まえがき

両親と Stefanie に
Simon
両親の愛と励ましに
Christian
Kristin，Buddy，Paul に
Peter
家族に
Dominic

World Wide Web はここ 20 年の間に急速に発展を遂げ，データの共有と収集，さらには公開方法は劇的に変化した．企業，公共機関，および個人ユーザらがありとあらゆるデータを発信し，また一方で新しいコミュニケーション・チャンネルが誕生しており，人間の行動にかかわる膨大な量のデータが記録されるようになっている．かつて社会科学においては，データが観測できず，また数も集められなかったことが問題であったのに，今ではデータに溢れかえっているのだ．ただ，これはこれで問題がある．たとえば，データを収集し分析する伝統的な技法では，この膨大かつ複雑に入り組んだデータに対処しきれない．こうしたデータを扱う必要性から，データをさばく達人たる「データサイエンティスト」という概念が誕生し，大学だけでなく企業からも求められる人材像となっているのである．

World Wide Web が隆盛をきわめる中，第 2 の技術動向が登場してきた．オープンソース・ソフトウェアの普及であり，R はその 1 つの例である．計量分析を主とする社会科学分野では，すでに R は広く利用されているソフトウェアである．R はユーザによる開発が積極的に進められており，パッケージという拡張機能が頻繁に開発・公開されている．R は単にフリーで使える統計解析用ソフトウェアであるというだけではない．R は他のプログラミング言語やソフトウェアと連携させられるので，どのような形態のデータが与えられてもスムーズに扱えるのである．

きわめて個人的な感想だが，社会科学における現在のデータ環境の問題点を要約すればこうなるだろう．

- 予算はわずかしかない.
- データを自分の手で集める時間も意思もない.
- だが，質の高い最新のデータを豊富に必要としている.
- またデータの収集から解析までの一連の作業を再現可能にし，レポートにまとめたい.

過去には，多種多様なデータ源から手作業でデータを集める必要があり，非常に不便な思いをしてきた．その際，コードの作成や，コピー&ペーストによってデータが損なわれることがないよう神経を擦り減らしてきた．さらにエラーが入りやすく，また面倒で死ぬほど退屈なデータの収集にもうんざりしていた．こうした経験から，われわれはデータ収集とレポート作成の両方を 1 つのソフトウェアで同時に実現する手法に移行し始めたのであった．R は，これを実現できるソフトウェア環境である．R を利用することで，日常のデータ解析作業の前後に生じる手間が大幅に改善される.

もちろん R が勝手に調査データを収集してくれたり，あるいは実験を即座にやってくれるという意味ではない．本書で紹介する技法によって，お金のかかる調査や実験，あるいはアルバイトの雇用からは解放されることになるだろうが，単にコスト抑制以上のことも学ぶことができるだろう．つまり，現代のさまざまなデータ分析を遂行するのに役立つ技法である.

筆者らは，オンラインのデータを収集することが，伝統的なデータ収集に比べ，単にコストパフォーマンスが優れていると考えているだけではない．むしろ，現代のアクチュアルな事象についてデータを収集する唯一の方法ではないかとすら考え始めている.

また作業をプログラミング言語によって実行することで，信頼性や再現性，時間の節約，そしてデータの質が担保される．生産性が高まってくると，コードを書くことや，煩雑で退屈きわまりない手作業をアルゴリズムで解決してしまうことが楽しくなってくるだろう.

要するに，本書で紹介する技法を習得するには手間がかかるが，その恩恵は計り知れない．データ分析の作業が大幅に改善され，効率的で質も高められるのだ.

本書からは学べないこと

目次から，本書の内容はおおよそ把握できるだろうが，読者が必要とする情報が本書で取り上げられているかどうか不明な場合もあるだろう．そこで，本書からは学ぶことのできない項目を挙げておこう.

まず本書では R の初歩は説明していない．R の入門書は多数出版されており，またネット上に多数掲載されている．そのため本書ではあえて解説しなかった．とはいえ，R を使ったことのない読者のために，次節で簡単に R について説明しているので安心していただきたい.

また，本書の内容は Web スクレイピングやテキストマイニングを実践するための唯一の解説書だと主張する気もない．これらに特化したさまざまなソフトウェアの使い方を解説

しているわけでもないし，R が最適のツールだというわけでもない．他のプログラミング言語を利用した方がよいケースもある．たとえば PHP，Python，Ruby あるいは Perl である．本書が読者の役に立つかどうかは，今現在 R を利用しているか，あるいは R を利用する意思があるかどうかによる．さもなければ，別のソフトウェアに頼った方がよいかもしれない．逆に R を使っているのであれば，それらの言語を習得することなく，慣れ親しんだ環境で課題を実現できるようになるだろう．

本書はデータサイエンスの専門書というわけではない．別に良書が多数出版されている．ただし，こうした入門書では，えてしてデータサイエンスを実行するために必要となるデータの取得方法については説明されていないことが多い．本書はデータサイエンスに必要な事前の準備を手助けするだけでなく，入手した情報を管理し，あるいはアップデートする手順も解説している．

そして重要なことだが，本書を熟読したとしても，読者が個別の関心を完全に解決する方法を得られるわけではない．データ収集のプロセスは課題ごとに異なるため，まったく同一の手順でデータが得られるわけではなく，時にはまったく異なるアプローチが必要になることもある．ただ本書で紹介する実行例やケーススタディにあるコードから，読者は自身が必要とするデータを収集するのに役立つコードを自身で書けるようになるだろう．

なぜ R なのか

本書で取り上げる課題を解くツールとして R が最適だと判断できる理由は多数あるが，ここでは特に以下を挙げておこう．

- R はフリーで利用しやすい．いつでもダウンロードしてインストールできる．高価な市販のソフトウェアを使わずにすむので，雇用主にソフトウェアの購入資金を求める必要もなくなる．
- R は統計解析を主な目的としたソフトウェアであるため，多種多様な分野で人気を博している．たとえば社会科学，医科学，心理学，生物学，地理学，言語学，そしてさまざまなビジネス分野で R は利用されている．
- R はオープンソースであるため，コードがどのように動くのか確認することができ，また簡単に修正することができる．つまり，ソフトウェアの修正が開発チーム以外の第三者にも開かれている．これは R 本体の開発に積極的に貢献できるということよりも，R の機能を拡張するパッケージをユーザ自身が開発できるため，こうしたパッケージの恩恵を多くの R ユーザが享受できるという側面が大きいだろう．公開されているパッケージの数は今も増え続けており，また頻繁に更新されている．CRAN Taskview (https://cran.r-project.org/web/views/) には特に人気のあるパッケージがジャンルごとに紹介されているので，参照するとよいだろう．
- R は通常の処理については十分に速い．SPSS や Stata などの統計ソフトウェアで複雑

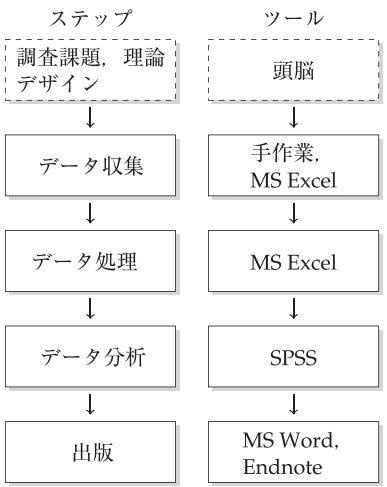

図 1　R を使わない場合の研究手順

なモデルを実行すると，結果が返ってくるまで数日かかるので，その間，休暇を取っていたという読者もいるかもしれない．R ではそのような心配はまず要らない．また「一度の実行時には 1 つのデータしか処理できない (one session，one data frame)」ということもない．さらに，R には処理を高速化する仕組みがあり，たとえば **Rcpp** パッケージを利用すると，C 言語のコードを R で実行できる．

- R にはデータを可視化する多種多様な機能が備わっている．データ収集という観点からはメリットを感じないかもしれないが，日常の分析業務では非常に重要な機能になるだろう．収集したデータの妥当性を検証する上で，データの可視化は重要で無視できないステップであることを本書では示す．グラフは巨大なデータを概観するのに非常に役立つツールである．
- R での作業は基本的にコマンドラインで行われる．これは初心者にはデメリットに思えるかもしれない．しかし，再現性を担保するにはこの方法しかない．マウスのクリック操作では無理なのだ．
- R は OS を選ばない．Windows，Mac OS，Linux のいずれでも動作する．
- R は他のソフトウェアに依存しない．最初から最後まで通して R で作業できる．本書を手に取った読者はプログラミングの愛好家ではないかもしれないが，あるトピックないしデータに強い関心があるだろう．ならば他の言語を学ぶことに時間を割いて，分析に着手するのを遅らせる必要はない．図 1 では研究の一連の手順をツールと関連させて示している．段階ごとにいろいろなソフトウェアを使い分けているのがわかるだろう．そのため，仮にデータ収集のステップにミスがあった場合，もう一度ソフトウェアを起動して，操作し直す必要に迫られる．R を使った分析プロセスでは，すべての手順が同じソフトウェア環境で実行されることになる（図 2）．Web スクレイピングとテキ

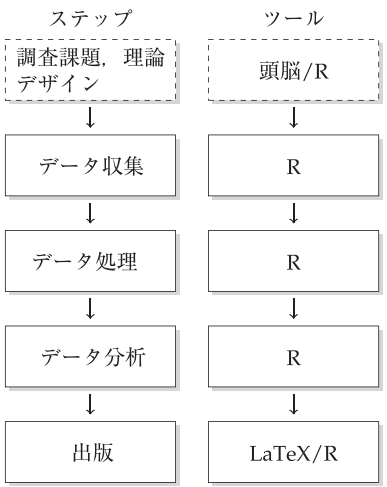

図 2　R を導入した場合の研究手順

ストマイニングに関連していえば，他のソフトウェアやプログラミング言語を学ぶ必要はない．習得すべきはマークアップ言語である HTML や XML の基礎であり，あるいは正規表現や XPath の仕組みである．そして実際の処理はすべて R で行うことができる．

R 初心者のための入門書

R 入門については多数の良書が出版されているが，ここでは以下を推奨しよう．

- Crawley, Michael J. 2012. *The R Book*, 2nd edition. Hoboken, NJ: John Wiley & Sons.
- Adler, Joseph. 2009. *R in a Nutshell. A Desktop Quick Reference*. Sebastopol, CA: O'Reilly. Teetor, Paul. 2011. *R Cookbook*. Sebastopol, CA: O'Reilly.

市販されている書籍に限らず，インターネット上にも多数の情報がある．たとえば Code School で公開されているオンラインコンテンツである `http://tryr.codeschool.com/` は秀逸だ．

さらに，Quick-R (`http://www.statmethods.net/`) は R の基本コマンドを調べるのに役立つだろう．また `http://www.ats.ucla.edu/stat/r/` からはフリーで利用できるデータやコードがダウンロードできる．

R は今も改良が続けられているソフトウェアなので，常に最新の情報に注意を払うとよい．Planet R (`http://planetr.stderr.org/`) では最新のパッケージ情報が得られる．R-Bloggers (`http://www.r-bloggers.com/`) では，さまざまな分野から R に関連したブログの記事を集めて紹介している．それらは経済学や生物学，地理情報など多岐にわたり，ほとんどの記事で投稿内容を再現するコードも公開されている．もちろん，自動データ収集に関する話題も投稿されている．

R 本体にヘルプは用意されているが，トラブルに遭遇した場合にはあまり役に立たないことも多い．むしろ Stack Exchange や Stack Overflow (http://stackoverflow.com) のようなオンラインフォーラムで相談する方が解決を得られやすい．問題が複雑な場合は，GitHub (http://github.com) で相談するのも良い手段だ．また，特定のトピックに特化したメーリングリスト (SIG) もある (http://www.r-project.org/mail.html)．さらには，オフラインの勉強会が読者の身近な街で開催されていることがあるかもしれない (http://blog.revolutionanalytics.com/local-r-groups.html)．

最後に，CRAN の Task View には Web 関連の最新の技術を R で実装したパッケージなどが紹介されているので，確認されたい (http://cran.r-project.org/web/views/WebTechnologies.html)．

本書の表記

本書は実践向けの指南書であり，実際に R を操作しながら読まれることを想定している．コードが明瞭になるよう，本書ではフォントを使い分けている．R のパッケージはボールド体で，また R の関数と変数はタイプライター体を利用している．本文中でも同様である．たとえば summary() と表記する．コードブロック全体を罫線で囲み，行番号を付与する．なお本書では R>で，コードの入力開始位置を表すプロンプトを表現している．出力部分にはプロンプトは表示されない．たとえば以下のようになる．

```
1 R> hello <-"hello, world"
2 R> hello
3 [1] "hello, world"
```

本書のサポートサイト

本書で利用したデータや掲載コード，また補足を http://www.r-datacollection.com で公開している．読者は本書に掲載されたコードをダウンロードして実行することができる．本書の練習問題への解答や，誤植についてもサイトで確認することができる．

免責事項

本書では Web スパイダーを取り上げていない．スパイダーとはインターネットを巡回して，ページのコンテンツを片っ端から収集するプログラムのことである．クローラーとも呼ばれる．こうしたクローラーである Google's Googlebot の手順を知りたいということであれば，本書は適切な選択肢ではない．本書では，特定のサイトの特定のコンテンツをスクレイピングするための技術に焦点を当てており，根こそぎコンテンツを取得するような振舞い

は想定していない．そして，本書で習得した技術をどのように利用しようとも，その責任は読者にある．本書で紹介しているコードを少し加工するだけで，Web サイトの管理者が困惑するような処理が行われてしまう可能性がある．そこで，Web からデータを収集する場合の基本的なマナーとして以下を提案しておきたい．

1. データ源を常に意識し，可能な限り報告書には明記する[1)]．
2. Web で取得したデータを再公開しようとする場合はコピーライトを確認すること．自身で収集したデータでない場合，再公開にあたってはデータのもともとのオーナーの許可を得る必要があるだろう．
3. 法に触れることはしない．データ収集でできること，あるいはできないことを知るには，法律関連のブログである Justia BlawgSearch (`http://blawgsearch.justia.com/`) を参照するとよい．ここで「Web スクレイピング」をキーワードに，最新の判決や規制に関する情報を得ることができる．最後の砦は 1990 年に設立された電子フロンティア財団 (Electronic Frontier Foundation; `http://www.eff.org/`) で，デジタル社会下の消費者あるいは公衆の権利を守るための団体だが，読者がここに頼るような状況にまで陥らないことを望まずにはいられない．

なお本書 9.3.3 項で，Web からコンテンツをスクレイピングする際に推奨されるマナーについて，もう少し詳細に説明している．

謝辞

本書の企画では多くの方々のお世話になった．ここでそれらの方々にお礼を述べたい．Peter Selb からは，第 3 のデータ収集技法に関する講座を開設することを勧められた．彼の励ましを受け，筆者らがそれぞれ散発的に行っていた実践をこうして 1 冊の本にまとめることができた．また，本書執筆の段階で内容についてコメントを寄せてくれた方々にも感謝したい．ここでは特に，Christian Breunig，Holger Doering，Daniel Eckert，Johannes Kleibl，Philip Leifeld，and Nils Weidmann らの名前を挙げたい．彼らのおかげで本書に適切な素材を集めることができた．

本書は，コンスタンツ大学で 2012 年と 2013 年の夏季に開催されたセミナー「第 3 のデータ収集技法」と「World Wide Web 時代のデータ収集」で利用されたテキストを原型としている．参加した学生たちからのフィードバックに感謝するとともに，一般的でないトピックや R，そしてやっかいな正規表現に耐え抜いた気力を称えたい．

また，マンハイムで 2012 年の 12 月に開催されたワークショップ「政治改革のためのデータに基づく研究：R による自動データ収集」と，2013 年 4 月にチューリッヒで開催されたワークショップ「R によるオンラインデータの自動収集」への参加者たちにも感謝した

1) ：この問題についての見解を筆者らは Hemenway & Calishain (2003) による Spidering Hacks (Hack #6) に負っている．

い．特にチューリッヒでの開催を援助してくれた Bruno Wüeste と Fabrizio Gilardi の名前をあげておきたい．

自動データコレクションというテーマで 1 冊の本を書き上げるのは本当に時間のかかる仕事であった．この間，筆者らは博士論文執筆にも追われていたが，本書のために多くの時間を割くことになった．筆者らの指導に当たっていた Peter Selb，Daniel Bochsler，Ulrich Sieberer，Thomas Gschwend らの忍耐と援助には深く感謝している．また，Christian Rubba には本書のために助成金を得るのにお世話になった (Grant Number 137805).

本書はまた，パッケージ開発者らの仕事に多くを負っている．彼らの絶え間ない努力によって，新しい研究手法が拓かれたといえる．こうした開発者全員の名前をあげる余白がないのは残念だが，ここでは Duncan Temple Lang と Hadley Wickham の 2 人には言及しておきたい．また本書を製作するにあたっては Yihui Xie が開発したパッケージのお世話になった．

出版企画から校正までの間には Heather Kay，Debbie Jupe，Jo Taylor，Richard Davies，Baljinder Kaur らがアドバイスを授けてくれた．

最後に，筆者らの友人と家族への謝辞を述べて序文を閉じたい．すなわち，Karima Bousbah，Johanna Flock，Hans-Holger Friedrich，Dirk Heinecke，Stefanie Klingler，Kristin Lindemann，Verena Mack，and Alice Mohr，Simon Munzert，Christian Rubba，Peter Meisner，Dominic Nyhuis に心から感謝したい．

訳注：原書のサポートサイトは http://www.r-datacollection.com/ であり，掲載コードは http://www.r-datacollection.com/materials/ にある．ただし原書の記述は一部が内容的に古くなっており，翻訳の時点で掲載コードのいくつかが動作しなくなっていた．そのため，これらを修正したコードを翻訳サポートサイトとして公開している．

https://github.com/IshidaMotohiro/ADC

GitHub になじみのない読者は，サイト右上にある Clone or download ボタンを押して，Download ZIP を選んでパソコンに取り込んでほしい．ファイルは文字コードを UTF-8 に設定してあるが，ダウンロードファイルに含まれる Windows.zip を解凍すると，文字コードを CP932（いわゆる Shift-Jis）に変換したファイルがあるので，そちらを利用されたい．
ただし，公開されたスクリプトを実行することで発生したトラブルに対して，共立出版また訳者ともいかなる責任も取れないので了承されたい．