのようにして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，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収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，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。

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かったりしていると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）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，システムでも実験データファイルを適当な場所に置かれるとそれを見つけて保存することはできないし，そもそもシステム化するとある一定のルールに従って，データを登録してもらうことになる。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，システム化したって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。

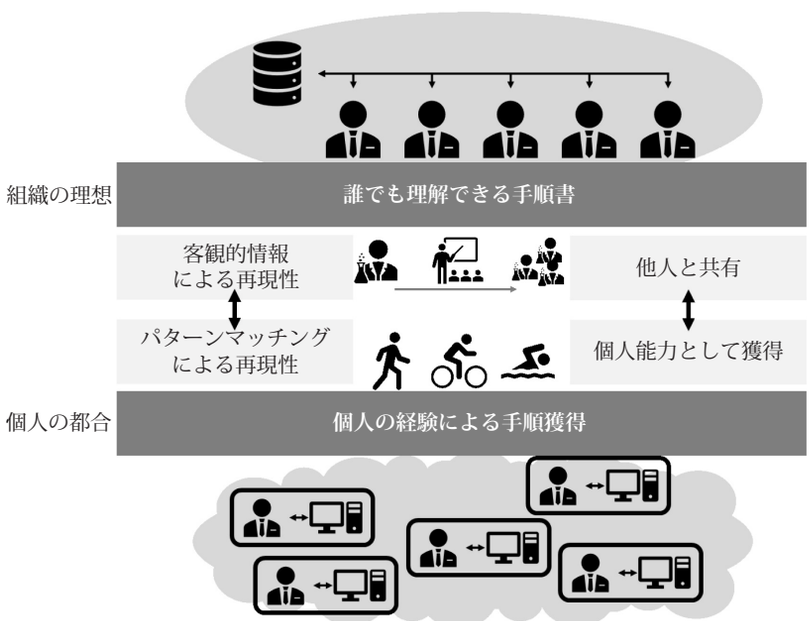

図 8　経験による手順獲得と文章による手順書作成

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，エクセル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。エクセル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，高機能なエクセルなどは，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。エクセル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。

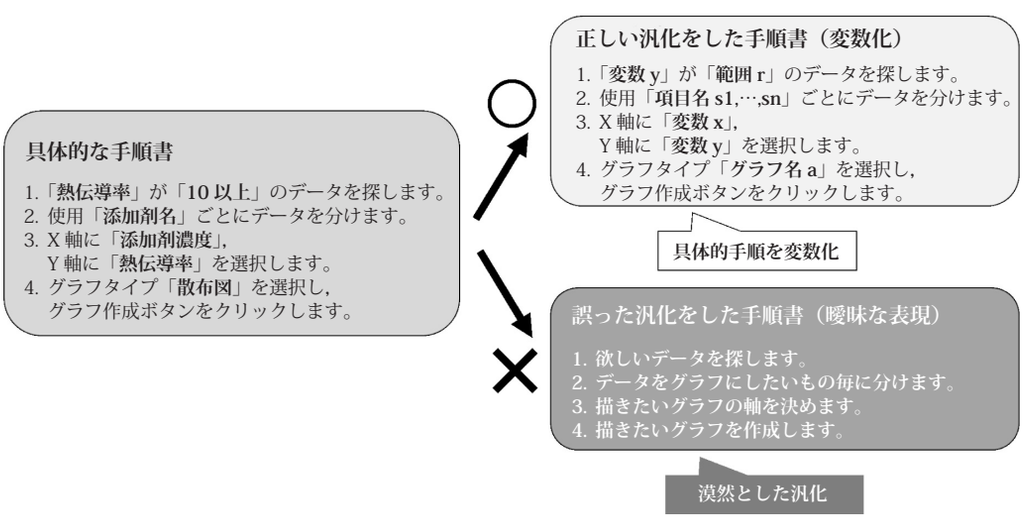

図 9　実用的な汎用的手順書の作り方

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。

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。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同じかどうかも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，そのような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？」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。

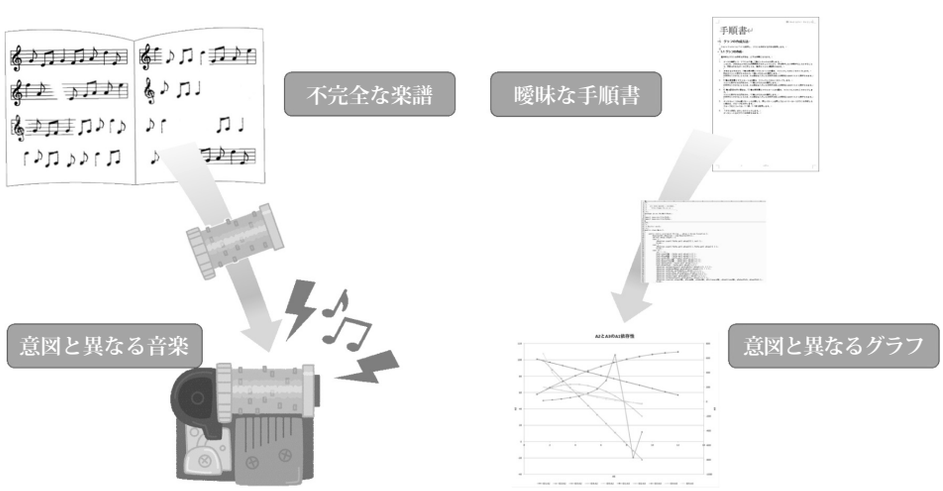

図 10　手動で行えていないことをシステムに組み込んではいけない

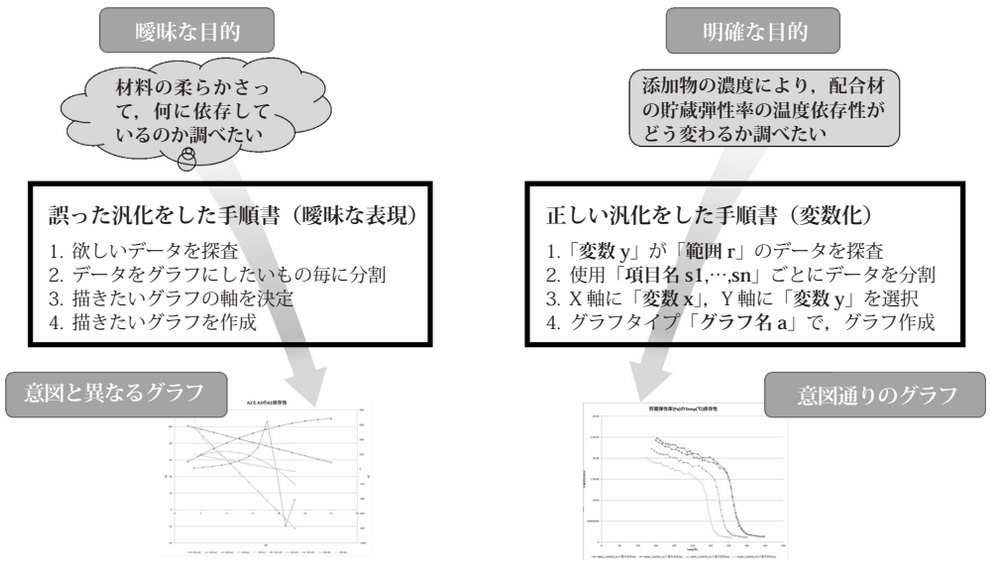

図 11　データ探査や処理の目的を明確化することの重要性

これらの「なぜ」の真の目的を明確化するためには，障害の原因分析に良く使われている「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，真因に近づけるはずで，「なぜなぜ分析」自身の方法は，Web 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，真の目的からデータ探査やデータ処理方法を見つめ直し，最適でないことが分かれば手順書を改善することが必要である。

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。