

This is a section of [doi:10.7551/mitpress/12200.001.0001](https://doi.org/10.7551/mitpress/12200.001.0001)

# The Open Handbook of Linguistic Data Management

**Edited By:** Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller, Lauren B. Collister

## Citation:

*The Open Handbook of Linguistic Data Management*

**Edited By:** Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller, Lauren B. Collister

**DOI:** 10.7551/mitpress/12200.001.0001

**ISBN (electronic):** 9780262366076

**Publisher:** The MIT Press

**Published:** 2022



The MIT Press

# The Open Handbook of Linguistic Data Management

**Open Handbook in Linguistics Series**

Series Editor: Heidi B. Harley

*The Open Handbook of Linguistic Data Management*, Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller,  
and Lauren B. Collister

# **The Open Handbook of Linguistic Data Management**

**Edited by Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller, and Lauren B. Collister**  
**Foreword by Sarah G. Thomason**

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

© 2021 The Massachusetts Institute of Technology

This work is subject to a Creative Commons CC-BY-NC license. Subject to such license, all rights are reserved.



This book was set in Stone Serif and Stone Sans by Westchester Publishing Services.

Library of Congress Cataloging-in-Publication Data

Names: Berez-Kroeker, Andrea L., editor. | McDonnell, Bradley James, editor. | Koller, Eve, editor. | Collister, Lauren B., editor.

Title: The open handbook of linguistic data management / edited by Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller and Lauren B. Collister.

Description: Cambridge, Massachusetts : The MIT Press, [2021] | Series: Open handbooks in linguistics series | Includes bibliographical references and index.

Identifiers: LCCN 2020044363 | ISBN 9780262045261 (hardcover)

Subjects: LCSH: Computational linguistics. | Natural language processing (Computer science) | Data mining.

Classification: LCC P98 .O64 2021 | DDC 410.285—dc23

LC record available at <https://lcn.loc.gov/2020044363>

To our parents  
Ellen and Bill  
Bob and Kris  
Nancy and Sanford  
Greta and Tom  
Jon and Chris

and to linguists, present and future.



## Contents

Series Foreword xi

Foreword by Sarah G. Thomason xiii

### I Conceptual Foundations, Principles, and Implementation of Data Management in Linguistics

- 1 Data, Data Management, and Reproducible Research in Linguistics: On the Need for *The Open Handbook of Linguistic Data Management*** 3  
Andrea L. Berez-Kroeker, Bradley McDonnell, Lauren B. Collister, and Eve Koller
- 2 Situating Linguistics in the Social Science Data Movement** 9  
Lauren Gawne and Suzy Styles
- 3 The Scope of Linguistic Data** 27  
Jeff Good
- 4 Indigenous Peoples, Ethics, and Linguistic Data** 49  
Gary Holton, Wesley Y. Leonard, and Peter L. Pulsifer
- 5 The Linguistic Data Life Cycle, Sustainability of Data, and Principles of Solid Data Management** 61  
Eleanor Mattern
- 6 Transforming Data** 73  
Na-Rae Han
- 7 Archiving Research Data** 89  
Helene N. Andreassen
- 8 Developing a Data Management Plan** 101  
Susan Smythe Kung
- 9 Copyright and Sharing Linguistic Data** 117  
Lauren B. Collister
- 10 Linguistic Data in the Long View** 129  
Laura Buszard-Welcher
- 11 Guidance for Citing Linguistic Data** 143  
Philipp Conzett and Koenraad De Smedt

### 12 Metrics for Evaluating the Impact of Data Sets

 157

Robin Champieux and Heather L. Coates

### 13 The Value of Data and Other Non-traditional Scholarly Outputs in Academic Review, Promotion, and Tenure in Canada and the United States

 171

Juan Pablo Alperin, Lesley A. Schimanski, Michelle La, Meredith T. Niles, and Erin C. McKiernan

### II Data Management Use Cases

#### 14 Managing Sociolinguistic Data with the Corpus of Regional African American Language (CORAAAL)

 185

Tyler Kendall and Charlie Farrington

#### 15 Managing Data for Integrated Speech Corpus Analysis in *SPeech Across Dialects of English (SPADE)*

 195

Morgan Sonderegger, Jane Stuart-Smith, Michael McAuliffe, Rachel Macdonald, and Tyler Kendall

#### 16 Data Management at the uOttawa Sociolinguistics Laboratory

 209

Shana Poplack

#### 17 Managing Legacy Data in a Sociophonetic Study of Vowel Variation and Change

 221

James Grama

#### 18 Managing Sociophonetic Data in a Study of Regional Variation

 237

Valerie Fridland and Tyler Kendall

#### 19 Data Management Practices in an Ethnographic Study of Language and Migration

 249

Lynnette Arnold

#### 20 Managing Conversation Analysis Data

 257

Elliott M. Hoey and Chase Wesley Raymond



- 21 Managing Sign Language Data from Fieldwork** 267  
Nick Palfreyman
- 22 Managing Data in a Language Documentation Corpus** 277  
Christopher Cox
- 23 Managing Data for Writing a Reference Grammar** 287  
Nala H. Lee
- 24 Managing Lexicography Data: A Practical, Principled Approach Using FLEx (FieldWorks Language Explorer)** 301  
Christine Beier and Lev Michael
- 25 Managing Data from Archival Documentation for Language Reclamation** 315  
Megan Lukaniec
- 26 Managing Data for Descriptive and Historical Research** 327  
Don Daniels and Kelsey Daniels
- 27 Managing Historical Data in the Chirila Database** 335  
Claire Bower
- 28 Managing Historical Linguistic Data for Computational Phylogenetics and Computer-Assisted Language Comparison** 345  
Tiago Tresoldi, Christoph Rzymiski, Robert Forkel, Simon J. Greenhill, Johann-Mattis List, and Russell D. Gray
- 29 Managing Computational Data for Models of Language Acquisition and Change** 355  
Matthew Lou-Magnuson and Luca Onnis
- 30 Managing Sign Language Acquisition Video Data: A Personal Journey in the Organization and Representation of Signed Data** 367  
Julie A. Hochgesang
- 31 Managing Acquisition Data for Developing Large Sesotho, English, and French Corpora for CHILDES** 385  
Katherine Demuth
- 32 Managing Phonological Development Data within PhonBank: The Chisasibi Child Language Acquisition Study** 391  
Yvan Rose and Julie Brittain
- 33 Managing Oral and Written Data from an ESL Corpus from Canadian Secondary School Students in a Compulsory, School-Based ESL Program** 401  
Philippa Bell, Laura Collins, and Emma Marsden
- 34 Managing Second Language Acquisition Data with Natural Language Processing Tools** 411  
Scott A. Crossley and Kristopher Kyle
- 35 Managing Data Workflows for Untrained Forced Alignment: Examples from Costa Rica, Mexico, the Cook Islands, and Vanuatu** 423  
Rolando Coto-Solano, Sally Akevai Nicholas, Brittany Hoback, and Gregorio Tiburcio Cano
- 36 Managing Transcription Data for Automatic Speech Recognition with Elpis** 437  
Ben Foley, Daan van Esch, and Nay San
- 37 Managing Data and Statistical Code According to the FAIR Principles** 447  
Laura A. Janda
- 38 Managing Synchronic Corpus Data with the British National Corpus (BNC)** 453  
Stefan Th. Gries
- 39 Managing Data in Sign Language Corpora** 463  
Onno Crasborn
- 40 Managing Sign Language Video Data Collected from the Internet** 471  
Lynn Hou, Ryan Lepic, and Erin Wilkinson
- 41 Managing Data from Social Media: The Indigenous Tweets Project** 481  
Kevin P. Scannell
- 42 Managing Semantic Norms for Cognitive Linguistics, Corpus Linguistics, and Lexicon Studies** 489  
Bodo Winter
- 43 Managing Treebank Data with the Infrastructure for the Exploration of Syntax and Semantics (INESS)** 499  
Victoria Rosén and Koenraad De Smedt
- 44 Managing Data in a Formal Syntactic Study of an Underinvestigated Language (Uzbek)** 513  
Vera Gribanova

<b>45</b>	<b>Managing Data for Theoretical Syntactic Study of Underdocumented Languages</b>	<b>523</b>
	Philip T. Duncan, Harold Torrence, Travis Major, and Jason Kandybowicz	
<b>46</b>	<b>Managing Experimental Data in a Study of Syntax</b>	<b>531</b>
	Matthew Wagers	
<b>47</b>	<b>Managing Web Experiments for Psycholinguistics: An Example from Experimental Semantics/Pragmatics</b>	<b>539</b>
	Judith Degen and Judith Tonhauser	
<b>48</b>	<b>Managing, Sharing, and Reusing fMRI Data in Computational Neurolinguistics</b>	<b>547</b>
	Hiroyuki Akama	
<b>49</b>	<b>Managing Phonological Data in a Perception Experiment</b>	<b>557</b>
	Rory Turnbull	
<b>50</b>	<b>Managing Speech Perception Data Sets</b>	<b>565</b>
	Anne Cutler, Mirjam Ernestus, Natasha Warner, and Andrea Weber	
<b>51</b>	<b>Managing and Analyzing Data with Phonological CorpusTools</b>	<b>575</b>
	Kathleen Currie Hall, J. Scott Mackie, and Roger Yu-Hsiang Lo	
<b>52</b>	<b>Managing Phonological Inventory Data in the Development of PHOIBLE</b>	<b>589</b>
	Steven Moran	
<b>53</b>	<b>Managing Data in a Typological Study</b>	<b>597</b>
	Volker Gast and Łukasz Jędrzejowski	
<b>54</b>	<b>Managing Data for Descriptive Morphosemantics of Six Language Varieties</b>	<b>609</b>
	Malin Petzell and Caspar Jordan	
<b>55</b>	<b>Managing Data in TerraLing, a Large-Scale Cross-Linguistic Database of Morphological, Syntactic, and Semantic Patterns</b>	<b>617</b>
	Hilda Koopman and Cristina Guardiano	
<b>56</b>	<b>Managing AUTOTYP Data: Design Principles and Implementation</b>	<b>631</b>
	Alena Witzlack-Makarevich, Johanna Nichols, Kristine A. Hildebrandt, Taras Zakharko, and Balthasar Bickel	
	Contributors	643
	Index	651

