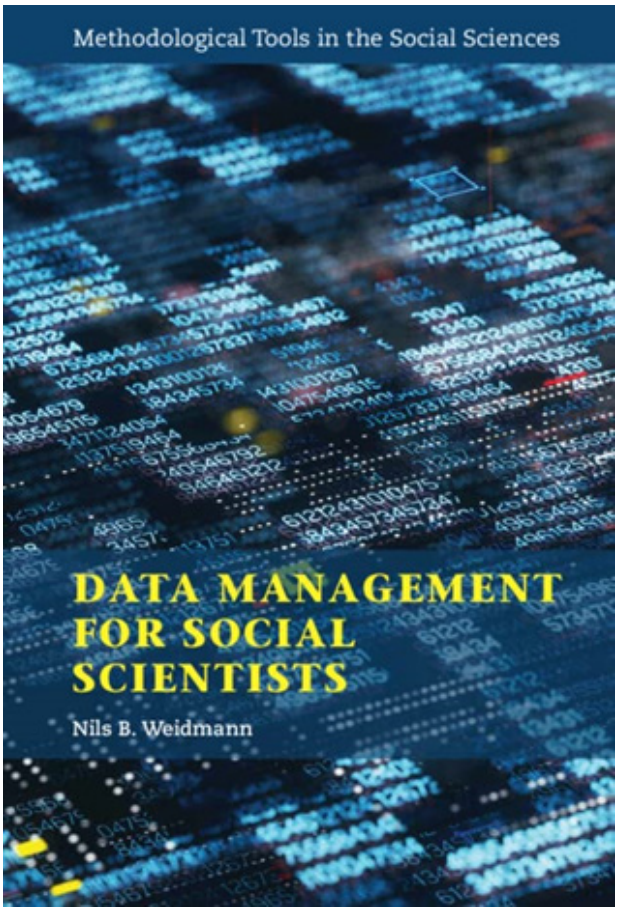
J Data Sci. Info. Citation Studies., 2023; 2(3):273-278.
https://www.jcitation.org

Book Review

Data Management for Social Scientists: From Files to Databases



Data Management for Social Scientists: From Files to Databases; by Nils B. Weidmann. Cambridge University Press, 2023. ISBN: 9781108990424, Open Access Book, DOI: 10.1017/9781108990424.

INTRODUCTION

The "data revolution" has presented social science researchers with a plethora of exciting new avenues of inquiry. It is becoming increasingly possible to record traces of social and political interactions digitally, which results in enormous volumes of fresh data that are made available for research. The concept of big data has gained significant attention in scientific disciplines (Floridi, 2012). However, despite the potential for big data and novel data sources to revolutionize research in the social sciences, their practical value and relevance to certain subject areas are not fully reached due to a lack of awareness (Kitchin, 2014; O'Brien, Sampson, & Winship, 2015). While some domains of the social sciences have a more unified approach to research using well-defined scientific methods and hypothesis testing to verify or falsify theories, other subjects, such as sociology (Leavy, 2014) and human geography, have a broader range of philosophical approaches (Kitchin, 2013; DeLyser, & Sui, 2014). However, with access to increasingly larger and more diverse datasets, social scientists have progressively turned to data-driven research to examine intricate hypotheses, enhance predictive accuracy, cultivate novel knowledge, and advance their comprehension of social dynamics. The field aims to extract meaningful insights and discern patterns from the vast and intricate datasets, with

DOI: 10.5530/jcitation.2.3.37

Publishing Partner : EManuscript Tech. [www.emanuscript.in]

the overarching objective of comprehending social phenomena, predicting future trends, and contributing to informed policy decisions. Data-driven social research holds the potential to furnish fresh insights and influence policy formulation across diverse domains encompassing public health, education, politics, and economics. Nonetheless, the attainment of dependable findings within this realm necessitates meticulous adherence to methodological rigor, proficient data management and analysis, ethical contemplation, and the judicious interpretation of research outcomes.

Nils B. Weidmann's "Data Management for Social Scientists: From Files to Databases" is a valuable contribution to the field of social science research. In a time where vast amounts of data are available for research, this book offers practical guidance on how to manage, process, and analyze data using established techniques and technologies from computer science. The book covers the entire range of data management techniques, from flat files to database management systems. The book aims to guide social scientists in managing their research data.

The book is divided into five distinct sections: *Part I (Introduction)* provides an introduction to data management and its importance in social science research.In this introductory section, the book delves into the role of data preparation and management within the research cycle. It also provides an overview of the software used, primarily focusing on R and database management systems (DBMS). The book emphasizes the importance of conceptualizing data as a combination of information and structure, covering data structures such as tables. *Part II (Data in Files)* covers the basics of data management, including file organization, documentation, and data cleaning. Real-world examples are introduced in the second part of the book, which focuses on file-based data processing, file formats, and data management using tools like MS Excel and R. *Part III (Data in Databases)* delves into more advanced topics such as relational databases, programming, and data visualization. This section explores specialized data systems, specifically relational databases, addressing the technical aspects of handling large datasets collaboratively. In *Part IV (Special Types of Data)*, a comprehensive examination is presented, focusing on various data types, with particular emphasis on Spatial Data, Textual Data, and Network Data, their multifaceted applications with R and database system usage. Within *Part V (Conclusion)*, the ultimate section of this study, an exploration of best practices in the realm of data management is undertaken, encompassing both general recommendations and specific strategies such as collaborative efforts in Research Data Management (RDM), as well as the dissemination of research data and code to facilitate reproducible scientific endeavors. The book's goal is to offer a comprehensive overview of various data management tools and techniques, equipping readers with the knowledge to choose the best software and workflow for their research needs.

The book's orientation is towards social scientists who may not have a background in computer science or programming but need to manage data for their research data. The chapterization of the book is clear and logical, making it easy for readers to navigate and find the information they need. Overall, the book is a valuable resource for social scientists looking to improve their data management skills in the age of data driven research. The book is structured into five distinct sections, each comprising a total of fourteen chapters. This organization is designed to systematically investigate and comprehensively cover all facets related to Research Data Management, analysis, and visualization. The initial three chapters constitute the introductory section.

Chapter 1 of the book *(Motivation)* primarily discusses the transformation of empirical social science research methods in recent decades. It highlights the need for efficient data processing in the context of modern digital technology, emphasizing the importance of preparing data for analysis. The chapter outlines the research cycle, comprising data collection, data processing, and data analysis, and explains the significance of data processing as a distinct step. It also covers the advantages of proper data documentation, including enhanced convenience, replicability, transparency, scalability, and versatility. The chapter introduces the concept of data management systems (DBMS) as a means to address the limitations of file-based data storage. It identifies the target audience for the book, which includes social scientists working with empirical data from various disciplines, and acknowledges the requirement of basic R knowledge for readers to effectively use the book's content. The chapter does not provide a fundamental introduction to R but focuses on practical data management concepts applicable to a wide range of statistical software.

Chapter 2 *(Gearing Up)* of the book lays the foundational groundwork for effective data management. It introduces the essential tools and environment setup for readers, commencing with the installation of the R statistical toolkit and emphasizing the added value of RStudio for efficient data analysis. The concept of project environments and file organization is presented to facilitate the handling of diverse datasets. Readers are guided on R package management and the use of R environments. Additionally, the chapter delves into the importance of the PostgreSQL database system for advanced data management needs, addressing key installation considerations. In summary, the chapter prepares readers with the technical tools and knowledge necessary for forthcoming data management exploration.

Chapter 3 *(Data = Content + Structure)* introduces fundamental concepts of data and highlights tables as the primary data structure in the social sciences. It emphasizes that data is information systematically coded to represent real-world aspects and discusses how to distinguish data content from its logical structure. Tables, which consist of rows (observations) and columns (variables), are presented as the primary structure for social science data,

ensuring comparability of different entities, such as countries or years. The chapter demonstrates how to access, update, add, and delete data within tables in the R environment. Additionally, it advises that designing tables that grow downward, not sideways, is more effective, ensuring a consistent and non-redundant data structure. It also suggests that splitting data into separate tables is beneficial when dealing with different types of entities to prevent data redundancy and inconsistency. The chapter concludes by highlighting the importance of considering table structure when collecting and processing data, as it significantly influences the efficiency and organization of research workflows.

Chapter 4 *(Storing Data in Files)* discusses the importance of storing data in files, which is a fundamental aspect of working with data in social science projects. It explains that data is typically stored temporarily in a computer's main memory when being processed in software like R, Excel, or Stata, but for long-term and persistent storage, data needs to be saved in files on the hard disk or other storage solutions. The chapter introduces various file formats for storing tabular data, emphasizing the CSV format as a convenient and flexible option due to its compatibility with most software packages. It also mentions other file formats like Excel, Stata, and SPSS, as well as R's own file formats, which are specialized and should be used only when necessary. The chapter concludes with recommendations, including understanding file-based storage, familiarizing oneself with different file formats, organizing directories and files consistently, and preferring generic, software-independent file formats for data storage to ensure accessibility and long-term usability.

Chapter 5 *(Managing Data in Spreadsheets)* introduces the use of spreadsheets, focusing on Microsoft Excel, for data management in social science research. Spreadsheets are versatile tools for entering data, performing calculations, and visually presenting information. However, the chapter highlights significant limitations when it comes to managing research data. Spreadsheets lack a predefined structure, making it challenging to maintain consistency in tables and ensure data accuracy. Manual interactions and the absence of a record of data processing steps also hinder replication and can lead to errors. The chapter advises against relying on spreadsheets for most social science data management tasks due to these limitations, but it acknowledges their continued use in research. The chapter concludes with recommendations for using spreadsheets more effectively, emphasizing practices like using one table per sheet, maintaining a rectangular table structure, employing proper variable names, validating data values, and avoiding the use of formatting elements for storing information to minimize potential issues in data management and analysis. While spreadsheets may have certain applications, especially in manual data entry, these recommendations aim to help users mitigate common problems and challenges when transitioning data for analysis in statistical tools like R or Stata.

Chapter 6 *(Basic Data Management in R)* delves into the basics of data management using the R statistical toolkit, emphasizing the core functionality of R for data storage, processing, and analysis. While spreadsheets have their limitations, R provides a powerful alternative where commands are entered in the R programming language to process data, offering script-saving capabilities and better Reproducibility. The chapter focuses on base R features and functions for data management, highlighting the importation of data from various sources into R data frames, merging datasets, and aggregating data. While this chapter covers essential data processing techniques within R, it also foreshadows the next chapter's discussion on the "tidyverse," a more user-friendly approach to data manipulation. Key takeaways from this chapter include the importance of understanding base R, the significance of data frames as R's primary tabular data structure, the need for careful data importation to preserve column types, and the usefulness of simple packages like do By for standard data manipulation operations. This foundational knowledge in data management with R is a significant improvement over spreadsheet-based workflows, offering researchers more control and Reproducibility in their data handling processes.

Chapter 7 *(R and the tidyverse)* introduces the concept of the tidyverse, a collection of interconnected R packages designed for more efficient and user-friendly data management compared to base R. The tidyverse operates on tabular data structures and uses a syntax that's easier to remember, making code more readable and reproducible. This chapter provides an overview of the tidyverse's core packages, tidyr and dplyr, for managing and manipulating data. While the tidyverse extends beyond data management into areas like data visualization with ggplot2, this chapter focuses on two specific packages for data-related tasks. The tidyverse employs a pipe operator that simplifies the workflow for various data operations, enhancing code clarity and minimizing the need for intermediate results. While the tidyverse is recommended for most data tasks due to its elegance and efficiency, understanding both base R and the tidyverse is valuable for versatility. Key recommendations include using the pipe operator, avoiding mixing approaches within a script, and being aware of potential conflicts and installation issues when using the tidyverse. Additionally, it emphasizes letting R handle conversions between long and wide tables rather than manual adjustments.

Chapter 8 *(Introduction to Relational Databases)* delves into the world of relational databases, presenting an alternative to the standard file-based data storage common in social science research. In this chapter, the focus is on relational database management systems (DBMS), which are designed to store, manipulate, and efficiently retrieve data. The chapter explains how relational databases organize data into tables linked to one another, emphasizing data integrity and providing efficient data operations. This database approach offers several advantages over

file-based storage, including organized data, data consistency, improved performance, and collaborative work with multiple researchers. Users access these databases using Structured Query Language (SQL) to interact with the data. While the chapter introduces relational databases with a single table, it sets the stage for expanding to multiple tables in the next chapter. Key takeaways from the chapter include an understanding of client-server setups, recognizing the differences between SQL and R programming philosophies, distinguishing between SQL and R's built-in database functions, and understanding the different functions for sending commands and retrieving data in R.

Chapter 9 *(Relational Databases and Multiple Tables)* explores the utilization of relational databases and their application involving multiple tables. The chapter extends the single-table example by incorporating a second table containing information about political parties alongside the existing elections table. By incorporating data from the PopuList project, which identifies populist parties in Europe, the chapter exemplifies how this additional data can be used to analyze electoral gains in Europe over recent years. Working with multiple tables in a relational database presents specific challenges. One key aspect involves conceptualizing the database structure, determining the necessary tables and their contents, emphasizing the importance of eliminating redundant data. The incorporation of "Entity-Relationship" models, which hinges on real-world entities and their inter relatedness, plays a pivotal role in shaping the design of the relational database. Technical challenges include dynamic data combination through SQL joins and maintaining data consistency across tables. The use of foreign keys and primary keys is crucial for preserving referential integrity within the database. However, the chapter emphasizes the importance of structuring data effectively, using primary keys for all tables, making use of integrity checks, and avoiding unnecessary merging of tables. The subsequent chapter addresses collaborative data work and efficient data search, concluding the basic introduction to relational databases in this book.

Chapter 10 *(Database Fine-Tuning)* delves into the operational aspects of relational databases, building upon the foundational knowledge presented in Chapters 8 and 9. While previous chapters focused on the conceptual aspects of data and its storage, this chapter explores two crucial operational dimensions. The *first* aspect addresses the efficient management of large datasets, highlighting how databases, such as PostgreSQL, outperform file-based workflows when handling extensive data. The introduction of search indexes is examined, providing the means for the rapid retrieval of specific entries within a table. These indexes, akin to a book's index, expedite data retrieval by directly pinpointing relevant records. Indexes offer a substantial performance boost, particularly when dealing with large datasets, as they operate in a logarithmic time complexity. The chapter emphasizes that while indexes enhance retrieval, they do increase database storage size. Additionally, adding new records or updating indexed columns may slightly affect performance, but this can be mitigated by creating indexes after data insertion. The *second* operational dimension relates to collaborative data management, allowing multiple users to access and manipulate the database. Relational databases offer fine-grained access control, enabling the customization of read and write privileges for various users. The chapter explores user privilege management within PostgreSQL, showcasing its value in collaborative scenarios. This feature is particularly useful for research teams working with shared data and various levels of access. This chapter underscores the importance of fine-tuning relational databases to optimize performance and enhance data management. Lessons learned to include the significance of indexing large datasets, transparent user privilege management for collaborative work, challenges in tracking data changes, and the utility of graphical database client tools for direct database access. Relational databases prove to be powerful tools for data storage and processing, especially in research settings.

Chapter 11 *(Spatial Data)* delves into the world of spatial data, a fundamental component of Geographic Information Systems (GIS). Spatial data are characterized by their association with geographic coordinates, allowing for precise location representation on Earth's surface. This chapter primarily focuses on vector spatial data, which includes points, lines, and polygons, each containing both attribute information and spatial coordinates. The chapter explores the essential distinction between geographic coordinate systems and map projections, crucial for transforming the Earth's three-dimensional surface into a two-dimensional map. Geographic coordinate systems rely on radial coordinates, such as longitude and latitude, whereas map projections involve methods for mapping Earth's surface onto a flat plane, with various projections optimized for specific purposes. While specialized GIS software exists (like ArcGIS, QGIS), the chapter underscores the utility of utilizing R and PostgreSQL, notably through the PostGIS extension, for managing and analyzing spatial data. An applied example involving spatial overlay operations is presented, showcasing the spatial join process for linking entries based on spatial relationships. However, this chapter provides a comprehensive introduction to spatial data, emphasizing vector data and their integration with GIS software. The chapter highlights the significance of choosing appropriate coordinate systems and projections. It illustrates the practicality of utilizing R and PostgreSQL with PostGIS in social science research for efficient spatial data management and analysis. The chapter concludes with recommendations for further exploration and data visualization to enhance spatial analysis capabilities.

Chapter 12 *(Text Data)* delves into the realm of textual data, distinct from the tabular data structure explored in prior chapters. Text data encompasses written statements, reports, and transcribed spoken language, making it crucial for comprehending a diverse

range of topics relevant to social scientists. From parliamentary debates to social media hashtag analysis, and journalism framing of social issues, text data plays a pivotal role. The chapter outlines how text analysis has gained prominence within the social sciences, transitioning from more complex linguistic analyses to simplified approaches like word frequency analysis and keyword searches. The focus of this chapter is the preparatory stages involved in handling textual data prior to analysis. These stages encompass data representation, storage, and retrieval. Textual data is frequently organized into collections of documents, forming a corpus, with each document carrying associated metadata. The chapter demonstrates the translation of this textual data and its metadata into a structured tabular format, facilitating further analysis. It distinguishes structured (metadata) from unstructured (text) data, emphasizing the intricacies of extracting information from unstructured data. The chapter elaborates on digital storage methods for textual data, outlining various file formats and storage approaches, including CSV for individual files and single-file storage with variations. An applied example involving United Nations General Debate speeches over time illustrates the practical application of handling text data using R and PostgreSQL. The chapter concludes by highlighting the two main text processing approaches: treating text as extended strings and employing Natural Language Processing (NLP) methods, discussing the relevant packages and databases for such operations. In summary, this chapter provides a foundational understanding of textual data, emphasizing the initial stages of data preparation and storage, with insights into different processing approaches. It equips researchers in the social sciences with the knowledge required to work effectively with textual data and lays the groundwork for subsequent text analysis methods.

Chapter 13 *(Network Data)* explores network data, a critical component in the social sciences that goes beyond individual entities and focuses on the relationships between them. These relationships are represented as network structures, which are a form of graph consisting of nodes (entities) and edges (relations). The chapter illustrates the two primary types of graphs: un-directed and directed. Un-directed graphs connect nodes symmetrically, while directed graphs attribute direction to edges, allowing them to represent flows and one-way relationships. To enhance the representation of network data, the chapter introduces attributes attached to nodes and edges. It explores two methods for storing network data: adjacency matrices and adjacency lists, with the latter being the more versatile choice, resembling structured tabular data. The chapter employs the igraph package in R and a relational database, specifically PostgreSQL, to process and analyze network data. A practical application concerning the relationship between democracy and international trade highlights the value of network data in addressing complex research questions. The chapter concludes by providing recommendations for working with network data, emphasizing the preference for adjacency lists, the use of tabular data formats, and the importance of consistency between node and edge datasets. Additionally, it suggests exploring graph databases for handling large networks efficiently.

Chapter 14 *(Best Practices in Data Management)* summarizes the key takeaways from the book's exploration of data management techniques and tools. It highlights two fundamental aspects of good data management: data structure and data manipulation. Data structure is crucial, and in the social sciences, tabular data formats are common. While spreadsheets provide flexibility, they may lead to data inconsistencies. Relational databases offer a more structured approach by specifying column types and table structures. The chapter emphasizes the preference for "long" tables over "wide" tables and the importance of distributing data across separate tables linked by unique identifiers to avoid redundancy. Data manipulation workflow is another essential aspect. Transparency and replicability are crucial, and the use of scripts, like R, for documenting data processing steps is encouraged. The chapter also discusses when to choose file-based workflows versus database-driven ones, which may be necessary for large or complex datasets. It provides recommendations for collaborative data management, favoring version control systems over shared drives, and highlights the challenges and ethical considerations when disseminating research data and code. It recommends using specialized portals like Dataverse and the Open Science Foundation for data sharing. The chapter concludes by reflecting on the value of comprehensive data management in handling increasingly complex and large datasets in the social sciences and emphasizes the importance of structured, tabular data while hinting at the growing relevance of unstructured data, particularly in text analysis.

An Appraisal

This work presents a new set of issues for how social scientists organize and process the data from their research. This book covers the full spectrum of methods for managing data, from flat files to database management systems, and everything in between. Drawing on a diverse assortment of real-world applications, it illustrates how well-established strategies and tools derived from the field of computer science can be utilized in the field of social science research projects. The book begins with basic tools like spreadsheets and continues on to file-based data storage and processing before moving on to more advanced data management technologies like relational databases. The book relies on a range of software tools, including R and RStudio for statistical analysis. Microsoft Excel is mentioned but cautioned against data management practices. The book provides detailed installation instructions and sample datasets, which can be accessed on the companion website. The book is designed for users to progressively explore software tools, with the possibility to start with R and RStudio and later delve into more advanced database systems as needed. In the last section of the book, more

sophisticated subjects are discussed, such as network data, text as data, and spatial data.

One of the strengths of the book is its accessibility. It is part of the "Methodological Tools in the Social Sciences" series, which aims to provide practical instruction for applying methods and getting them right. The book strikes a balance between the theory underlying the methods and their implementation. The author provides extensive examples of applications of the methods covered in the book and makes technical code and data available to aid in replication and extension of the results. In addition, the book excels in its emphasis on the organization and structure of data. Relational databases force us to think about data structure much more than we commonly do in social science data analysis. We need to explicitly define data structures before we can use them, and this makes us think about what information the individual tables should contain, how many we need, and how they are linked. Even if readers later move on to less-structured data, they will do so being fully aware of the strengths and weaknesses of the different approaches. The book also covers some basic techniques for managing large amounts of data, which are essential as datasets become bigger. Relational databases allow the author to illustrate how a client-server setup works, which is becoming increasingly necessary as data management becomes more complex due to the amount of data that needs to be processed.

Although the book offers an extensive array of content and provides many examples, it can be overwhelming for readers who are new to the topic. The technical language used in the book may be difficult for those who are not well-versed in computer science. Additionally, while the book introduces advanced subjects like spatial data, text data analysis, and network data, the level of detail provided on these topics is not as comprehensive as that offered for fundamental data management principles. Consequently, readers interested in delving deeper into these areas may need to supplement their learning with additional resources.

The book offers limited discourse concerning the ethical dimensions that emerge in the context of data management, including matters pertaining to data privacy, informed consent, and data sharing. In light of the potential for social science research to exert tangible effects on the real world, this represents a noteworthy omission. While the book briefly acknowledges the significance of research Reproducibility, it would be advantageous to incorporate a more comprehensive examination of methodologies for guaranteeing that research outcomes can be replicated by others. This has become especially pivotal in light of the growing emphasis on open science and research transparency.

This book is a timely intervention in a field that is rapidly advancing, and it will be useful for academics and researchers in the social sciences, particularly in the fields of business and management, internet studies, data sciences, political studies, urban sociology, law, media and cultural studies, sociology, and cultural anthropology studies. In addition to that, it will be of tremendous interest to new researchers who are looking for insights on social research. However, the book is a valuable resource for social scientists who want to manage, process, and analyze data using established techniques and technologies from computer science. The book provides practical guidance on how to manage data, from simple tools such as spreadsheets to more powerful data management software such as relational databases. It is accessible and discursive, and makes technical code and data available to aid in replication and extension of the results. Overall, this book is highly recommended for social scientists who want to improve their data management skills.

REFERENCES

DeLyser, D., & Sui, D. (2014). Crossing the qualitative-quantitative chasm III: Enduring methods, open geography, participatory research, and the fourth paradigm.*Progress in Human Geography*,*38*(2), 294–307. https://doi.org/10.1177/0309132513479291

Floridi, L. (2012). Big data and their epistemological challenge.*Philosophy and Technology*,*25*(4), 435–437. https://doi.org/10.1007/s13347-012-0093-4

Kitchin, R. (2013). Big data and human geography: Opportunities, challenges and risks.*Dialogues in Human Geography*,*3*(3), 262–267. https://doi.org/10.1177/2043820613513388

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts.*Big Data and Society*,*1*(1). https://doi.org/10.1177/2053951714528481

Leavy, P. (Ed.).(2014).*The Oxford handbook of qualitative research*. Oxford University Press.

O'Brien, D. T., Sampson, R. J., & Winship, C. (2015). Ecometrics in the age of big data: Measuring and assessing "broken windows" using large-scale administrative records.*Sociological Methodology*,*45*(1), 101–147. https://doi.org/10.1177/0081175015576601

Weidmann, N. B. (2023).*Data management for social scientists: From files to databases*. Cambridge University Press.

Reviewed by

Sanjoy Kar
Institute of Development Studies Kolkata (IDSK),
Salt Lake City, Kolkata, West Bengal, INDIA.
Email: sanjoy@idsk.edu.in
ORCID: 0000-0001-5050-5945