

Big Data, Little Data, No Data

Scholarship in the Networked World

Christine L. Borgman

**The MIT Press
Cambridge, Massachusetts
London, England**

© 2015 Christine L. Borgman

All rights reserved. No part of this book may be reproduced in any form by any electronic or mechanical means (including photocopying, recording, or information storage and retrieval) without permission in writing from the publisher.

MIT Press books may be purchased at special quantity discounts for business or sales promotional use. For information, please email special_sales@mitpress.mit.edu.

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

Detailed Table of Contents

Preface xvii

Acknowledgments xxi

Part I: Data and Scholarship 1

1 Provocations 3

Introduction 3

Big Data, Little Data 4

Bigness 5

Openness 7

The Long Tail 8

No Data 10

Data Are Not Available 11

Data Are Not Released 11

Data Are Not Usable 13

Provocations 13

Conclusion 15

2 What Are Data? 17

Introduction 17

Definitions and Terminology 18

Definitions by Example 19

Operational Definitions 20

Categorical Definitions 21

Degrees of Processing 21

Origin and Preservation Value 23

Collections 25

Conceptual Distinctions 26

Sciences and Social Sciences 26

Humanities 27

Conclusion 28

3 Data Scholarship 31

Introduction 31

Knowledge Infrastructures 32

The Social and the Technical 35

Communities and Collaboration 36

Knowledge and Representation 37

Theory, Practice, and Policy 38

Open Scholarship 39

Open Access to Research Findings 39

Open Access to Data 42

Open Technologies 45

Converging Communication 47

Data Metaphors 47

Units of Data 50

Documents of Record 51

Conclusion 52

4 Data Diversity 55

Introduction 55

Disciplines and Data 56

Size Matters 58

Project Goals 58

Data Collection 60

Data Analysis 61

When Are Data? 62

Distance Matters 64

Sources and Resources 64

Metadata 65

Definitions and Discovery 66

Communities and Standards 68

Provenance 70

External Influences 71

Economics and Values 71

Property Rights 75

Ethics 77

Conclusion 79

Part II: Case Studies in Data Scholarship 81

5 Data Scholarship in the Sciences 83

Introduction 83

Research Methods and Data Practices 83

Science Cases 84

Astronomy 85

Size Matters 86

Big Science, Little Science 86

Big Data, Long Tail 87

When Are Data? 90

Sources and Resources 91

Telescopes 91

Electromagnetic Spectrum 92

Celestial Objects 93

Astronomy Data Products 93

Knowledge Infrastructures 94

Metadata 94

Coordinate Systems 95

Celestial Objects 96

Data Archiving 97

Publications 98

Provenance 99

External Influences 100

Economics and Value 100

Property Rights 100

Ethics 101

Conducting Research in Astronomy 102

The COMPLETE Survey 102

Research Questions 103

Collecting Data 103

Analyzing Data 104

Publishing Findings 104

Curating, Sharing, and Reusing Data 105

Sensor-Networked Science and Technology 106

Size Matters 106

When Are Data? 108

Sources and Resources 109

Embedded Sensor Networks 109

Physical Samples 111

Software, Code, Scripts, and Models 111

Background Data 111

Knowledge Infrastructures	112
Metadata	112
Provenance	113
External Influences	113
Economics and Value	113
Property Rights	114
Ethics	115
Conducting Research with Embedded Sensor Networks	116
Research Questions	117
Collecting Data	117
Analyzing Data	119
Publishing Findings	119
Curating, Sharing, and Reusing Data	120
Conclusion	121
6 Data Scholarship in the Social Sciences	125
Introduction	125
Research Methods and Data Practices	126
Social Sciences Cases	127
Internet Surveys and Social Media Studies	128
Size Matters	128
When Are Data?	129
Sources and Resources	129
Knowledge Infrastructures	131
Metadata	132
Provenance	133
External Influences	135
Economics and Value	135
Property Rights	136
Ethics	136
Conducting Internet Surveys and Social Media Research	137
Research Questions	138
Collecting Data	139
Analyzing Data	140
Publishing Findings	141
Curating, Sharing, and Reusing Data	142
Sociotechnical Studies	143
Size Matters	144
When Are Data?	144
Sources and Resources	145
Field Observations and Ethnography	145
Interviews	146

Records and Documents	146
Building and Evaluating Technologies	147
Knowledge Infrastructures	147
Metadata	148
Provenance	148
External Influences	149
Economics and Value	149
Property Rights	149
Ethics	150
Conducting Sociotechnical Research in CENS	150
Research Questions	151
Collecting Data	152
Analyzing Data	154
Publishing Findings	155
Curating, Sharing, and Reusing Data	156
Conclusion	157
7 Data Scholarship in the Humanities	161
Introduction	161
Research Methods and Data Practices	162
Humanities Cases	164
Classical Art and Archaeology	164
Size Matters	165
When Are Data?	166
Sources and Resources	166
Physical versus Digital Objects	167
Digital versus Digitized	167
Surrogates versus Full Content	167
Static Images versus Searchable Representations	168
Searchable Strings versus Enhanced Content	169
Knowledge Infrastructures	170
Metadata	171
Provenance	172
Collections	173
External Factors	176
Economics and Value	176
Property Rights	178
Ethics	178
Conducting Research in Classical Art and Archaeology	179
Research Questions	180
Collecting Data	181
Analyzing Data	182

Publishing Findings	184
Curating, Sharing, and Reusing Data	184
Buddhist Studies	186
Size Matters	187
When Are Data?	187
Sources and Resources	188
Primary versus Secondary Sources	188
Static Images versus Enhanced Content	189
Knowledge Infrastructures	189
Metadata	190
Provenance	191
Collections	191
External Factors	192
Economics and Value	192
Property Rights	193
Ethics	193
Conducting Research in Buddhist Studies	194
Research Questions	195
Collecting Data	196
Analyzing Data	196
Publishing Findings	197
Curating, Sharing, and Reusing Data	199
Conclusion	200
 Part III: Data Policy and Practice	 203
 8 Sharing, Releasing, and Reusing Data	 205
Introduction	205
Supply and Demand for Research Data	207
The Supply of Research Data	208
To Reproduce Research	209
Defining Reproducibility	209
Determining What to Reproduce	209
Detecting Fraud	210
Resolving Disputes	211
To Make Public Assets Available to the Public	211
To Leverage Investments in Research	212
To Advance Research and Innovation	212
The Demand for Research Data	213
Scholarly Motivations	214
Publications and Data	215

Communicating Research	215
Publishing Research	216
Data as Assets and Liabilities	217
Releasing Data	218
Representation and Mobility	219
Provenance	220
Acquiring Data to Reuse	222
Background and Foreground Uses	222
Interpretation and Trust	223
Knowledge Infrastructures	224
Repositories, Collections, and Archives	225
Private Practice	227
Human Infrastructure	228
Intractable Problems	229
Disciplinary Knowledge Infrastructures	229
Sciences	230
Astronomy	230
Sensor Networked Science and Technology	231
Genomics	233
Social Sciences	235
Internet Research	235
Sociotechnical Research	235
Humanities	236
Classical Art and Archaeology	236
Buddhist Studies	237
Conclusion	237
9 Credit, Attribution, and Discovery of Data	241
Introduction	241
Principles and Problems	243
Theory and Practice	245
Substance and Style: How to Cite	245
Theories of Citation Behavior: What, When, and Why to Cite Objects	248
Meaning of Links	248
Selecting References	249
Theorizing and Modeling Citation Behavior	250
Citing Data	251
Clear or Contested: Who Is Credited and Attributed?	252
Naming the Cited Author	252
Negotiating Authorship Credit	253
Responsibility	255
Credit for Data	256

Name or Number: Questions of Identity	258
Identifying People and Organizations	258
Identity and Discovery	260
Identifying Objects	261
Theory Meets Technology: Citations as Actions	264
Risks and Rewards: Citations as Currency	266
Conclusion	268
10 What to Keep and Why	271
Introduction	271
Provocations Revisited	273
Rights, Responsibilities, Roles, and Risks	273
Data Sharing	275
Publications and Data	278
Data Access	281
Stakeholders and Skills	283
Knowledge Infrastructures Past, Present, and Future	285
Conclusion	287
References	289
Index	361