はじめに

ビッグデータは現在大きな関心を集めているが、学術的探求のためにはリトルデータも同じく重要である。データの絶対量が増大する中で、個々の観測を点検する能力は低下している。観測者は関心対象の現象からさらに一歩前に進まなければならない。新たなツールと新たな視点が求められている。しかし、ビッグデータは必ずしもより良いデータではない。観測者がデータの原点から離れれば離れるほど、どのように収集されたか、どのように処理、整理、変換されたか、どんな仮定や目的を念頭に置かれたか、といったそれらの観測が意味することを判定するのが難しくなるはずである。研究者はたいてい、自らきっちりと点検を行えるより小さな量のデータを好む。データが見つかっていないまたは見つけられない場合、研究者はデータを全く持たないと言えるだろう。

研究データは、開発される商品の枠をはるかに超えるものであるが、商品未満のものでもある。資金提供機関、雑誌、および研究機関のデータ管理計画、データ公開要求や善意に基づく方針は、分野を超えたデータあるいはデータ活動の多様性をおよそ考慮に入れていない。データが何であるかの例をリスト化する以外にデータを定義しようと努める方針はほとんどない。学術活動に携わる多くの利害関係者の競合する誘因や動機を考慮する方針はさらに少ない。データは同じ時にさまざまな人々にとってさまざまなものであり得る。データは、管理、蓄積、交換、結合、マイニング、そしておそらくは公開される資産であり得る。データは、管理、保護、または破壊される負債でもあり得る。データは慎重なまたは内密の扱いが必要で、公開されれば高いリスクを伴うものであり得る。その価値は直ちに明らかになることも、ずっと後でなければわからないこともあるかもしれない。一部には無期限にキュレートする投資に見合うものがあるが、多くは一時的な価値しか持たない。数時間あるいは数ヶ月の内に、技術進歩や研究の最前線はある種の観測が持つ価値を消し去ってしま

う。

学問におけるデータの役割を理解するための第一歩は、データは決してモノではないことを認識することである。データは独自の特質を備えた自然のオブジェクトではない。むしろ、データは、研究または学問のために現象の証拠として用いる、観測、オブジェクトあるいは他の実体の表現である。それらの表現は、研究者、状況、そして時を経て変化する。科学、社会科学、人文学の全般で、研究者は多くの場合それらのデータが何かについての合意のないまま、データの作成、利用、分析、解釈を行う。何かをデータとして概念化することそれ自体が学術的行為である。学問には、証拠、解釈、議論が必要である。データは目標に対する手段であり、その目標とは一般に雑誌論文、図書、会議論文、あるいは学術的承認に値する他の産物である。データを考慮に入れない研究はほとんどない。

ガリレオはノートにスケッチした。19 世紀の天文学者はガラス板に撮影した。現在の天文学者はフォトン（光子）のデータ取得にデジタル機器を用いている。一般消費者向けカメラで撮られた夜空のイメージは、スペースミッションによるイメージと調整可能である。データ記述とマッピングのための表現について、天文学者の合意が形成されているためである。天文学は、標準化、ツール、アーカイブにたくさん投資してきたので、数世紀にわたって収集された観測結果を統合できる。しかし、天文学の知識インフラは完全にはほど遠く、十分に自動化されていない。情報専門家が、天文学やその他のデータの組織化とアクセスの一元化に重要な役割を担っている。

文献とデータの関係は多様であり、この点が学術コミュニケーションの枠組みの中で研究データが活発に研究されている所以である。データの作成は計画的、長期的で、時の経過と共に価値が高まる資源の山を蓄積するかもしれない。また、発生の時点で入手可能であれば何であれ現象の標徴を捉える、その場しのぎで偶然に頼るものかもしれない。天文学、社会学、民族学のいずれであれ、どれだけ首尾よく研究プロトコルを明確化できたとしても、データのコレクションは次のデータの選択に影響を与える各局面の知見と共に確率的かもしれない。どんな分野においても、研究者に成るためにはデータの評価方法を学び、信頼性と妥当性を判定し、研究室、現場、またはアーカイブの状況に適

応するための方法を学習することが不可欠である。知見を報告する文献は、その知見を当該領域の文脈に組み入れると共に、読者の専門知識に加える。そこでは、議論、方法、結論を理解するために必要な情報が提供される。研究の追試のために必要な詳細は、読者はその分野の方法に通じていることを前提としているため、多くの場合省略される。追試と再現性はデータ公開の一般的な論拠であるが、選り抜きの分野でのみ妥当で、そしてそれらの分野においてさえ達成が難しい。どの学術生産物が保存に値するかを決定するのは、より難しい問題である。

データ管理、公開、共有のための方針は、学術研究におけるデータの複雑な役割を覆い隠し、領域内および領域間の活動の多様性をほぼ無視している。データの概念は、科学、社会科学、人文学の各領域でそして各領域内で、大きく異なっている。ほとんどの分野では、データ管理は教わるよりも身に付けるものであり、このことがその場限りの解決をもたらしている。研究者は多くの場合、自らのデータの再利用に大きな困難を感じている。そうしたデータを見知らぬ他者の予期せぬ目的のために役立てるのはさらに困難である。データ共有は、数分野だけで規範となっているにすぎない。実践するのがきわめて難しく、誘因はごく少なく、そして知識インフラへの大規模な投資が必要であるからである。

本書は、学者、研究者、大学の管理者層、資金提供機関、出版者、図書館、データアーカイブ、政策担当者をはじめとする広い範囲の利害関係者を読者として想定している。第Ⅰ部では、データの概念、学術、知識インフラ、研究活動の多様性に関して議論を呼び起こすよう、四つの章でデータと学問についての枠組みを明確化する。第Ⅱ部は三つの章から構成され、それぞれで科学、社会科学、人文学におけるデータの学問を調査する。これらの事例研究は並列構造となっており、それにより領域を超えた比較がもたらされる。まとめの部は、三つの章でデータに関する政策や実践を扱い、なぜデータの学問があまりに多くの困難な問題を提示するのかを探る。これらには、データの公開、供給、再利用、クレジット、帰属、発見、何をなぜ保持するのかといった問題が含まれる。

学問とデータには、長く、撚り合わさった歴史がある。どちらも新しい概念

ではない。新しいのは、学術のプロセスからデータを抽出し、それらを他の目的に活用しようという取り組みである。研究データの利用に関連するコスト、便益、リスク、報奨は、競合する利害関係者の中で再分配されつつある。本書の目標は、そうした関係者間に、より確かな情報に基づく充実した議論を呼び起こすことにある。問われているのは、学問の将来である。

クリスティン L. ボーグマン
Los Angeles, California
May 2014