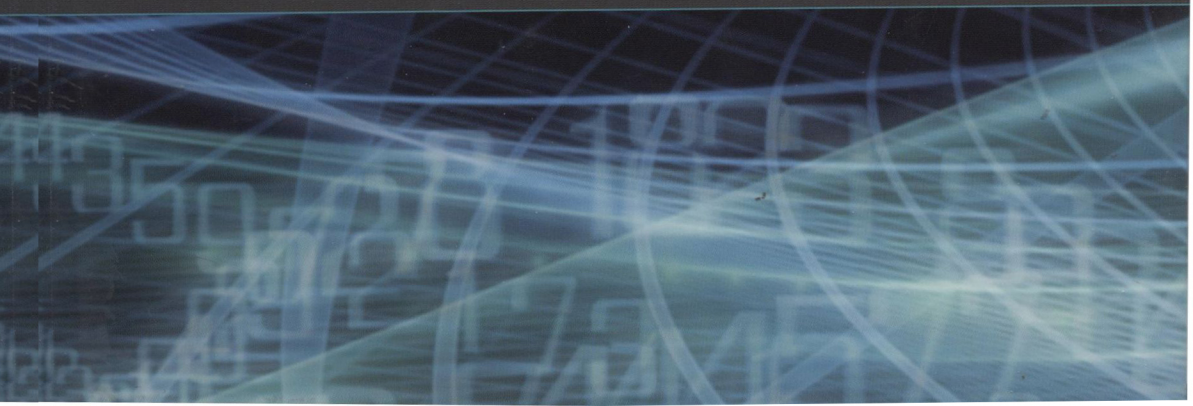
BIG DATA, LITTLE DATA, NO DATA

SCHOLARSHIP IN THE NETWORKED WORLD

Christine L. Borgman



Big Data, Little Data, No Data

Scholarship in the Networked World

Christine L. Borgman

The MIT Press
Cambridge, Massachusetts
London, England

This book was set in Stone Sans and Stone Serif by the MIT Press. Printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data

Borgman, Christine L., 1951–
Big data, little data, no data : scholarship in the networked world / Christine L. Borgman.
pages cm
Includes bibliographical references and index.
ISBN 978-0-262-02856-1 (hardcover : alk. paper)
1. Communication in learning and scholarship—Technological innovations. 2. Research—Methodology. 3. Research—Data processing. 4. Information technology. 5. Information storage and retrieval systems. 6. Cyberinfrastructure. I. Title.
AZ195.B66 2015
004—dc23

2014017233

10 9 8 7 6 5 4 3 2 1

Big Data, Little Data, No Data

For Betty Champoux Borgman, 1926–2012,
and Ann O'Brien, 1951–2014

Preface

Big data begets big attention these days, but little data are equally essential to scholarly inquiry. As the absolute volume of data increases, the ability to inspect individual observations decreases. The observer must step ever further away from the phenomena of interest. New tools and new perspectives are required. However, big data is not necessarily better data. The farther the observer is from the point of origin, the more difficult it can be to determine what those observations mean—how they were collected; how they were handled, reduced, and transformed; and with what assumptions and what purposes in mind. Scholars often prefer smaller amounts of data that they can inspect closely. When data are undiscovered or undiscoverable, scholars may have no data.

Research data are much more—and less—than commodities to be exploited. Data management plans, data release requirements, and other well-intentioned policies of funding agencies, journals, and research institutions rarely accommodate the diversity of data or practices across domains. Few policies attempt to define *data* other than by listing examples of what they might be. Even fewer policies reflect the competing incentives and motivations of the many stakeholders involved in scholarship. Data can be many things to many people, all at the same time. They can be assets to be controlled, accumulated, bartered, combined, mined, and perhaps to be released. They can be liabilities to be managed, protected, or destroyed. They can be sensitive or confidential, carrying high risks if released. Their value may be immediately apparent or not realized until a time much later. Some are worth the investment to curate indefinitely, but many have only transient value. Within hours or months, advances in technology and research fronts have erased the value in some kinds of observations.

A starting point to understand the roles of data in scholarship is to acknowledge that data rarely are *things* at all. They are not natural objects with an essence of their own. Rather, data are representations of

observations, objects, or other entities used as evidence of phenomena for the purposes of research or scholarship. Those representations vary by scholar, circumstance, and over time. Across the sciences, social sciences, and the humanities, scholars create, use, analyze, and interpret data, often without agreeing on what those data are. Conceptualizing something as data is itself a scholarly act. Scholarship is about evidence, interpretation, and argument. Data are a means to an end, which is usually the journal article, book, conference paper, or other product worthy of scholarly recognition. Rarely is research done with data reuse in mind.

Galileo sketched in his notebook. Nineteenth-century astronomers took images on glass plates. Today's astronomers use digital devices to capture photons. Images of the night sky taken with consumer-grade cameras can be reconciled to those taken by space missions because astronomers have agreed on representations for data description and mapping. Astronomy has invested heavily in standards, tools, and archives so that observations collected over the course of several centuries can be aggregated. However, the knowledge infrastructure of astronomy is far from complete and far from fully automated. Information professionals play key roles in organizing and coordinating access to data, astronomical and otherwise.

Relationships between publications and data are manifold, which is why research data is fruitfully examined within the framework of scholarly communication. The making of data may be deliberate and long term, accumulating a trove of resources whose value increases over time. It may be ad hoc and serendipitous, grabbing whatever indicators of phenomena are available at the time of occurrence. No matter how well defined the research protocol, whether for astronomy, sociology, or ethnography, the collection of data may be stochastic, with findings in each stage influencing choices of data for the next. Part of becoming a scholar in any field is learning how to evaluate data, make decisions about reliability and validity, and adapt to conditions of the laboratory, field site, or archive. Publications that report findings set them in the context of the domain, grounding them in the expertise of the audience. Information necessary to understand the argument, methods, and conclusions are presented. Details necessary to replicate the study are often omitted because the audience is assumed to be familiar with the methods of the field. Replication and reproducibility, although a common argument for releasing data, are relevant only in selected fields and difficult to accomplish even in those. Determining which scholarly products are worth preserving is the harder problem.

Policies for data management, release, and sharing obscure the complex roles of data in scholarship and largely ignore the diversity of practices

within and between domains. Concepts of data vary widely across the sciences, social sciences, and humanities, and within each area. In most fields, data management is learned rather than taught, leading to ad hoc solutions. Researchers often have great difficulty reusing their own data. Making those data useful to unknown others, for unanticipated purposes, is even harder. Data sharing is the norm in only a few fields because it is very hard to do, incentives are minimal, and extensive investments in knowledge infrastructures are required.

This book is intended for the broad audience of stakeholders in research data, including scholars, researchers, university leaders, funding agencies, publishers, libraries, data archives, and policy makers. The first section frames data and scholarship in four chapters, provoking a discussion about concepts of data, scholarship, knowledge infrastructures, and the diversity of research practices. The second section consists of three chapters exploring data scholarship in the sciences, social sciences, and humanities. These case studies are parallel in structure, providing comparisons across domains. The concluding section spans data policy and practice in three chapters, exploring why data scholarship presents so many difficult problems. These include releasing, sharing, and reusing data; credit, attribution, and discovery; and what to keep and why.

Scholarship and data have long and deeply intertwined histories. Neither are new concepts. What is new are efforts to extract data from scholarly processes and to exploit them for other purposes. Costs, benefits, risks, and rewards associated with the use of research data are being redistributed among competing stakeholders. The goal of this book is to provoke a much fuller, and more fully informed, discussion among those parties. At stake is the future of scholarship.

Christine L. Borgman
Los Angeles, California
May 2014

Acknowledgments

It takes a village to write a sole-authored book, especially one that spans as many topics and disciplines as does this one. My writing draws upon the work of a large and widely distributed village of colleagues—an "invisible college" in the language of scholarly communication. Scholars care passionately about their data and have given generously of their time in countless discussions, participation in seminars and workshops, and reading many drafts of chapters.

The genesis of this book project goes back too many years to list all who have influenced my thinking, thus these acknowledgments can thank, at best, those who have touched the words in this volume in some way. Many more are identified in the extensive bibliography. No doubt I have failed to mention more than a few of you with whom I have had memorable conversations about the topics therein.

My research on scholarly data practices dates to the latter 1990s, building on prior work on digital libraries, information-seeking behavior, human-computer interaction, information retrieval, bibliometrics, and scholarly communication. The data practices research has been conducted with a fabulous array of partners whose generative contributions to my thinking incorporate too much tacit knowledge to be made explicit here. Our joint work is cited throughout. Many of the faculty collaborators, students, and postdoctoral fellows participated in multiple projects; thus, they are combined into one alphabetical list. Research projects on scholarly data practices include the Alexandria Digital Earth Prototype Project (ADEPT); Center for Embedded Networked Sensing (CENS); Cyberlearning Task Force; Monitoring, Modeling, and Memory; Data Conservancy; Knowledge Infrastructures; and Long-Tail Research.

Faculty collaborators on these projects include Daniel Atkins, Geoffrey Bowker, Sayeed Choudhury, Paul Davis, Tim DiLauro, George Djorgovski, Paul Edwards, Noel Enyedy, Deborah Estrin, Thomas Finholt, Ian Foster,

James Frew, Jonathan Furner, Anne Gilliland, Michael Goodchild, Alyssa Goodman, Mark Hansen, Thomas Harmon, Bryan Heidorn, William Howe, Steven Jackson, Carl Kesselman, Carl Lagoze, Gregory Leazer, Mary Marlino, Richard Mayer, Carole Palmer, Roy Pea, Gregory Pottie, Allen Renear, David Ribes, William Sandoval, Terence Smith, Susan Leigh Star, Alex Szalay, Charles Taylor, and Sharon Traweek. Students, postdoctoral fellows, and research staff collaborators on these projects include Rebekah Cummings, Peter Darch, David Fearon, Rich Gazan, Milena Golshan, Eric Graham, David Gwynn, Greg Janee, Elaine Levia, Rachel Mandell, Matthew Mayernik, Stasa Milojevic, Alberto Pepe, Elizabeth Rolando, Ashley Sands, Katie Shilton, Jillian Wallis, and Laura Wynholds.

Most of this book was developed and written during my 2012–2013 sabbatical year at the University of Oxford. My Oxford colleagues were fountains of knowledge and new ideas, gamely responding to my queries of "what are your data?" Balliol College generously hosted me as the Oliver Smithies Visiting Fellow and Lecturer, and I concurrently held visiting scholar posts at the Oxford Internet Institute and the Oxford eResearch Centre. Conversations at high table and low led to insights that pervade my thinking about all things data—Buddhism, cosmology, Dante, genomics, chirality, nanotechnology, education, economics, classics, philosophy, mathematics, medicine, languages and literature, computation, and much more. The Oxford college system gathers people together around a table who otherwise might never meet, much less engage in boundary-spanning inquiry. I am forever grateful to my hosts, Sir Drummond Bone, Master of Balliol, and Nicola Trott, Senior Tutor; William Dutton of the Oxford Internet Institute; David de Roure, Oxford eResearch Centre; and Sarah Thomas, Bodley's Librarian. My inspiring constant companions at Oxford included Kofi Agawu, Martin Burton, George and Carmella Edwards, Panagis Filippakopoulos, Marina Jirotka, Will Jones, Elena Lombardi, Eric Meyer, Concepcion Naval, Peter and Shirley Northover, Ralph Schroeder, Anne Trefethen, and Stefano Zacchetti.

Others at Oxford who enlightened my thinking, perhaps more than they know, include William Barford, Grant Blank, Dame Lynne Brindley, Roger Cashmore, Sir Iain Chalmers, Carol Clark, Douglas Dupree, Timothy Endicott, David Erdos, Bertrand Faucheux, James Forder, Brian Foster, John-Paul Ghobrial, Sir Anthony Graham, Leslie Green, Daniel Grimley, Keith Hannabus, Christopher Hinchcliffe, Wolfram Horstmann, Sunghee Kim, Donna Kurtz, Will Lanier, Chris Lintott, Paul Luff, Bryan Magee, Helen Margetts, Philip Marshall, Ashley Nord, Dominic O'Brien, Dermot O'Hare, Richard Ovenden, Denis Noble, Seamus Perry, Andrew Pontzen, Rachel

Quarrell, David Robey, Anna Sander, Brooke Simmons, Rob Simpson, Jin-Chong Tan, Linnet Taylor, Rosalind Thomas, Nick Trefethen, David Vines, Lisa Walker, David Wallace, Jamie Warner, Frederick Wilmot-Smith, and Timothy Wilson.

Very special acknowledgments are due to my colleagues who contributed substantially to the case studies in chapters 5, 6, and 7. The astronomy case in chapter 5 relies heavily on the contributions of Alyssa Goodman of the Harvard-Smithsonian Center for Astrophysics and her collaborators, including Alberto Accomazzi, Merce Crosas, Chris Erdmann, Michael Kurtz, Gus Muench, and Alberto Pepe. It also draws on the research of the Knowledge Infrastructures research team at UCLA. The case benefited from multiple readings of drafts by professor Goodman and reviews by other astronomers or historians of astronomy, including Alberto Accomazzi, Chris Lintott, Michael Kurtz, Patrick McCray, and Brooke Simmons. Astronomers George Djorgovski, Phil Marshall, Andrew Pontzen, and Alex Szalay also helped clarify scientific issues. The sensor-networked science and technology case in chapter 5 draws on prior published work about CENS. Drafts were reviewed by collaborators and by CENS science and technology researchers, including David Caron, Eric Graham, Thomas Harmon, Matthew Mayernik, and Jillian Wallis. The first social sciences case in chapter 6, on Internet research, is based on interviews with Oxford Internet Institute researchers Grant Blank, Corinna di Gennaro, William Dutton, Eric Meyer, and Ralph Schroeder, all of whom kindly reviewed drafts of the chapter. The second case, on sociotechnical studies, is based on prior published work with collaborators, as cited, and was reviewed by collaborators Matthew Mayernik and Jillian Wallis. The humanities case studies in chapter 7 were developed for this book. The CLAROS case is based on interviews and materials from Donna Kurtz of the University of Oxford, with further contributions from David Robey and David Shotton. The analysis of the Pisa Griffin draws on interviews and materials from Peter Northover, also of Oxford, and additional sources from Anna Contadini of SOAS, London. The closing case, on Buddhist scholarship, owes everything to the patient tutorial of Stefano Zacchetti, Yehan Numata Professor of Buddhist Studies at Oxford, who brought me into his sanctum of enlightenment. Humanities scholars were generous in reviewing chapter 7, including Anna Contadini, Johanna Drucker, Donna Kurtz, Peter Northover, Todd Presner, Joyce Ray, and David Robey.

Many others shared their deep expertise on specialized topics. On biomedical matters, these included Jonathan Bard, Martin Burton, Iain Chalmers, Panagis Filippakopoulos, and Arthur Thomas. Dr. Filippakopoulos

read drafts of several chapters. On Internet technologies and citation mechanisms, these included Geoffrey Bilder, Blaise Cronin, David de Roure, Peter Fox, Carole Goble, Peter Ingwersen, John Klensin, Carl Lagoze, Salvatore Mele, Ed Pentz, Herbert van de Sompel, and Yorick Wilks. Chapter 9 was improved by the comments of Blaise Cronin, Kathleen Fitzpatrick, and John Klensin. Paul Edwards and Marilyn Raphael were my consultants on climate modeling. Sections on intellectual property and open access benefited from discussions with David Erdos, Leslie Green, Peter Hirtle, Peter Murray-Rust, Pamela Samuelson, Victoria Stodden, and John Wilbanks. Christopher Kelty helped to clarify my understanding of common-pool resources, building on other discussions of economics with Paul David, James Forder, and David Vines. Ideas about knowledge infrastructures were shaped by long-running discussions with my collaborators Geoffrey Bowker, Paul Edwards, Thomas Finholt, Steven Jackson, Cory Knobel, and David Ribes. Similarly, ideas about data policy were shaped by membership on the Board on Research Data and Information, on CODATA, on the Electronic Privacy Information Center, and by the insights of Francine Berman, Clifford Lynch, Paul Uhlir, and Marc Rotenberg. On issues of libraries and archives, I consulted Lynne Brindley, Johanna Drucker, Anne Gilliland, Margaret Hedstrom, Ann O'Brien, Susan Parker, Gary Strong, and Sarah Thomas. Jonathan Furner clarified philosophical concepts, building upon what I learned from many Oxford conversations. Will Jones introduced me to the ethical complexities of research on refugees. Abdelmonem Afifi, Mark Hansen, and Xiao-li Meng improved my understanding of the statistical risks in data analysis. Clifford Lynch, Lynne Markus, Matthew Mayernik, Ann O'Brien, Katie Shilton, and Jillian Wallis read and commented upon large portions of the manuscript, as did several helpful anonymous reviewers commissioned by Margy Avery of the MIT Press.

I would be remiss not to acknowledge the invisible work of those who rarely receive credit in the form of authorship. These include the funding agencies and program officers who made this work possible. At the National Science Foundation, Daniel Atkins, Stephen Griffin, and Mimi McClure have especially nurtured research on data, scholarship, and infrastructure. Tony Hey and his team at Microsoft Research collaborated, consulted, and gave monetary gifts at critical junctures. Thanks to Lee Dirks, Susan Dumais, Catherine Marshall, Catherine van Ingen, Alex Wade, and Curtis Wong of MSR. Josh Greenberg at the Sloan Foundation has given us funds, freedom, and guidance in studying knowledge infrastructures. Also invisible are the many people who invited me to give talks from the book-in-progress and those who attended. I am grateful for those rich opportunities

for discussion. Rebekah Cummings, Elaine Levia, and Camille Mathieu curated the massive bibliography, which will be made public as a Zotero group (Borgman Big Data, Little Data, No Data) when this book is published, in the spirit of open access.

Last, but by no means least, credit is due to my husband, George Mood, who has copyedited this manuscript and everything else I have published since 1977. He usually edits his name out of acknowledgments sections, however. Let the invisible work be made visible this time.