

データを共有するという文化

Let data speak to data

Nature Vol.438(531)/1 December 2005



いまやウェブツールを使うことで、論文の発表と並行してデータの共有や非公式の議論ができるようになった。しかし、そのメリットを十分に享受するために、研究者たちは文化を共有し、データベースをめぐる見通しを再考しなければならない。

自分の生データをアップロードして他人と共有し、自分のブログのインパクトファクターを高める道に進むか、それとも研究活動を断念するか。そのいずれかの選択を迫られる時代はまだ到来していないが、ブログのような個人的な情報発信ツールから電子実験ノートにいたるさまざまなウェブ技術が、ウェブの性格を、大型図書館のようなものからユーザー主導型の共同作業場を提供するものへと様変わりさせつつある。

この動きは、やがて数多くの研究分野に抜本の変革をもたらすことになるだろう。バイオインフォマティクスや高エネルギー物理学などの分野では、すでにこのような変革が起きている。いまから10年前、たとえば天文学の分野では、それぞれの研究グループが独自の観測データを保有し、その研究成果を個別に発表していた。それが今では、大規模なデータセットを中心にして研究が進み、データはコード化されて、学界全体が共有して利用できるようになっている。だが、より小規模で多様な研究者集団が互いにデータを共有しようとするには、データのタイプとフォーマットがきわめて多いためむずかしい課題が残る。

この現状を変えうる重要な技術変革が、集中型データベースから「ウェブサービス」への移行だ。ウェブサービスとは、データとソフトウェアへのアクセスを簡素化する公開インターフェースのことである（稼働中のウェブサービスの一例として、www.ebi.ac.uk/xembl/index.html）。こうしたインターフェースができたことで、最近まで熟練したプログラマーが扱ってきた領分であるデータ処理と解析も、プログラミングの基礎知識をもつだけで自動化できるようになった。

コンピュータで読解、操作でき、データベース間で「対話」ができるようなフォーマットによって、さまざまな種類のデータが保存されるようになり、その数はますます増えている。これによってユーザーは、複数のソースから得たデータを抽出、解釈する技術を利用し、まったく新しいデータ製品やサービスを作り出せるようになった。

たとえば、生物多様性の研究では、集中型の巨大データベースを構築するのではなく、必要なデータが取められている複数の既存のデータベースからデータを取得し、特定の種に関連するすべてのデータ（分類、遺伝子配列、地理的分布など）をまとめあげるといったことが考えられる。データベースは個人や研究室の研究プロジェクトの成果であることが多い

め、不安定な資金提供や、人材や研究室が新分野へ移ることによる影響を受けやすいという問題があるが、データを分散化すれば、そうした問題の解決にも役立つ。

研究分野に特化したデータベースが重要な役割を果たすことに疑いの余地はないが、研究者は、将来的なデータの維持管理が保証され、だれもが容易に利用できる大規模な総合的データ保管についても考える必要がある。それが理想論だと思うなら、インターネットアーカイブ（Internet Archive）とクリエイティブ・コモンズ（Creative Commons）によって創設された「OurMedia」というサービスを考えてみるとよい。これは、どのようなデジタル著作物でも（ホームビデオや休日写真でさえ）無料で保存し、データの共有を永続的に行うサービスだ。2005年11月には、Google社がグーグルベース（Google Base）を立ち上げた。これは、Googleの巨大なプラットフォームに何でもアップロードできるようにしたサービスである。

このようなサービスの登場により、公開データについての新たな考え方も必要になっている。ウェブサービスは、コンピュータがリアルタイムで自由にデータにアクセスできることを前提としている。GenBankや数多くの大型データベースがデータへの無制約のアクセスを保証する反面、数多くの研究機関で書類によるデータ使用の許可手続きが存続しており、ウェブサービスによるデータ利用の妨げとなっている。

収集に莫大な投資を要したデータについては、論文発表までの間は、関係する研究者がアクセス権を独占するとしてもその正当性が認められるかもしれない。しかし、秘密にしておく必要のない膨大なデータが存在するのでもまた事実である。クリエイティブ・コモンズのライセンス（<http://creativecommons.org/license> 参照）によって、データの再利用に関する権利内容と著作権表示を規定したうえで、コンピュータによるアクセスを妨げることなくデータを公開する方法もあるのだが、ほとんどの組織はそれを知らないようだ。

ウェブサービスが研究者に力を与えていくと考えられるなかで、このようなビジョンの達成にとって最大の障害は文化的な要素になるだろう。研究者は、今後も絶えず競争にさらされる。しかし、現代の研究者が世界にひろがる学界に対して貢献できることの幅をひろげる、データを共有する者にとって意味のある著作権表示方法の開発が非常に重要なのだ。