手元に溢れる情報を如何に管理するか？
— 研究生活のためのインフォマティクス

Hints for Sustainable Data Management and An Effort for Data Sharing for Researchers.
Key-words: Informatics, Data management, Repository, e-Science

轟　眞市
Shin-ichi TODOROKI (National Institute for Materials Science)

1.　はじめに

日々収得する大量のデータを持続可能な手段で管理する方法論は、研究分野を問わず求められている。筆者がコンビ研究への参画を通じて身につけた方法論のエッセンスを示すとともに、研究プロジェクト内でのデータ共有を視野に入れたインフラ構築に関して e-Science を目指した取り組みを紹介する。

PC に使われるハードディスクの容量は、年を追うごとに指数関数的に増大している。2 年で 2〜3 倍の勢いで増えているはずなのに、買い足しても、買い換えても、じきに満杯になってしまうのは、誰しも経験されていることであろう。溜め込むデータは増え続ける一方で、溜まったデータを活用する人間側の能力は、果たして進歩しているのであろうか？

筆者がコンビナトリアル材料科学プロジェクトに参加したのは 11 年前のこと。装置から吐き出される大量のデータに立ち向かうために身につけた技術をまとめてみる。それらはコンビプロジェクトを離れた今でも役立っている。

言うまでもなく、この手のノウハウは十人十色、さまざまなものがあるはずで、全ての方法論を把握し比較することは不可能だし、意味が無い。しかし、筆者が今まで破綻せずに続けてきた方法論を、そのベースになる考え方を抽出しつつ示すことは無駄ではないだろう。表面的には古くさく見えるかもしれないが、その根底に流れる基本方針を参考にしていただければ、

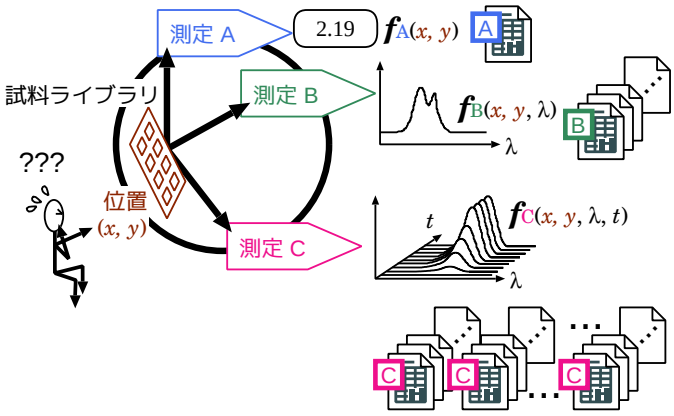


図 1: コンビナトリアルテクノロジーによって集積化した試料ライブラリを，複数の方法で評価した場合に得られる多次元データ。任意の変数や関数間の相関を取るためには、その度に大量のファイルを処理しなければならない。

この先 10 年、自分に合った持続可能なやり方を構築するのに役立つのではないだろうか。

ここでひとつ注意しておかねばならないことがある。これから紹介するデータ管理手法は、あくまで個人で閉じた運用をするものであり、他人との共有は考慮していない。とはいえ、研究グループ内での閉じた共有にも需要があることは心得ている (例えば、ポスドクや学生が残した実験データの引き継ぎなど)。そこで本稿の後半に、筆者の把握している範囲でのそれらの動向をまとめておく。

前半の話題については、2 つの視点に分けて論ずる。コンビナトリアル材料科学に携わると避けて通れない多次元データの取扱い方についてと、より一般的な視点からデータコレクション全体の管理方法についてである。

2.　多次元データ処理

研究開発プロセスをコンビナトリアル化するにあたっては、試料作製から分析・評価・解析までを一貫して設計しないと意味が無い。1 個の試料ライブラリ内に多数の試料セグメントを集積化できたとしても、その後のプロセスも自動化・並列化しない限り、それらが律速段階となって全体のパフォーマンスは向上しない。

これはデータ処理工程も例外ではない。試料ライブラリには試料セグメントが線状や面状に集積されているので、分析・評価データの次元数は、位置座標の分だけ増加する。また、同一の試料ライブラリに対して

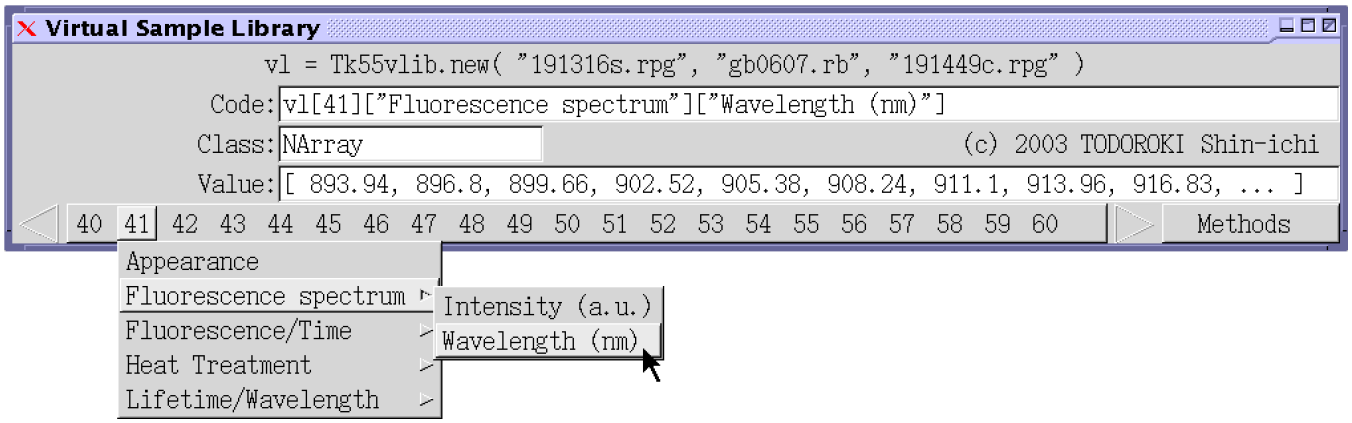


図 2: 仮想試料ライブラリの構造を GUI で表現した例。試料ライブラリ中の座標「41」に対する「Fluorescence spectrum(蛍光スペクトル)」のうちの「Wavelength (nm)」を呼び出している状態。

複数の測定を行い、それらの相関を調べようと思えば、必然的に多次元データ処理が必要になる (図 1 参照)。

多次元データ処理を行うにあたり、筆者が留意したのは次の 2 点である。

1. 多次元空間の中から必要なデータにアクセスする際には、直感に反しない方法をとること。

2. 操作においては GUI(グラフィカルユーザインターフェース) を使わないこと。

これらは一見、あい矛盾する要件の様に思える。後者が必要な理由は、マウス等を使った人手を介する作業を廃し、自動化を徹底するためである。そのため、何らかのプログラミング言語を使ってデータを操作するとになる。

前者は、そのプログラミング言語の機能を用いて実験データを規格化・抽象化することを意味する。ある特定のセグメントで測定したデータを抜き出すのに、その当時セーブしたファイル名やデータ形式を思い出す手間を介するのは時間の無駄である。一度全データを仮想的な容器に収容し、試料セグメントの座標や測定データの種類を指定するだけでデータにアクセスできる様にするのが効率的である。

その容器は、既存の多次元データ処理ライブラリを使っても良いし、自分で設計しても良い。11 年前の筆者は、「仮想試料ライブラリ」という容器を自作した [1, 2]。専門用語を使えば、オブジェクト指向プログラミングにおいて「試料ライブラリ」クラスを定義したのである。

その容器の構造を直感的に理解していただくために、便宜的に GUI を使って説明する。図 2 に、試料セグメントが一列に並んだ仮想試料ライブラリの例を示す。これは、ウインドウの再下行に並んだ座標の一つを選択し、その試料セグメントに対して収納されているデータのひとつを選択した状態である。選択されたデータの中身は Value 欄に表示されており、その値を呼び出すための記法は Code 欄に示されている[1]。つまりプルダウンメニューを辿る要領で順番に記述して、目的のデータに到達する。

実際にグラフ等を作成するには、この記法を使って呼び出したデータをグラフ描画ライブラリ[2]に転送すれば良い。この仮想試料ライブラリに実際のデータを放り込む操作によって、データの規格化・抽象化を図るのである。各種測定データファイル[3]を読み込ませる手順をあらかじめ定義しておく必要があるが、一度設定すればそれ以降は各データファイルの存在について気にする必要がなくなるのである。

おそらく読者のほとんどの方々は、グラフを作成する際には専用のアプリケーションソフトウエアをお使いになっていることであろう。そのソフトウエアが高度なプログラミング機能を有していれば、同等なことは実現できるはずである。GUI を介した手作業の排除と直感的なデータへのアクセスを両立させれば、大量のデータ処理は恐れるに足りない。

なお筆者は超高速動画の画像解析にもこの手法を応用している。動く被写体を捉えた画像データは、xy 座標、光強度、時間の 4 次元データであり、それが撮影回数分蓄積される。文献 [3] のオンライン版に掲げて

[1] オブジェクト指向スクリプト言語の Ruby を用いた。
[2] Ruby/PGPLOT を用いた。
[3] ウインドウの 1 行目の括弧内にファイル名が指定されている。

ある動画を作成するにあたっては、多次元データを収容する容器[4]を介して、必要なデータをプロットしている。前人未到のデータを取り扱う際には、装置メーカが提供している解析ソフトウエアでは間に合わないことはよくあることである。確かに商品版のソフトウエアは GUI が優れているが、痒いところに手を届かせるためには、GUI を経由しない手段も身につけておいた方が良い。

3. データコレクション管理

実験データをディスクに保存するときに、どのような方針を立てておられるであろうか？筆者が留意しているのは次の 3 点である。

1. ファイル名をタイプする時に迷いが生じない単純な命名ルールに則ること。
2. 後でそのファイルを探し出すための検索手段が確保されていること。
3. データを丸ごとバックアップする際に、ファイルの新旧を容易に判別できること。

これを満たす方針は多々考えられるが、叩き台として筆者の方針を示しておく [4]。

1. および 3. 全ての実験データを、その内容に関わらず時間順に保存する。
   (a) ファイル名は測定日時を表す 6 桁の数字を用い、実験条件との対応関係を電子化実験ノートに記録する (図 3 右参照)。
   (b) フォルダ名は測定月日を表す 4 桁の数字を用い、実験内容が直前のものと異なる場合には新しいフォルダを作成する (図 3 左参照)。
2. 検索するにあたっては、自分の記憶を辿って日時を特定するか、あるいは電子化実験ノートに搭載された全文検索機能を使う。

ファイル名を無機質な 6 桁の数字にすることに抵抗を感じる方もおられるかもしれない。しかし、実験条件等の情報をファイル名に盛り込むのは避けた方が良い。ファイル名が長くなると、他のファイルとの差を見つけるのに苦労する。他と共通する条件の記述を省略してしまうと、後で省略された条件を思い出すのに

[4]Numerical Ruby NArray を用いた。

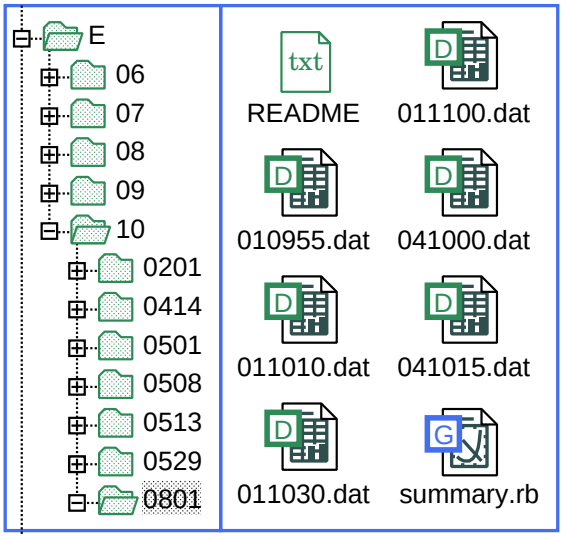


図 3: 実験データを保存するフォルダ構造の例。

結局実験ノートを参照することになる。それならば、最初から実験条件を漏れなくノートに書き留めておき、そこからファイル名を参照した方が良い。

電子化実験ノートを利用するために、筆者はパスワード認証機能を搭載させたブログサーバを立ち上げている [5]。Web ブラウザを通して利用できるので、LAN に繋がった端末ならどこでも書き込み閲覧ができる。

なお筆者は必要に応じて、実験ノートの内容や実験データを、職場や自宅のパソコンとの間でネットワーク越しに相互に自動転送できる様に設定してある [6]。どのパソコンを使っていても、同じハードディスクが見える様にしておくと、気分良く仕事ができるし、バックアップが常に取られている安心感がある。

一昔前であれば、この種の自動化を実現するには専門知識が必要であったが、昨年あたりからクラウド技術を活用した誰でも使えるフリーなサービスが登場してきた (Dropbox や ZumoDrive など)。容量制限や課金、機密保持の問題が気になるのであれば、自前で構築するのが良いと思う。なお、スティックメモリに入れてファイルを持ち運ぶのは、紛失の危険性が避けられないし、移動の度にファイル転送の手間がかかるので、避けた方が良い。

4. 実験データ共有手法に関する動き

さて、以上のルールは個人に閉じて運用してきたからこそ、破綻無くやってこれたのである。これを、他人が見ても分かるものにしようとするなら、結構な手間が掛かることは容易に想像できる。誰がアクセスし

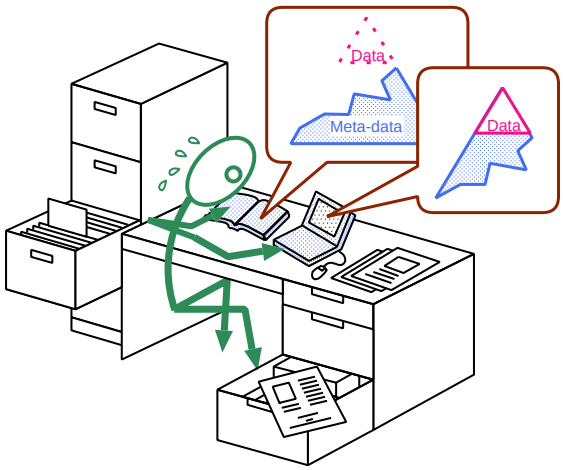


図 4: 人が去ってデータは死ぬ。データ単体ではただのゴミであり、そのデータが記録されたコンテクスト (Meta-data) と対応させないと生きてこない。

ても迷いの無いルールを決める必要があるし、同時にアクセスしてきた時にも破綻しない様に処理しなければならない。

また、他人が登録した実験データ単体を見せられても、その意味を読み取ることは非常に困難である。実験ノートに記された諸々の条件 (コンテクスト、あるいは Meta-data) を理解してこそ、そのデータが語ることが引き出せるのである。どこの研究室でも、学生やポスドクが去った後に残されたデータを解読するのに苦労されているではないだろうか？(図 4 参照)。

研究室や研究プロジェクト内で実験データのオンライン共有をしたい、という夢はしばしば語られるが、そのためには実験ノートの共有も視野に入れる必要があるし、機密保持が確立したインフラの構築や構成員への継続的な教育が必要となる。それにはかなりの資金が必要であり、戦略的に投資を回収するような研究プロジェクトでも無い限り、その実現は難しい。大学や研究独法全体での導入は縁遠い話の様に聞こえる。

ところがまったく別のセクタから、実験データ共有を視野に入れたインフラを立ち上げる動きが出てきた。その原型は、昨今大学等の図書館がこぞって導入している「機関リポジトリ」である[5]。そもそもこれは、研究機関が生産した研究成果物をオンラインで提供するために世界共通の規約[6]で設計されたデータベースである。その流通単位は、論文や報告書、プレゼンテーションスライド等の文書ファイルであり、研究者 (あるいは代理の図書館員) が自らの意思でそれらを公開するセルフアーカイビングによって成り立っている。このリポジトリシステムを発展させ、e-Science の舞台として使おう、という動きが現れてきているのである。

本誌読者の方々はおそらく耳慣れないであろう e-Science という用語は、高度に発展したコンピュータネットワークを活用した科学を意味する。既に、天文学や素粒子物理学の分野では、大規模な観測装置が排出する観測データを国際的に共有するインフラを整えている。このようなインフラを他の研究分野にも広めるには、現在普及しつつあるリポジトリシステムに実験データを登録し、それにアクセスできるユーザの範囲を個別に設定できるようにすれば良い、というのがアイデアの基本である。

つまり、流通単位を文書から実験データに細分化し、研究グループ内でのみ閲覧を許可する仕組みを導入するのである。実験データに適切なメタデータを付与して検索し易くし、またそれに連動するブログ等を実験ノートとして活用すれば、小規模なグループ内でのデータ共有が可能となる。

e-Science 推進側の狙いはそれだけに留まらない。論文が刊行されれば、それに付随する実験データも公開してもらい、他の研究者による検証や再利用に供することも想定している。研究者が望むなら、論文に掲載できなかった付随的なデータを公開しても良い。このような実験データのセルフアーカイブが一般的になれば、今後リポジトリ上の実験データを引用する論文が現れることになるだろう。

筆者が所属する物質・材料研究機構でも、NIMS eSciDoc というリポジトリシステムを構築中である [7]。eSciDoc とは、ドイツの Max Planck Digital Library が開発・公開している e-Science のためのシステム基盤ソフトウエアの名称であり、NIMS は共同研究プロジェクト立ち上げて、自組織に必要な機能の追加を進めている。2 年前の秋に機関リポジトリ機能の試験公開を開始して以降着々と進歩をとげ、組織内の指定したメンバー間で文書を共有する仕組みはこの夏にほぼ完成する見込みである[7]。

[5] http://maps.repository66.org/ でその分布を世界地図上で見ることができる。

[6] Open Archives Initiative (OAI) が策定したメタデータ収集のためのプロトコル (OAI-PMH)。

[7] 外部の方々が NIMS eSciDoc の機能に触れていただくのに良いサービスは、今年 6 月に運用を開始した研究者総覧システム (http://samurai.nims.go.jp) である。人事 DB や業績評価 DB と連携して所属研究者の研究紹介ホームページを自動生成するものである。各人がセルフアーカイブした文書もダウンロードできる。

今後このシステムが材料科学における e-Science を育む基盤となるか否かは、研究者のニードに寄り添ったシステム改良を続けていけるかにかかっている。筆者の考えは楽観的である。国際標準の規格に則って、オープンなシステム基盤の上に構築されている点や、所蔵物の永久保存と検索の容易性を最優先に考える図書館の文化から生まれてきたシステムである点は、持続可能性にプラスに働くに違いない。むしろ、研究者自身の保守性や時間的な余裕の無さが一番の障害になる様に思う。

5. おわりに

手元に集めた情報の管理手法に関して、筆者が関心を抱くきっかけになったのは、就職した年に刊行された、野口悠紀雄著「『超』整理法」[8] であった。それ以来、「検索する手段を確保しておけば、情報を保存する際に整理する必要はない。」と「情報は一ヶ所にまとめよ (ポケット一つの原則)」を基本原則に研究生活を送っている。その当時は、主として紙媒体に記録された情報に対する管理手法として捉えられていたが、電子媒体上の情報管理にも有効であり、ご利益を受けつづけている。

これが他人と共有する情報にも適用できれば良いのだが、それは不可能と野口氏は断じている。それをカバーするのが情報通信技術であることは疑いない。e-Science の動向も含めて常にチェックしていきたいものである。

文 献

[1] 轟 眞市：“仮想試料ライブラリによる多次元データ管理”, コンビナトリアルテクノロジー—明日を開く ’もの作り ’の新世界 (鯉沼 秀臣, 川崎雅司（編）), 丸善, 東京, 第 7.2 章, pp. 195–202 (2004). (ISBN4-621-07447-4).

[2] S. Todoroki: “Object-oriented virtual sample library: a container of multi-dimensional data for acquisition, visualization and sharing”, Meas. Sci. Technol., **16**, 1, pp. 285–291 (2005).

[3] S. Todoroki: “In situ observation of modulated light emission of fiber fuse synchronized with void train over hetero-core splice point”, PLoS ONE, **3**, 9, p. e3276 (2008).

[4] 轟 眞市：“研究生活のためのインフォマティクス (2) 実験結果のトレーサビリティ—汝のディスクを闇で満たすなかれ”, マテリアルインテグレーション, **21**, 11, pp. 76–77 (2008).

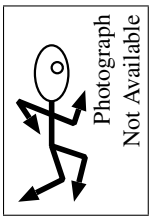
[5] S. Todoroki, T. Konishi and S. Inoue: “Blog-based research notebook: personal informatics workbench for high-throughput experimentation”, Appl. Surface Sci., **252**, 7, pp. 2640–2645 (2006). (和訳「ブログを基にした実験ノート: 個人の研究活動を効率化する情報環境」をセルフアーカイブ公開中。).

[6] 轟 眞市：“研究生活のためのインフォマティクス (1) ポケットひとつの原則—ファイルは手ぶらで運ぶもの”, マテリアルインテグレーション, **21**, 10, pp. 68–69 (2008).

[7] 谷藤 幹子, 高久 雅生, 大塚 真吾, 轟 眞市：“材料系研究所におけるリポジトリシステムの実践と将来”, 情報管理, **51**, 12, pp. 888–901 (2009).

[8] 野口 悠紀雄：“「超」整理法”, 中央公論社 (1993). ISBN 4-12-101159-7 (中公新書 1159).

筆者紹介



轟 眞市 (とどろき しんいち)
1993 年京都大学大学院工学研究科博士後期課程修了. 同年 NTT 入社. 1998 年 科学技術庁無機材質研究所入所を経て、2001 年 4 月より現職. 高強度光を伝搬する光ファイバにおける損傷現象の研究に従事.
[連絡先] 〒 305-0044 つくば市並木 1-1
物質・材料研究機構 光材料センター
URL: http://www.geocities.jp/tokyo_1406/