訳者あとがき

かつて社会科学者はデータを収集するために訪問調査やアンケート回収に多大なコストを払ってきた．ところがインターネットが整備された現代では，多種多様なデータがネット上に公開されるようになっている．こうしたデータが，研究課題を解決するのに十分な質と量を備えていることも珍しくない．

本書のテーマである Web スクレイピングは，ネット上に公開されているデータを R などのプログラミング言語を使って収集する技術である．これにより，大量のデータを取得するのに何度もブラウザを起動する必要はなくなる．さらには，データの取得を定期的にコンピュータで実行できるようになる．ネット上のデータは増大し続けるので，管理する工夫も必要となる．大規模なデータを効率的に保存かつ操作するにはデータベースの導入が欠かせない．Web スレイピングもデータベースも背後にある技術は複雑だが，R を利用するとプログラミング経験の浅いユーザにも比較的簡単に実践できるようになる．また R には豊富なデータ分析・グラフィクス作成機能があり，さらにはレポート（プレゼンテーション）の作成までもカバーできる．

つまり本書は，単に R を使ってインターネットからデータを収集・保存する方法を紹介した入門書ではない．むしろ，その後の分析やレポート作成までのすべてのプロセスを効率化あるいは自動化する技法を実践的に解説した専門書である．分析に関連しては，データの前処理に役立つ正規表現や，ドキュメントから統計的な手法によって知見を引き出す技術であるテキストマイニングも取り上げられている．

本書の構成は大きく 3 つに分かれ，第 1 部では基本技術に焦点があてられている．R を利用することで Web スクレイピングとデータ操作は簡単に実現できるが，その背景技術について知識を深めておくと，構造の複雑な Web サイトに遭遇した場合でもデータをピンポイントに抽出できるようになる．続く第 2 部は，ここまで習得した技術の実践編となっており，現実に遭遇するケースのそれぞれに対処する方法が詳しく紹介されている．そして第 3 部は，やや複雑な研究課題を想定し，データの収集から分析，そして効果的なグラフィクスの作成までを詳細に解説している．

原著は Simon Munzert, Christian Rubba, Peter Meißner, Dominic Nyhuis 著 *Automated*

Data Collection with R: A Practical Guide to Web Scraping and Text Mining, Wiley，2015 年初版である．翻訳では，まえがきと第 1 章，2 章を石田が，そして 3, 4 章を工藤，5 章を石田，6, 7, 8 章を牧山，9, 10, 11 章を熊谷，そして 12 章から 17 章を高柳が分担している．訳者らが翻訳に着手したのは 2016 年のことであるが，この時点で原書に掲載されたコードのかなりの部分が正しく動作していなかった．そこで各章の担当者によって必要な訂正をコードに施している．また，これに伴い原書の文章についても訳者の判断で修正している場合があることをお断りしておく．

最後になるが，翻訳・出版にあたっては，共立出版の石井徹也氏に非常にお世話になった．また訳文の原稿整理にリッチハニカム社の協力をえることができた．ここに記して謝意を表したい．

2017 年 4 月

訳者を代表して

石田基広