オープンリサーチデータと データプラットフォーム

国立情報学研究所所長
東京大学生産技術研究所教授
喜連川 優（きつれがわ まさる）

今年5月、米国における基礎研究推進をつかさどるNSF(米国国立科学財団)はBig Ideasと銘打って、将来に向けた9つのテーマを最重要課題として取り上げた。その1番目に挙げられたテーマは「Harnessing Data for Science and Engineering」である。[注1] 2012年3月に「Big Data Initiative」を立ち上げた米国において、その4年後、再度データの重要性をうたったことは注視に値しよう。昨今、第3次AI(人工知能)ブームが起こりつつあるが、膨大なデータが利用可能になったことが、最大のドライバといえる。大量のデータがなくては、今日のパワフルなAIは機能しない。

データドリブンソリューションが次々とその威力を発揮するなかで、質の高いデータを保有することの重要性が認識されつつある。今やデータは最も貴重なアセットである。第5期科学技術基本計画におけるSociety 5.0が描く新しい社会像のなかでも、データの環境整備は中核に据えられるべき対象といえよう。

新たなデータアセット：オープンリサーチデータ

オープンソース、オープンコースウエア、オープンアクセスをはじめ、これまで多様な観点で「オープン」が語られてきた。オープンデータは、データに対してのオープン化の方向感と位置付けられている。オープンガバメントが指向されるなか、政府データのオープン化については、すでに議論が多々なされている。ここでは、オープンサイエンスにおけるオープンデータに注目したい。この課題は、最近つくばで開催されたG7科学技術大臣会合でも取り上げられ、継続的な話題となっている。サイエンスにおける基礎研究は、現在では国家が公的資金を研究開発費として供給する場合が多いが、オープンサイエンスの根底にある考え方は、国家間で研究成果をよりオープンに利活用することにより科学発展の大きな加速がされ、イノベーションの創出につながることが期待できるというものだ。最初に議論されたのは、論文・報告書等を途上国も含め誰でも読めるオープンアクセスである。これについては、オープンアクセスジャーナルが発刊されるなど一定程度進展が見られ、現在は、次のステップとして、論文成果の元になったデータのオープン化に注目が集まっている。論文に記載される内容は、最終的な研究成果を凝縮した知識のテキストによる表出であり、知の再現は必ずしも容易ではない。ITの進化とともに、紙媒体による出版という制約から解き放たれ、論文だけではなく、研究成果に直結するエビデンスとなる「データ」も一緒に提出しようという考えが自然に生まれてきた。さらに、データが論文に付随すると考えるのではなく、データはより多数の論文、すなわち、知を生み出す源であるとの認識が広がり、データそのものの価値を第一義的にクローズアップするオープンリサーチデータの潮流を生み出しつつある。最近ではデータを説明し公開することを目的とするデータジャーナルの誕生をはじめとして極めてホットな話題が耳目を集めている。従来、研究者の手元に埋もれてしまいがちであった研究データが新たなデータ資源となり、広く利用可能になることは望ましい方向といえる。日本学術会議[注2]においても、この議論が進められている。もちろん、「オープン」といってもすべてを公開することを意味するわけではなく、国家レベルや機関レベルなどでしっかりとしたオープン・クローズ戦略を練ることは必須である。

データプラットフォームの構築と学と産による活用

あらゆる分野においてデータの価値の重要性が強く認識されるに至ったなかで、現在データを入れる器としての強力な「データプラットフォーム」の構築が注目されている。EUではEuropean Open Science Cloudの構築に向けた議論が進んでいる。欧州デジタル単一市場戦略では、データの流通を基に欧州圏の社会・経済全般の発展を目指している。[注3]

Google、Amazon、そしてFacebookに代表される近年のIT企業が、膨大なデータの積極的利活用から巨大な富を生み出したことからも明らかなように、いかなる分野においても、しっかりとしたデータ基盤の構築が要となる時代となった。〝学〟が生み出した学理に根差すオープンリサーチデータと、〝産〟の有するデータとを融合し新価値を創造する時代の到来が予見される。それを機動的に実現するためのプラットフォームの構築が望まれる。研究者の属する大学や研究所、あるいは、研究資金を提供する機関等がバラバラにデータ格納庫を構築することは経済的に非効率であり、スケールメリットを追求することが大切である。データに対する取り組みは研究分野によって異なるが、その先駆的取り組みで知られる米国のNIH(国立衛生研究所)が、約20PB(ペタバイト)のデータを保持するなど、すでにゲームは始まっている。日本の環境系データベースDIASも同程度の規模を有する。

今年3月に東京で開催されたRDA(Research Data Alliance)総会で基調講演を行ったが、多くの国からの参加者の多さに驚いた。当日、会場では、EUからSCIENCE 2.0というデジタル時代の科学の変革についての言及もなされていた。サイエンスの研究プロセス全体を徹底的にIT化することにより、科学的発見の圧倒的な効率化と高度化を目指そうとするものである。巨大研究データが最も重要な役割を果たすという意味であり、オープンリサーチデータへの取り組みはその第一歩といえる。

共創的データインフラの有無が勝負を決める時代へ

〝学〟の成果を実践的に〝産〟が利用できる環境が、国家を強くすることは自明であろう。わが国では北海道から沖縄まで大学・研究所は、今年4月より100Gbps(ギガビット／秒)[注4]という超高速ネットワークで接続され、巨大研究データへ縦横無尽にアクセス可能な環境が生まれる。まさに絶好のタイミングであり、産と学の機動的連携を実現する戦略的な取り組みが期待される。いかなる研究開発においても、研究ビッグデータ(情報爆発)を柔軟に料理できる共創的データインフラの有無が勝負を決める時代となった。国家戦略上、この認識は極めて重要である。

(注1)詳細は、http://www.sciencemagorg/news/2016/05/nsf-director-unveils-big-ideas-eye-next-president-and-congress参照

(注2)詳細は、http://www.scj.go.jp/ja/info/kohyo/pdf/kohyo-23-t230.pdf参照
(注3)詳細は、https://ec.europa.eu/digital-single-market/en/economy-society-digital-single-market参照
(注4)詳細は、https://www.sinet.ad.jp/参照