

第8回 SPring-8 先端利用技術ワークショップ

「オープンサイエンスに向けたオープン解析プラットフォーム」
参加報告国立研究開発法人理化学研究所 放射光科学総合研究センター
杉本 宏

1. はじめに

2017年1月7日に神戸芸術センターにおいて「オープンサイエンスに向けたオープン解析プラットフォーム」と題するワークショップが(公財)高輝度光科学研究センター(JASRI)主催で開催された。本ワークショップは、同会場で日本放射光学会年会・放射光科学合同シンポジウムが開催される直前ということもあり、多数のSPring-8ユーザーおよび放射光施設関係者の参加があった。オープンサイエンス(インターネットを活用して科学研究の成果を社会に対してオープンに共有することで、より効果的に発展させようという試み)は、世界的に大きな流れとなっている。公的資金によって大量の実験データを生み出しているSPring-8でも、オープンデータの考え方を取り入れる必要があるのではないかという論理は多くの人が理解できるだろう。その一方で、オープンにするデータの切り分け方や、膨大な量の生のデータに対応するための技術的な課題も多いはずである。今後、新しく開発される測定方法やデータ解析技術の進展も見据えて考えていかなければならない。本ワークショップでは、データの公開を念頭にしつつ、まずはSPring-8でのオンラインデータ解析の基盤を今後どのように整備していくか、そしてユーザーの利便性を向上させるにはどのような観点が必要かを議論することを目的として、各分野から実際の利用研究におけるデータ処理の事例が紹介された。本稿では、各講演の概要を紹介した上で筆者の感想を述べたい。

2. 講演

冒頭にオーガナイザーのJASRIの八木直人氏からの趣旨説明として、オープンサイエンスの推進についての内閣府や日本学術会議による検討状況が紹介さ

れた。具体的な推進策がまだ明確ではなく、各コミュニティでの議論を進める必要性が挙げられている。また、SPring-8の施設としては、単なる実験データの公開だけではなく、それを解析するための大容量データサーバと、オンラインソフトウェアの提供が一体となったデータ解析プラットフォームの整備が、利用者にもメリットが大きいという提案が行われた。

最初の講演者として、理研の横田秀夫氏から画像処理手法の開発者としての観点に加え、科学分野における大量のデータの一元管理の方法やオンライン解析の経験に基づいたシステム構築の方法、そしてオープンサイエンスの展望について講演があった。画像処理の技術の応用として、親子判別、交通の監視、戦車の追跡、車両の自動運転など様々なアルゴリズムがあるが、学術研究への応用については、処理の前提となる条件が未定である場合や、細胞のようにダイナミックに変化する難しい観察対象というのが特徴的のことだった。また商用クラウドの利用のメリットについて力説し、横田氏らが構築した画像処理プラットフォームでは、計算機リソースの整備や大量のデータを維持管理するためのコストの問題を解決するために、Amazon Web Serviceを利用している例が挙げられた。

JASRIの松下智裕氏の講演では、原子分解能ホログラフィーの測定制御と解析ソフトウェアの開発経緯が説明された。本手法は、結晶中のドーパントなどの原子構造や、触媒の反応中心での機能発現機構の解明を行うために用いられる。1998年頃から開発をスタートし、測定画像のデータフォーマットの汎用化、そして応用対象の模索から始まった。現状は、松下氏が現在参画している新学術領域研究グループの強力な推進力によって、幅広い分野の研究者がこのソフトを利用する機会を得ている。現在、研究分野の発展によ

りユーザーが拡大しており、成長期の入り口にさしかかったという位置づけでとらえている。これまでのフリーウェアという配布形態からクラウド上での利用に切り替える展望が述べられた。また、ソフトウェア開発には、戦略と組織化が最も重要であると結論した。

JASRI の伊藤真義氏の講演では、コンプトン散乱実験のデータ解析の手順を解説し、画像データを逐次処理していく際に必要な解析サーバの性能と大容量ストレージの必要性を強調した。世界的にみても、実働しているコンプトン実験ステーションは SPring-8 BL08W 高エネルギー非弾性散乱ビームラインのみであり、解析ソフトウェアの開発・保守は、現在、BL 担当者が一人で行っている。SPring-8 に施設で整備された共用データサーバがあれば、少なくともサーバ管理は BL 担当者の手から離れるメリットがある。さらに、検出器からの生データは大容量 (884 GB/day) であるため、オープンデータに際してはそれなりのデータストレージが必要だと述べられた。生データの保存が可能であれば、測定後の再解析も容易になり、条件によってバックグラウンドの低減、分解能の向上が見込まれる。

(株) 豊田中央研究所の林雄二郎氏の講演では、SPring-8 BL33XU 豊田ビームラインにおいて構築した X 線マイクロビームを用いた走査型の 3D X-ray diffraction microscopy (3DXRD) 顕微鏡装置によるデータ記憶や解析方法についての解説が行われた。鉄鋼材料の内部の不均一性を定量的に解析する為に、結晶粒のスケールでの実験と理論計算に基づいたデータ処理が行われる。解析作業は SPring-8 内のサーバに研究所からリモート接続を介して行っており、約 12 時間かけて測定した 1.6 TB 程度の回折斑点の情報から構造の再構成に要する計算時間は、小規模クラスタ計算機で 1 週間ほど要する。解析には無償公開されているソフトウェアを利用していた。

東京大学の篠原佑也教授からは、X 線光子相関分光実験 (XPCS) とデータ解析について講演が行われた。コヒーレント X 線によるスペックルを時間分解測定し、散乱強度の時間相関を計算することで、系のダイナミクスに関する情報を得る手法である。ゴム中でのナノ粒子の揺らぎといった現象を観測できる例が挙げられていた。世界的には XPCS 用の新しい検出器を作る動きがあるが、データ転送や保存の方法も検出器

の開発と同様に重要であると述べられた。実験で得られるデータは 4 TB 程度とのことだった。実際の放射光施設での測定では、観測されるスペックルのゆらぎが試料内部の本質的なものなのか、蓄積リングの電子軌道のゆらぎ等に起因するものかどうかをその場で判断する必要性が述べられた。また、実験が終わって帰った後も生データにアクセスして新たな解析法を試したい状況があるため、オープン解析プラットフォームが切望されるとのことだった。ただし、汎用性のない独自の規格や、GUI 化のデメリット、さらに、実験前に検出器制御や解析のスクリプトなどをテストできるような仮想サーバの整備といった具体的な要望が挙げられた。

JASRI の熊坂崇氏からは、生命科学分野におけるオープンデータの流れについての現状と課題についての解説が行われた。バイオサイエンスの領域では、ゲノム DNA の塩基配列情報や、メタボローム解析を筆頭に、莫大な情報量をまとめた公開型のデータベースが多数存在している。放射光関連では、タンパク質の立体構造情報のデータベースである Protein Data Bank (PDB) や、X 線イメージングの CXIDB が知られている。また、講演ではタンパク質結晶構造解析における試料調製から分子構造情報まで、データの取得や一連の解析の自動化が進んでいることが説明された。今後は完全自動化の方向へ進むため、これとオープンデータが融合していくという展望が述べられた。

3. 感想

本ワークショップは、オンラインデータ解析の現状を把握した上で将来のユーザーの利便性や新たな技術の発展に寄与するにはどうしたら良いかについて考える初めての機会であったと思う。オープンソースや学術論文などのオープンアクセスに比べてオープンデータについての議論はまだ入り口の段階で、各利用分野によってデータの互換性、サイズ、解析方法、記録方法も千差万別であることが浮き彫りになり、このことが本テーマの議論を複雑にしていると感じた。その一方で、SPring-8 のユーザーや BL 担当者が BL で実際にどのような測定や処理を行った後、データをどのように所属機関へ持ち帰っているかなど、現状と今後の展望についての現場の声を聞ける貴重な機会であった。

筆者が専門とする構造生物学（特にタンパク質の結晶構造解析）の分野の特徴についてももう少し触れた上で、私見を述べたい。熊坂氏の講演にもあったように、この分野ではオープンソースやオープンデータの概念が比較的早い段階から取り入れられてきた。1970年代に設立された PDB は、研究者が自身の解析結果の妥当性を示すため、そして科学研究を発展させるためには、不可欠なデータベースである。その有益性に疑いの余地はない。そういえば、つい昨年に話題になった例では、構造精密化用プログラムの開発チームが、すでに PDB で公開されていた構造座標に大きな問題があることを見つけた。タンパク質領域にダミー原子として多数の水分子が置かれているなど、全体的に正確性を欠いた分子モデルだった。このような間違いは、位相決定のために分子置換法に用いた初期モデルのバイアスに起因したとはいえ、経験者なら考えられない間違いである。10 年程前には、手系が逆の電子密度に無理矢理タンパク質のモデルを組んだ例もあった。これらは、論文の査読や PDB による構造 validation にも限界があることを見せつけた。この分野ではほとんどのジャーナルで PDB へのデータ登録が論文受理の必須条件になっていることから、X 線回折画像という生データを登録しなさいと言われたとしても、技術的な問題がクリアされれば、研究者側に違和感がないと考えられる。タンパク質の立体構造解析のための回折データの測定はルーチ的な要素が多いことから、国外の放射光施設では遠隔測定が主流となっている。従って、オンラインデータ解析プラットフォームはあって当然のシステムである。このように、完全自動化と遠隔測定が進むのは時代の流れだが、人材育成の観点からは少々不安な気持ちになる。

4. おわりに

オープンサイエンスは一般論としては正論だが、それぞれのコミュニティでの議論をまとめるには時間を要する。どの分野でも共通しているのは、データの量とメタデータの標準化である。そして、SPRing-8 ユーザーにとっての利便性とオープンサイエンスの推進意義が両立しなければ成功しない。何らかのインセンティブがないと人はなかなか従わないからである。技術的にもまだ課題が多いため、拙速に導入すべきではないというのが大部分の意見のようだ。一方、

SPRing-8 のデータ解析プラットフォームについては、近い将来に検出器から出力されるデータが毎秒 100 GB に迫る大容量となると、生データを自分の所属機関に持ち帰ることや転送することは実質的には不可能となる。従って、本ワークショップで議論された効率の良いデータ解析のためのプラットフォームの必要性がいつそう認識されるだろう。オープンデータを見据えた利便性の高いシステムの構築については、ユーザーと施設の双方が協力して問題解決に取り組むことが重要である。



プログラム

- 「趣旨説明」 八木 直人 (JASRI)
- 「クラウド画像処理システムの開発と複数装置情報の統合解析」 横田 秀夫 (理化学研究所)
- 「原子分解能ホログラフィーの測定制御・解析ソフトウェア」 松下 智裕 (JASRI)
- 「コンプトン散乱実験とデータ解析」 伊藤 真義 (JASRI)
- 「走査型 3DXRD 顕微鏡法とデータ解析」 林 雄二郎 (豊田中央研究所)
- 「XPCS 実験とデータ解析」 篠原 佑也 (東京大学)
- 「タンパク質結晶構造解析～結晶化・回折測定・データ解析」 熊坂 崇 (JASRI)
- 「大規模データを扱うその他の実験手法」 八木 直人 (JASRI)

杉本 宏 SUGIMOTO Hiroshi

(国) 理化学研究所 放射光科学総合研究センター
〒679-5198 兵庫県佐用郡佐用町光都 1-1-1
TEL : 0791-58-2817
e-mail : sugimoto@spring8.or.jp