ビッグデータとオープンデータ：これからの科学研究を支える二つのキーワード
“Big-data” and “Open-data”: Two key words to support future science

*村田 健史[1]
*Ken T. Murata[1]

1.情報通信研究機構
1.National Institute of Information and Communications Technology

フェーズドアレイ気象レーダはXバンド気象レーダの一つであるが、アレイ状のレーダを回転することで3次元の降雨構造を取得でき、これまでと比較にならないスケールのデータが出力されることを意味する。多くの気象学者は、例えば100TBを超える1年間の継続観測データを蓄積し、気象ビッグデータからの特定パターンの現象抽出（データマイニング）を行うようなデータサーベイ研究手法を主流としない。多くの気象学者は特定の興味深い事例を選択し、その現象を詳細に解析する事例研究に注力する。フェーズドアレイ気象レーダの全期間データを機械学習することで、気象学の予備知識がほとんどない研究者が現象のパターン分類をするとしたら、気象学家はどう思うであろう。「気象学の知識なしにデータ解析ができるわけがない」と言うのがおそらく多くの研究者（専門家）の率直な意見であろう。しかし、実はビッグデータの世界では、知識がビッグデータを解析するのではなく、ビッグデータが知識を凌駕することがあるということが分かってきた。これは、おそらく多くの科学研究分野においては衝撃的なアプローチであり、このような方法が有効であるとはにわかには信じられないかもしれない。
自然言語処理においてもこのようなビッグデータ手法が有効であることが、最近明らかになってきた。現在のWebアプリケーションでコンピュータが文章を自動翻訳する技術では、言語特有の文法に基づいた判定をしない。あらかじめコンピュータにあらゆるタイプの文章をデータベース化しておき、判断対象の文章に出現する単語ごとにその単語が出現する文章からその単語の意味を類推する手法がとられている。これまでに科学技術がその礎としてきた演繹的手法とは全く異なるアプローチが、ビッグデータ指向型科学においては数々の成功を収めて始めているのである。もちろん、すべての科学研究において、ビッグデータ指向型研究手法が有効であると主張する者はいない。ここで考えるべきは、「専門家にしかデータは理解できない」という固定概念は捨てなくてはならないという事である。様々な機械翻訳コンテストに優勝したGoogleの言語翻訳チームは、中国語もアラビア語も話せない「非専門家集団」であった。
データ指向型科学ではデータを研究のスタートポイントにする手法であると前述したが、特にビッグデータ指向科学（data driven scienceとも言われる）はデータの中にすべての情報があるという発想に基づいている。このようなビッグデータ指向型科学の事例は多くの書籍やメディア等で紹介されている。「みんなの意見は案外正しい」ことが様々な分野で確かめられつつある。これは、インターネット社会では、近年「集合知（Collective Intelligence）」として示されている考え方である。ビッグデータ、オープンデータ、集合知、情報コモンズなどは、科学分野において一つの方向を目指しているように見える。そこに必要なことは、ビッグデータの処理技術とビッグデータのオープン化である。あらゆるデータがオープン化され、ビッグデータ処理技術を持つ者がビッグデータ指向型研究手法によりデータ解析を行うことで、これからどのような発見があるであろうか。それとも、そこには限界があるだろうか。その答えに到達するには、まずはすべての科学データのオープン化が進まねばならないのである。

