JLSC

ISSN 2162-3309 | JLSC is published by the Iowa State University Digital Press | http://jlsc-pub.org

Volume 12, 1 (2024)

Brief Reviews of Books and Products

Data Management for Social Scientists: From Files to Databases

Marla Hertz

Hertz, M. (2024). Data Management for Social Scientists: From Files to Databases [Brief Reviews of Books and Products]. *Journal of Librarianship and Scholarly Communication*, 12(1), eP17642. https://doi.org/10.31274/jlsc.17642

ISSN 2162-3309 10.31274/jlsc.17642

BRIEF REVIEWS OF BOOKS AND PRODUCTS

Book review: Weidmann, N. B. (2023). ***Data Management for Social Scientists: From Files to Databases*. Cambridge University Press. 228 pp. doi:10.1017/9781108990424. ISBN 9781108990424 (open access e-book).**

Data Management for Social Scientists: From Files to Databases delivers on its promise to serve as an introduction to research data management (RDM) in the social sciences by applying computer science tools to political science research questions. The author, Nils Weidmann, a professor of political science at the University of Konstanz in Germany, was trained in both computer and political science. His expertise in these two fields coalesce in this book to produce strong practical advice. Each chapter applies a data management concept to real-life examples using publicly available data sourced primarily from the United States and Europe.

This book centers on the processing steps between data collection and analysis, a critical but oft overlooked phase of research. The book purposely avoids study design, data collection methodology, data analysis, and data visualization, topics that have been reviewed elsewhere. As a result, the book has a defined narrow purpose: teach best practices to transform collected data into an analysis-ready format. Proper execution of this step is critical, as publishers and funders increasingly expect researchers to share raw data, metadata, and code alongside final research outputs.

Although this book draws on examples from political science, it is written to engage readers regardless of their prior knowledge. Additionally, data processing is by nature more universal than the project design and data analysis steps, and thus can be broadly applied within the social sciences and beyond. Therefore, the audience for this book is not limited to social scientists; it appeals to any researcher working with tabular data who seeks to educate themselves on data processing. I personally would have found this book valuable early in my career when I was struggling with table design and file organization. The book will also be of interest to anyone in a research support role, such as research data services librarians, who may provide data management and curation services. The book includes tutorials using popular open-source R software and extension packages that enhance the learning experience. While not essential to understand the book, the reader will gain more by following along. As someone with only a beginner-level proficiency in R, I was able to easily carry out the prompts even if I did not fully understand all the commands. This was aided by the addition of a companion website where readers can download a specific instance of R to run all the packages and data files necessary to complete the tutorials.

Data Management for Social Scientists is written so that it may be read in full or piecemeal. Each chapter addresses a single topic with a practical example and concludes with a summary to reinforce key points. The topics are introduced using both conventional social science terminology and corresponding database jargon to help the reader make connections. The book is organized into four parts, which gradually increase in complexity in terms of both the data types covered and computational tools employed.

Part 1 is a conceptual and practical introduction. It defines what data is and outlines the benefits of organizing data using specific tabular structures. It communicates the motivation behind writing the book and contextualizes where data processing fits into the greater data life cycle. The introduction also includes a chapter dedicated to software setup. For readers already convinced of the merits of data management, this section may be skipped or skimmed. However, for the inexperienced reader, this section serves as a solid introduction to the topic.

Part 2 covers strategies for processing data stored in files. It gives a basic overview of common file formats for tabular data and discusses best practices for managing data using Microsoft Excel, a familiar tool with limited capabilities for this purpose. Later chapters cover how to carry out analogous functions using both base R and the tidyverse package to import, merge, and aggregate tabular data. The author strongly recommends the usage of R for its capabilities to maintain a log of all actions performed on data, a practice that supports the rigor, reproducibility, and transparency of research.

Part 3 covers strategies for processing data stored in databases. To do so, it introduces relational databases, namely the PostgreSQL database management system (DBMS). Although this specialized tool may not be necessary for every researcher, it is useful for projects that deal with large amounts of data or require simultaneous access by collaborators. This section covers how to send data to and from a DBMS, assign keys, and create indices to reduce computing time.

While Parts 2 and 3 deal with pure tabular data, Part 4 addresses three specialized data types: spatial, text, and network. These data types can be converted into tabular forms to facilitate analysis. The real strength of this section is that it builds upon prior chapters. The topics in Parts 3 and 4 are organized from simple to complex and include options to process data in either R or the relational database.

The book ends with a rich conclusion chapter that provides a list of recommended practices. It also discusses two common data management challenges: managing collaborative projects and public dissemination of research data. Overall, the book argues that the structure of tabular data and how files are organized is paramount to facilitate analysis, foster collaborations, and share results. It empowers the reader to step outside their comfort zone of using spreadsheets and switch to R or a DBMS. This book spotlights the growing significance of RDM in the social science and is an excellent resource for those looking to embrace the movement.

AUTHOR BIOGRAPHY

Marla Hertz, PhD, is an associate professor and the research data management librarian at the University of Alabama at Birmingham. She oversees training and instruction on data management best practices for the university. Prior to entering librarianship, she was an academic researcher in microbiology and global health.