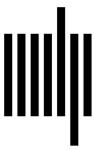
The Open Handbook of Linguistic Data Management

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

Foreword

Sarah G. Thomason

January 2020

This Handbook will be an invaluable addition to every linguist's toolbox. All linguists who deal with data (that is, all linguists, right?) need to know how to collect, transcribe, annotate, analyze, store, and share data, and we all need to pay close attention to ethical considerations that arise in most areas of data management. Those of us who do fieldwork must understand the importance of working with communities of language users, including issues of data ownership and communication. Those of us whose research connects with neighboring disciplines—psychology, anthropology, computer science, physics, and others—must ensure that our methods of data management are compatible with best practices in those disciplines as well as in linguistics. We all need to know about archiving our own data, and we all need to wrestle with the relatively new issue of citing data sets properly in our publications.

Some linguists are already managing their data in sophisticated ways, as the chapters in part II, "Data Management Use Cases," show. Others—and this is the category I belong to—know about a few of the topics covered here but are unfamiliar with most of them. As a fieldworker on a gravely endangered language, I know how to collect and transcribe data, and my field notes have a certain amount of metadata tagging. But my field notes, although digitized, don't even begin to approach the level of data management in any of the other ways discussed here. Phoneticians, computational linguists, and some fieldworkers, including authors in this Handbook, are way ahead of fieldworkers like me. But although some of us have more to learn than others, all linguists will learn vitally important things about data management from the chapters in this Handbook: about how to achieve the goals of FAIR, for instance, according to which research data should be findable, accessible, interoperable, and reusable; about different options for processing data, including organizing it for analysis; and about data management in single-language projects and data management in crosslinguistic studies.

I imagine the reason I was asked to write this foreword is that I wrestled with one aspect of this Handbook's topic back in 1994, when as the editor of *Language* I wrote an Editor's Department column that emphasized the need to ensure accuracy of the data in articles published in the journal. I consulted other journal editors and a few other colleagues who also felt strongly about accuracy before writing the column, and I found that they all shared my concerns—and that they all had horror stories similar to the ones I'd found in the course of my editing activities, among them misanalyzed data, data miscopied or even attributed to the wrong language, data extracted from a couple of examples and generalized to an entire language, and data drawn from secondary or tertiary sources without checking the original source. In those print-only days, it wasn't yet feasible for journals to publish supplementary materials online, so the author's examples comprised the only available data in a given article. Nowadays, with the possibility of publishing complete or at least extensive data sets online as supplementary materials, readers can check for consistency. With adequate metadata included, readers can also see how the data was collected and identify the original data sources, if relevant. Even this won't achieve the ideal, because (for instance) an author who was the only fieldworker for a particular language might have made mistakes in collecting the data. But it's a start.

Another issue, one that I didn't have to deal with as an editor but that I've certainly dealt with regularly as a teacher, is how to cite data sets. I've always given my students a template for citing my unpublished Montana Salish dictionary files, but readers of this Handbook will get a more informed view of best practices in the citation

of published and unpublished data sets. Yet another issue is how to teach administrators the value of both published and unpublished data sets so that they can be taken into account in hiring, tenure, and promotion cases. These and the many other topics covered in the Handbook will contribute much to the enlightenment of all linguists who do data-intensive research.

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://lccn.loc.gov/2020044363