Preface

More than a decade ago, while I was still a PhD student, Mike Ward encouraged me to develop a book project about relational databases for social science applications. The book proposal was not successful, but I never completely abandoned the idea. Later in my career, when working with many excellent students, I realized that there is still a huge need to establish data management as part of our quantitative social science curricula. Most of the training we offer in political science focuses on (oftentimes advanced) methods for statistical analysis and causal inference, but does not really help students get to the datasets required for this. As a result, "many social scientists will find themselves 'hacking together' datasets in a fundamentally ad hoc way," as one reviewer for this book commented on the status quo in our field. I hope that this book contributes to improving this.

In comparison to the original idea, the focus of this book has been expanded considerably, beyond relational databases. The first half of the book describes different tools to manage data in a file-based workflow, without interfacing with a dedicated database system. Yet, more technically advanced readers will wonder why I focus so much on databases in the second half of the book, given that this is – at least by computer science standards – a fairly old technology. Still, relational databases continue to be around, and they allow me to cover a number of key learnings that easily generalize beyond this technology. First, with the need to explicitly define data structures (tables) before we can use them, databases force us to think about data structure much more than we commonly do in social science data analysis. What information should the individual tables contain, how many do we need, and how are they linked? There are different

https://doi.org/10.1017/9781108990424.001 Published online by Cambridge University Press

ways to do this, and some are better than others. Even if readers later move on to less-structured data – which is becoming more and more common also in the social sciences –, they will do so being fully aware of the strengths and weaknesses of the different approaches. Relational databases also allow me to cover some basic techniques for managing large amounts of data, which are essential as our datasets become bigger. Indexing a table is a standard operation in a database, and we can nicely illustrate what we gain from it. Last, databases are a great way to demonstrate how a client-server setup works. As our data management becomes more complex, for example due to the amount of data we need to process, there is an increasing need to perform certain tasks on specialized servers rather than one's own personal computer. This makes it necessary to interact with these servers, which is something we do in this book using a relational database management system.

This book benefited from the help and support by several people and institutions. The initial development of the material was funded by the German Federal Ministry of Education and Research under the "b^3 – beraten, begleiten, beteiligen" project, and Lukas Kawerau, with his extensive skills as a computational social scientist, was essential in getting a first draft off the ground. I am grateful to Lars-Erik Cederman and the International Conflict Research group at ETH Zurich for hosting me during the Winter term 2019–2020, which gave me the opportunity to work intensively on this project. During this time, Luc Giradin with his joint computer science and social science background provided many useful comments and suggestions. At Konstanz, the members of my group (Frederik Gremler, Anna-Lena Hönig, Eda Keremoğlu, Sebastian Nagel, Stefan Scholz and Patrick Zwerschke) and the students in my courses on data management (summer term 2020 and 2021) were critical and constructive readers, and contributed greatly to the improvement of this book. Also, an early presentation of this project at our department's Center for Data and Methods (CDM) proved to be extremely helpful in setting the general scope, clarifying the main goals of the project. I am very grateful to Guy Whitten and Paul Kellstedt for including this book in the Methodological Tools in the Social Sciences series at Cambridge University Press, and for supporting me with their expertise and advice. The open access publication of this book was made possible through financial support from the University of Konstanz's Open Access Fund and the the Cluster of Excellence "The Politics of Inequality"

(EXC-2035/1–390681379). Lastly, I want to thank all the developers of the datasets and the free software used in this book. We often fail to realize that behind every database we use and every package we install, there is a person or a team investing so much time and effort for the benefit of the entire research community.