PF 分光分析ビームラインにおけるデータパイプライン構想

仁谷 浩明
KEK 物質構造科学研究所放射光実験施設 基盤技術部門 BL 制御開発チーム

PF の硬 X 線 XAFS ビームラインでは年間数万回の測定が行われているが、実験に携わる人数を考えると、そのすべてのスペクトルデータが活用されているとは考えにくい。主なボトルネックはほぼ手動で行われているデータ解析パートにあることは容易に想像がつく。一方で、データ取得時に測定器の適切な設定が行われず、品質の低いデータを取得してしまったため満足できるデータ解析が実施できなかったという例もあると思われる。さらには、データの管理が適切に行われていなかったため、学生が卒業してしまってデータを紛失してしまったという例もあるかもしれない。実験施設としてはデータが測定されるだけでなく、そのデータが解析され、論文等で発表されることでようやく施設としての評価がされるため、特に次の点が重要となる。①常に質の高いデータが得られること、②測定データが適切に保存されて必要時に遅滞なく取り出せること、③すべてのデータが解析されて数字が科学的な意味を持つこと。つまりはデータの取得から保管、解析、可視化までを一貫したルールの上でハンドリングすることで高品質のデータを余すことなく研究者に提供できることが理想である。このようなシステムは研究開発の分野ではビッグデータ解析などの発展とともに一般化してきており「データパイプライン」と呼ばれることもある。放射光界隈ではタンパク質結晶構造解析はすでにこのようなシステムの構築に取りかかっており、導入の遅れている分光分析分野でも導入を検討する時期に来ていると考える。

実際に導入を行おうとすると開発要素としては、データ収集パート、データ保管パート、データ解析パートの 3 つに分けられる。データ収集に関しては、様々な試料に対して最適の測定パラメータが提供されることや時間効率よくデータ測定ができるようにするなど、究極的には全自動測定システムの開発が一つの目標となる。データ保管に関してはどれだけのデータ量を想定するかで設置場所（物理的、ネットワーク的）やシステムの規模を検討しなくてはいけない。また、次のプロセスで使いやすいようにデータを整形することや、他のデータベースとの連携を考慮してインターフェースを設計する必要がある。データ解析に関してはディープラーニングを取り入れて、どこまで解析自動化が行えるかが重要である。いずれのパートもここ数年で高度に発展してきており、昔のように一人の技術者が最初から最後まで開発を行うということは難しくなってきている。従って現状の施設や放射光コミュニティのみではこのデータパイプライン構想は実現不可能である。逆に各分野の専門家とうまく連携することができれば、既存技術の組み合わせでも比較的短時間でシステム構築が可能であると考えられる。放射光コミュニティと専門家コミュニティの双方にメリットが出るようなプロジェクトの発展が望ましい。