Smith, Alexander D. 2022. A review of The Open Handbook of Linguistic Data Management. 2021. Edited by Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller, and Lauren B. Collister. Cambridge: MIT Press. ISBN 97802262045261.i–xiv. 671pp. *Glossa: a journal of general linguistics* 7(1). pp. 1–18. DOI: https://doi.org/10.16995/glossa.9261

Open Library of Humanities

A review of The Open Handbook of Linguistic Data Management. 2021. Edited by Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller, and Lauren B. Collister. Cambridge: MIT Press. ISBN 97802262045261.i–xiv. 671pp

Alexander D. Smith, National University of Singapore, SG, smith.alexander.david@gmail.com

As the field of linguistics becomes more dependent on data and access to data, linguists face the task of improving their data management practices to meet the demands of a growing open data movement. The Open Handbook of Linguistic Data Management (The Handbook) positions itself as a primary resource for linguists to improve data management. This article offers a review of The Handbook, which is praised for giving a foundational description of issues in data management facing our field as well as best practices for data management. Some criticisms of content choice and additional suggestions for data management improvements are also provided.

1 Introduction: Data management in linguistics and “The Handbook”

All linguists interact with data, regardless of sub-discipline, yet there are still sometimes major differences in data management practices both between sub-disciplines and between individual researchers within sub-disciplines. Historically, data management was something left to the individual, and in the pre-digital era this usually meant that relevant data sets were kept unpublished and in private collections. Dissemination of raw data might have occurred in publication, perhaps as an appendix to a research paper, but the independent distribution of data was uncommon and poor data management practices sometimes negatively affected the reliability of published materials (Thomason 1994). As technology has advanced, linguists have gained accesses to more data storage and dissemination resources, but actual data management practices have lagged behind (Berez-kroeker et al. 2018).

A lack of proper data management practices, data storage, and data availability can have a detrimental impact on the field. For example, while publishing about a lesser-studied language a researcher will make certain claims about the structure of that language. In response to such claims other researchers may want to analyze the datasets that informed the original analysis. Without proper data management and availability, however, researchers interested in asking questions will either have to solicit the data from the original author (who is not required to provide the data, who may provide only part of the story, or who may not be available), or find out for themselves by proposing, funding, and implementing their own investigation. In the end, if there is indeed some error in the original analysis, but the data set which informed that analysis is unavailable, the original and erroneous claims may be repeated for decades and may have a lasting impact on the field despite their invalidity.

Happily, recent trends in linguistic data management are improving how we record, store, report, distribute, and cite the data which drive our field. It is no longer acceptable in most publications to report on data without providing a certain minimum amount of detail on the data source. Gone are the days of what I call “take-my-word-for-it-linguistics” where claims are repeated without reasonable verification of those claims. It is no longer common practice to keep data in desk drawers, old notebooks, shoe boxes, personal computers, and the like (the “file-drawer problem”). Much more attention is given to the issue of data ownership and rights. Linguists are now expected to make their data sets widely available, and there are more resources for storage and dissemination than ever before.

As the field shifts towards more rigorous and open data management practices, linguists face new responsibilities and expectations when dealing with their own data and when dealing with data collected by others. For many linguists this change is welcome and the transition rather seamless. For others, there may be certain knowledge gaps which hold them back from fully embracing and implementing contemporary data management practices.

The need to improve one's own data management practices can be daunting, especially as it conflicts with other academic responsibilities. Without a definitive resource on data management, self-improvement in this area can be a long process. A handbook on data management for linguists is therefore needed, and *The Open Handbook of Linguistic Data Management* fills this gap with a valuable and densely packed volume dedicated to various issues in linguistic data management. It brings linguists up to speed with current practices and engages them in thoughtful discussion of problems in data management and possible solutions. It is honest about the state of the field and how it is changing. At 56 chapters, 109 contributors, and 671 pages, the volume is of a substantial size, but the chapters themselves are short enough to make for pleasant reading, and most who are interested in its contents will not be reading the entire volume front-to-back, but rather a selection of chapters here and there. Such is the nature of a handbook of this size and scope.

2 Reviewing The Handbook

This review article approaches The Handbook as a resource for linguists to improve their data management practices in the context of a field going through significant changes in its view of data management, storage, and accessibility. I judge it on its practicality, its timeliness, and its relevance to linguists working with data today, as well as its relevance for readers of Glossa. In addition to reviewing the volume, I offer personal insights and recommendations, especially for those who are just beginning the process of improving their data management practices.[1]

The Handbook is available for purchase as a hardback for US $250 as well as for free online at https://doi.org/10.7551/mitpress/12200.001.0001 with additional supplementary material in an online course at http://linguisticdatamanagement.org, discussed in more detail below.

The Handbook is divided into two parts. Part 1 is more concerned with the theories and best practices of data management as well as general recommendations for those working with data. It is highly valuable, especially for readers with less familiarity with data management and who may want to familiarize themselves with current recommendations and incorporate those practices into their own research. Part 2 contains a large number of case study chapters. The usefulness of these chapters will vary depending on how relevant they are for the readers particular needs.

[1] I am an Austronesianist, someone whose main focus is on the comparative study of Austronesian languages. My work is informed primarily through data sets that I compile from fieldwork. As an Austronesianist, I am particularly focused on historical and phonological study of Austronesian languages, language documentation, and language endangerment. I am familiar with the processes of gathering, organizing, and archiving large linguistic data sets and have several deposits available in the Kaipuleohone Archive (Kaipuleohone n.d.). In addition to these primary areas of research, I maintain an interest in theoretical approaches to the study of Austronesian languages. In writing this review, I try to both draw on my experience as a fieldworker as well as comment on the usefulness of the volume for those with more theoretical and experimental research programs.

Because of the size of the volume, this review must be restrictive on what chapters are covered. Part 1, *Conceptual foundations, principles, and implementation of data management in linguistics*, contains 13 chapters including the introduction. I provide a more complete review of this section and have comments on each chapter. Part 2, *Data management use cases*, includes chapters that are more restricted and specialized. In my review of Part 2 I tend to give less attention to individual chapters. I also briefly review the online-only content, followed by a review of the volume as a whole where I offer a more critical discussion. Finally, I end with a series of recommendations for those just beginning the process of improving their data management practices.

3 Part 1: Conceptual Foundations, Principles, and Implementation of Data Management in Lingusitics

I recommend Part 1 of The Handbook to any linguist with a desire to improve and update their data management practices. Some chapters are better prepared than others and some discuss topics that are more relevant than others, but in general Part 1 is comprehensive and covers foundational issues in data management of which we should all be aware. Part 1 covers topics that one would expect from such a volume, like archiving, data management plans, file storage and backups, ethics, open distribution, and the manipulation of data, but also some issues which one might not have expected, like copyright laws, extremely long time depths in data storage, and the academic review process.

The first chapter, *Data, data management, and reproducible research in linguistics: On the need for The Open Handbook of Linguistic Data Management*, describes the reasoning behind The Handbook and sets the stage for the remaining chapters. The chapter states that the major premise of The Handbook is that although all linguists work with data, many are untrained in data management. The remainder of The Handbook continues naturally from this premise, and is situated comfortably in an academic environment which is becoming more aware of the need for solid data management.

I have grouped the remaining twelve chapters of Part 1 into three rough "themes": The "data focused" chapters, the "ethics and law" chapters, and the "credit" chapters. The ordering of the chapters in the volume roughly follows these themes, although not exactly. The following discussion follows my theme grouping.

3.1 The "data focused" chapters

Chapters 2, *Situating linguistics in the social science data movement*, 3, *The scope of linguistic data*, 5, *The linguistic data life cycle, sustainability of data, and principles of solid data management*, 6, *Transforming data*, 7, *Archiving research data*, 8, *Developing a data management plan*, and 10, *Linguistic data in the long view* are the data focused chapters. They cover what linguistic data are, planning data management, and how data are created, manipulated, and stored. In many respects these are the most practical chapters; that is to say, they may directly influence and guide the reader's data management practices.

Chapter 2 situates linguistics within the social sciences, distinguishing it from "hard sciences". The labeling of linguistics as a social science enables a comparison between the state of data management between other social sciences and linguistics. The authors justify the labeling of linguistics as a social science by pointing out that linguistics deals with people and the data which we collect and analyze are, by the nature of our field, data originating from or engaged with human beings. I am happy with this justification. As a fieldworker, the data that I work with are the product of human interaction. They are gathered in certain contexts and certain relationships exist between field-worker and language-user. Even secondary or tertiary data that may be used by those not engaged in primary data gathering are the product of such human interactions.

In comparison to the data management practices of other social sciences, the authors show how linguistics lags behind, but that is not to say that linguistics is uniquely substandard. On the contrary, data management issues abound throughout the social sciences, for example, the "replication crisis" in social psychology (Gawne & Styles 2021: 14). The main difference between linguistics and other social sciences is timing; other social sciences began the process of improving their data management practices some years ago while linguistics is in many cases just beginning the process. Chapter 2 argues that linguistics stands to benefit from the lessons already learned in other social sciences.

The final point of chapter 2 is that changes must take place for linguistics to reap the full benefit of updated data management practices. Publishers tend to favor novel results, and there is not much room for replicative studies. Researchers themselves may feel unwilling to share data or make it fully open because they fear doing so may undermine their own research agenda. There are few immediate rewards for data openness since data sets are traditionally not credited as part of a researcher's scholarly contribution. Addressing these issues will take a cultural shift in the field. The chapter makes clear the dangers of the status quo and the need for data management changes. It sets the tone for the remainder of The Handbook, introduces the reader to the core issues of data management in linguistics, discusses where we lag behind and the cultural reckoning which needs to occur to change how we approach our data. It does less to provide clear solutions to these problems, and instead focuses more on making the reader aware of them.

Chapter 3, *The scope of linguistic data*, provides an overview of different types of linguistic data, how that data may be encoded, how it may be made available, and how new types of data are emerging in the field. Its main contribution is familiarizing the reader with these various types of data, which the author divides into data from observable linguistic behavior, analytical structures applied to data of language use, generalizations about languages and language (extracting typological generalizations from data), data on language users and situations, and linguistic metadata. Data from observable linguistic behavior is subdivided into documentation data, corpus data, and specialized data. The “specialized data” category includes grammaticality judgments, word judgments, and specialized recordings (such as recordings for phonetic analysis).

The initial separation of these types of data into “specialized” may come off as alienating. The author specialized in language documentation and seems to lump theoretically-inclined data into a separate category. The author’s arguments, however, is that these types of data should be treated like other types of data, as valuable resources that should be well managed, stored, and dispersed. I appreciate this point, and agree that types of linguistic data which are less often archived will benefit from better treatment. Data gathered in a highly controlled setting, or with specific and curated lists of sentences, or measurements taken in specific environments, lend themselves to replication. The nature of the interaction between researcher and language-user is highly relevant for correctly interpreting data and replicating experiments. Increased replicability should be a priority goal for linguists, and improved data management helps in that regard. Note, however, that although these points about better management of “specialized” data are legitimate, there is less discussion about how exactly to execute an improved data management plan for specialized research.

Chapter 5, *The linguistic data life cycle, sustainability of data, and principles of solid data management* was written by a librarian and thus gives the reader a view into a side of data management with which they may be unfamiliar. The chapter identifies the typical life cycle of data, 1) Planning research, 2) Collecting data, 3) Processing and analyzing data, 4) Publishing and sharing data, 5) Preserving data, and 6) Reusing data and uses this outline to inform a list of practical recommendations for data management which may be incorporated into current research. The practicality of this chapter makes it especially useful as an introduction, perhaps as part of the reading list for a course dealing with data. These practical elements include a discussion on file format types and use (choose open file formats like .txt or .csv, not proprietary formats like .docx or .xlsx), file naming (keep it simple, use the YYYY-MM-DD date format), data storage (3-2-1 rule), and metadata and documentation (keep proper records, use a standard metadata format, include a readme file).

The chapter is not ground-breaking in its recommendations, but anyone in the beginning stages of modernizing their data management practices will do well to reference this chapter as a guide to basic best practices.

Chapter 6 *Transforming data*, details the data transformation process; that is, taking raw data and altering it in some way, creating a new form of the same data. With chapter 6 the reader is introduced to the first case of overlap: Chapter 6, like chapter 5 before it, discusses open file formats in some detail. Chapter 5 focuses more on what file type is most suitable for what data-type, whereas chapter 6 is focused on more technical details of the file type itself, so there are subtle differences in the discussion, but the overlap is noticeable nevertheless and, since it immediately follows chapter 5, feels repetitive.

The more technical aspects of chapter 6 include discussion on encoding, direct data manipulation, merging and editing files, version control, and data corruption. Unfortunately, for readers whose main interaction with digital spaces is through the UI of programs like Word and who have little experience with command-line operations, some of the discussion may be beyond their current understanding. The authors provide relevant resources and recommendations for readers to familiarize themselves with these topics, but overall the chapter is best-suited for those with at least some preexisting familiarity with command-line operations and programming.

Chapter 7, *Archiving research data*, begins by discussing the advantages of archiving research data as well as the negative consequences of not archiving. For example, having access to an archive allows reviewers to give better feedback, and open data access can encourage replication studies. For the communities which provide the data, archives may serve as valuable tools for future language reclamation (Austin 2021). As attitudes towards data sets change, archived data may gain importance as a scholarly product. The chapter also discusses barriers to archiving research data, but these barriers overlap with points discussed at length in other chapters: lack of skills, lack of time, lack of reward, fear of having one's data open access, and unknowns regarding ethics and legality.

The core contribution of the chapter, however, is a practical how-to for archiving (sections 4 and 5), beginning with choice of repository, determining what data to archive, and dealing with personal information. The task of deciding on an archive is perhaps the most difficult for researchers who are not already familiar with relevant archives. Chapter 7 therefore includes information on archive-locating resources, like the Registry of Research Data Repositories (re3data n.d.) and the Open Language Archives Community (OLAC n.d.). Both resources help linguists filter reputable potential repositories by relevance to their own work and can be crucial in helping to decide where to store one's data.

Chapter 8, *Developing a data management plan*, is also a practical chapter. Data management plans (DMP) are increasingly necessary when applying for funding, but are also vital for any research plan, whether the data management plan is a requirement or not. The chapter provides the reader with extensive lists of the various elements of a data management plan. This includes creating the DMP itself, legalities and ethical issues, data storage, backup and security, documentation and metadata, dissemination, preservation, sharing, and creating a timeline.

A careful reader will be able to use the chapter directly as a step-by-step guide on the basics of creating their own DMP. For more junior researchers, it offers valuable guidance for creating one's first DMP. For more senior researchers, the chapter can be valuable for those who may have not been required to write DMPs but find themselves suddenly required to write one, or who simply want to improve their own practices. Chapter 8 is therefore one of the more immediately useful chapters in The Handbook.

Chapter 10, *Linguistic data in the long view*, discusses long-term data storage and its issues. My first impression of this chapter is that it is less practical than the others. It discusses data lifespans that extend to 10,000 years in the future. The chapter makes a good point about the safety of modern depositories, which are heavily reliant on consistent funding and management, and point out how a gap in funding could easily result in catastrophic data loss with little to no change of recovery from a modern digital archive. The main drawback of this chapter is that it discusses an issue which is beyond the scope of many linguists working with data archiving today. The topic of encoding data into DNA, for example, seems much less relevant for someone trying to organize and deposit language documentation data or someone trying to annotate an elicitation session. It is a chapter most relevant for archivists whose principle concern is data storage itself and not for those who are involved in providing that data. That is not to say that the topic of the chapter is not important. On the contrary, it is a very important topic. Do we expect our data to exist 10,000 years from now? If so, how will it be accessed and how will we ensure its survival? Given the practicality and here-and-now relevancy of the rest of the volume, and the target audience (linguists involved in linguistic research) it feels less relevant than other chapters.

The data focused chapters make for important reading, often dealing with basic and practical advice on data management for linguistic researchers. Readers who are already well-acquainted with such practices may still find some interesting points in these chapters, but there is clearly a bias towards researchers whose data management practices need updating. With the exception of chapter 10, which I feel is too far removed from the issues faced by linguists working with data today, the chapters in this section provide the reader with an excellent foundation.

3.2 The “ethics and law” chapters

Chapters 4, *Indigenous peoples, ethics, and linguistic data*, and 9, *Copyright and sharing linguistic data*, are the “ethics and law” chapters. These chapters are grouped together because they tackle a similar issue: the sharing of data and data ownership, from two very different perspectives. One is the perspective of indigenous people and the rights to their languages, data, and the ethics of sharing that data, and the other is the perspective of the (mostly western) legality of data sharing and copyright law.

Chapter 4 focuses on ethics and indigenous peoples’ rights. The chapter discusses important aspects of ethical research, such as utilizing empowering models of linguistic research, having greater participation and collaboration, and reducing the objectification of languages and their speakers.[2] Achieving a more inclusive ethics will involve not only changes to data use and management, but to the underlying power relationships that create those data. Indeed, the chapter centers not only the linguist's relationship to data but to the human beings who supply that data, and the inherently imbalanced nature of many of these relationships.

The rights of indigenous people and indigenous sovereignty over linguistic data are to some extent at odds with The Handbook's theme of open data. However, the apparent clash can be analyzed as a result of the prevailing mindset which has dominated linguistic field work, that linguistic data "belong" to the researcher. The chapter emphasizes that outsider linguists don't have innate rights to indigenous languages; rather, indigenous people have the right to share or not share their language with outsiders. The chapter does a good job of balancing the needs of linguistic research with the fundamental rights of indigenous communities.

Chapter 9 is essentially a review of copyright law, with a focus on linguistic data. It offers an overview on what copyright is, what copyright covers and where it applies, as well as exceptions to copyright. When comparing chapters 4 and 9, one can see the fundamental differences in their approach to data rights. Whereas ethical considerations of indigenous peoples' rights to data maintain a position that the ownership of data rests with the language users themselves, copyright law in many cases does not take this as a default position. Scholars must be aware of the shortcomings of copyright as it applies to indigenous peoples' languages, as well as when ethical considerations must be taken into account which go beyond copyright and legal considerations. It is, unfortunately, often up to the individual researcher to ensure ethical practice in this area, as the legal ownership of data does not always overlap with indigenous rights. Adhering to copyright and other western legal frameworks does not ensure that researcher interactions with indigenous people are ethical. The juxtaposition of these two chapters makes this clear.

3.3 The "credit" chapters

Chapters 11, *Guidance for citing linguistic data*, 12, *Metrics for evaluating the impact of data sets*, and 13, *The value of data and other non-traditional scholarly outputs in academic review, promotion, and tenure in Canada and the United States*, are the "credit" chapters. They ask questions like, "How are linguistic data and data sets cited?" and "How are they utilized (or underutilized) in academic review and evaluation?"

[2] Objectification of language refers to the common tendency that some linguists have to overvalue decontextualized language data and to under-report or ignore the context of language use. Objectification can also occur when languages and their users are reduced to typological characteristics, theoretical importance and interest, and speaker numbers. This tendency towards objectification can be difficult to overcome while exploring deep theoretical questions about Language. The chapter does not give direct guidance for theoreticians on this matter and in that respect the chapter is lacking. It sheds light on an issue but resolving the issue remains the task of individual linguists.

Chapter 11, like many of the preceding chapters, is highly practical. It offers a guide to researchers on how to properly cite data, especially data from archives and other repositories. The main point of the chapter is that the citation of data sets should not differ from citation of publications. Ideally, all data that appears in a publication should be properly cited such that interested researchers can easily access and evaluate the original data themselves. The chapter includes bibtex examples, and guidelines on how to incorporate data set citation into popular citation formats.

Chapters 12 and 13 deal with a different type of credit, but one which still intersects with citation and citation metrics, that is, evaluation metrics and data sets' role in academic review and promotion. Chapter 12 provides a historical analysis of citation metrics including the eye-opening fact that Impact Factors were developed primarily to "guide decision making for journal indexing and library collection development" (Champieux & Coates 2021: 158), and were not at all intended for use as a citation metric for individual scholars. Impact metrics effect all of us, from early career linguists to established scholars, administrators, and institutions, so understanding their intended use helps us understand how they are being misused.

With such a disconnect between the metrics which are used to measure our contribution and our actual contribution, something clearly should change. The chapter offers insightful guidance through types of metrics and how they interact with data, the usefulness of altmetrics, and a practical list of guidelines for scholars, evaluators, institutions, journals, and data repository managers to change their practices so that the metrics of scholarly contribution both better reflect the actual contribution of the author as well as incorporate data sets into those metrics.

Chapter 13 is unique among the Part 1 chapters. Most chapters can be classified as discussion pieces, reviews, statement pieces, and guidelines on best practices. Chapter 13, on the other hand, is a study on review, promotion, and tenure (RPT) practices in the United States and Canada, informed by an analysis of 864 RPT documents from 129 universities. The analysis targeted key words pertaining to the type of scholarly output (traditional, conventional, funding, arts, data, education, preprints, and more), and measured their mentions in internal RPT reports in each institution and discipline. The results of the analysis align with expectation; traditional categories far outweigh data mentions in RPT documents across the board. In doctorate-issuing institutions, for example, 93 percent of institutions mention traditional output but only 9 percent mention data.

The chapter provides a glimpse into the RPT process with which many junior scholars may not have experience. Disappointingly, the discussion is not as thorough as one might hope. Recommendations for what to do about a bias towards traditional outputs and against data sets are 1) cite data sets explicitly in materials presented for review and 2) follow good data management practices to facilitate inclusion of such materials. The repeated inclusion of such citations in review materials “…could, in time, lead to their inclusion among the explicitly recognized outputs” (Alperin et al. 2021: 179). I do not question the value of including such materials, but am less optimistic that inclusion will somehow passively lead to more explicit recognition. The recommendations place the burden of change on the shoulders of individual scholars, even though more direct advocacy for change from both scholars as well as sympathetic administrators will be necessary to change the over reliance on traditional output in academic review.

4 Part 2: Data Management Use Cases

The editors describe Part 2 as a snapshot of current practices in the form of data management use cases. This is the larger of the two parts, by a rather wide margin, but the reader will find that while the shorter Part 1 is densely packed with relevant discussion, Part 2’s utility will vary depending on the reader’s own research program. The same chapter may be inconsequential to one reader and indispensable to another.

Part 2 is surprisingly varied in its styling. Some chapters provide descriptions of the methods used in specific research projects. Other chapters are more like guides for best practices within certain research genres. Others are oriented towards specific software and feel like more of a how-to for linguists unfamiliar with the software. Within each chapter there are sometimes surprises, like chapter 49, *Managing phonological data in a perception experiment*, which mostly describes the methods and data management of a phonological perception experiment run by the author, but also includes a short guide to file management best practices (something that one might not expect from simply reading the chapter title).

The diversity of topics and focuses in Part 2 makes it somewhat hit-or-miss. Many readers will find something useful here, but it is not always obvious which chapters align most with one’s interests and the reader will have to do a bit of digging to zero in on the most relevant chapters. That said, the chapters are obviously organized such that similar sub-fields are grouped together. Chapters 14–20 ares sociolinguistic. Chapters 21–25 are more focused on documentation. Chapters 26–28 deal with historical and comparative linguistics, and chapter 28 specifically acts as a bridge chapter between the previous historical chapters and chapter 29, which is focused on computational linguistics. Chapters 30–34 deal with acquisition. Chapters 35–42 are a mix, but include forced alignment, speech recognition, and corpora. Chapters 43–46 deal with various approaches to syntactic data management, and chapter 46 bridges the gap between syntax and experimental chapters, which include chapters 47 and 48. Chapters 49–52 are on phonology. Chapters 53–56 are on typology.

Even though this sub-field structure exists, within each “section” we find chapters that are quite independent of one another. In the “sociolinguistic” chapters, for example, chapter 14 is a description of the creation of the Corpus of Regional African American Languages (CORAAL, Kendall & Farrington (2021)). It describes CORAAL's purpose, the processes involved in building CORAAL and within this section are helpful insights into data gathering, use of legacy materials, processing data and redaction of sensitive materials, metadata, and dispersal. Chapter 16 is a wide-ranging overview of the data management practices at the Sociolingusitics Lab at the University of Ottawa. It covers nearly all aspects of data management not for a single project, but as the guiding principles for an entire lab spanning over several decades of work. Chapter 17 describes the use of legacy materials in a longitudinal sociophonetic study. It includes the workflow of this specific study and covers topics like transcription and forced alignment (foreshadowing chapter 35 but not referencing it), data checking and cleaning, and data storage. Chapter 19 describes the data management practices in an ethically sensitive situation. It follows a similar descriptive course of other chapters, but diverges on its approach to storing and sharing research data, which must remain restricted due to ethical considerations. Chapter 20 is not a case study as some of the other chapters in this section but is instead a "how-to" for managing data for conversational analysis.

A similar diversity is found in the "syntax" chapters. Chapter 43 is essentially an introduction into tree banks as a computational tool for studying grammatical constructions in corpora, treebank software tools, and best practices for treebank creation, management, and citation. Chapter 44 is a case study and gives an overview of the data management, archiving, and distribution of grammatical judgments from a traditional elicitation-based syntactic study, but does not cover any aspects of recording since the author did not record elicitation sessions. Chapter 45 is similar, focusing on elicitation and data management in syntactic fieldwork, but is surprisingly lacking in discussion on archiving and data availability. Chapter 46 is another "how-to" chapter on data management in experimental syntactic data. This chapter in particular seems much more in the spirit of the volume, and discusses the issue of reproducibility in the management of experimental data in syntax. In general, these syntax chapters, like the sociolinguistic chapters, provide a range of topics such as software introduction, case studies, management of traditional elicitation data, and experimentation.

How each chapter will benefit a particular sociolinguist or syntactician will depend on how the focus of each chapter overlaps with the specific interests and research programs of the reader. Although The Handbook does contain dedicated chapters on syntactic data management, how those chapters relate to one's own work is sometimes questionable. Tree bank creation, for example, seems to overlap only minimally with theoretically-oriented syntactic research. The chapters that do talk about theoretical work and traditional elicitation each leave out details which could have been helpful to a wider audience. The most useful chapter is oriented towards experimental work.

The diversity of Part 2 is no doubt by design, and this means that while most readers will find some chapters in Part 2 useful, it is not necessary for the reader to give every chapter the attention that was given to the chapters in Part 1. It will be much more useful to identify potentially useful chapters, scan through then, then zero in on the most relevant chapters before potentially branching out. Still, some may have a difficult time finding exactly what they need in Part 2.

5 Online content

The Handbook is paired with an online self-study course at http://linguisticdatamanagement.org. The online course contains exercises linked to each chapter designed to help the reader practice and apply relevant skills. There are 13 lessons in total, each linked to a chapter in Part 1. The online lessons are short, and contain reviews of the original chapter, exercises, and suggested further readings.

How one approaches the online courses will depend on the individual. They might be useful in a classroom or for the independent reader with little familiarity with the subjects. They do not seem to be a primary component of The Handbook, however.

The most useful component of these online courses is how they link Part 1 chapters to relevant use-case chapters from Part 2. The online course linked to chapter 8, for example, cites chapters 34, 35, 43, and 45 as relevant case-use chapters to read alongside chapter 8. One wonders why these cross-references were not listed in the book itself and are only found in the online courses. Before reading the online courses, I found it sometimes difficult to determine which chapters in Part 2 would be most interesting and relevant for me and I found myself wishing for just such a cross-reference. Readers who only interact with the book itself and do not read the online courses will miss out on this information.

6 General comments on The Handbook

The Handbook does a good job of explaining the need for data management, and in Part 1, at least, it provides a solid foundation in data management principles. The area where The Handbook falls short is in Part 2, where it seems to lose some of its initial focus and the chapters become much more varied in both topic and utility. It seems that this was at least partially the intention of the editors who may have been attempting to make The Handbook both useful for individual scholars and broad in the topics that it covers. In some cases they were successful, but Part 2 can also be difficult to navigate. As already mentioned, some chapters in Part 2 read more as how-to guides for specific software and for data management within a certain field. I find these chapters to be the most useful, although there is no obvious way to single these chapters out. Other chapters detail the workflows of specific projects and these can be too project specific for their own good. The main issue with having numerous chapters describing specific projects is that the chapters themselves are often inapplicable to other projects, even projects that are in the same sub-discipline or utilize similar research methods. Since no two projects are exactly the same, detailing the specific protocols used in one's own research is less useful than giving more general advice that could be applied to a range of related projects.

A more generalizable approach to Part 2 might have had chapters like "Data management best practices for elicited data sets", "Data management best practices for theoretical syntacticians", "Data management best practices for experimental linguistic research", and so forth. Such chapters would be both immediately recognizable as relevant or irrelevant for the reader's own research and better suited to address issues likely faced by a wider range of researchers in the specific subdiscipline. Instead, Part 2 of The Handbook tends to opt for specificity without generalizability. There are chapters on syntactic data management, for example, but the chapters not necessarily generalizable to syntactic data management as a whole. Given the success that Part 1 had in generalizing data management for the entire field of linguistics, one does feel that something similar could have been done in Part 2. Restricting Part 2 in this way may have also reduced the overall number of chapters in the volume, making it more manageable for readers.

Overall, readers of The Handbook will finish with a deeper understanding of the need for data management and a strong grasp of the fundamentals of data management and best practices. These fundamentals can go a long way in improving one's own data management protocols. What will be lacking in some cases is a clear understanding of the application of these principles to typical situations in specific sub-disciplines.

Integrating The Handbook into one's own workflow, and using it as a resource for self-improvement is certainly possible. After reading Part 1, the researcher will be better prepared to discuss the issues surrounding linguistic data management, how to choose an archive, handle different types of data, use proper file types, create data management plans, and be ethical. The task of taking this knowledge and integrating it into one's own research is mostly up to the reader. Part 2 has some potential as a useful resource for integrating data management practices into existing research, but only if the existing research aligns with what is described in a particular chapter. For the readership of Glossa, those involved in experimental research will find the most benefit from the chapters in Part 2. Those whose research is not experimental or otherwise more purely theoretically oriented will have to spend more time customizing their own data management plans using Part 1 as a starting point.

Despite the shortcomings in Part 2, The Handbook remains a valuable resource. It is worth restating here that The Handbook is available for free online, and that readers can download only those chapters which are most relevant. This makes the above criticism on Part 2 less important for those who will access The Handbook online.

7 Putting The Handbook to use and starting the process of improving data management practices

In this final section I offer suggestions on how to best approach the task of updating one's data management protocols, especially for those who are starting from the beginning. Specific chapters from The Handbook are cited which provide appropriate reference chapters for each suggestion.

First, practice making data management plans for projects immediately, even if you do not have a current project. Several chapters from Part 1 can be referenced during this process, especially chapter 8. Consider the kinds of data that you are likely to deal with, how that data might be properly managed, and the feasibility of your data management plan, using chapter 5 as a reference for data type. While making such a plan keep in mind future uses of the data, like how the will data be used by others, how it will enhance the research (such as by allowing direct citation in publications to precise data sets in an archive), how you will ensure that the wishes of the people who provide the data are respected (chapters 4, 7, and 11). It is good practice to do as much as is reasonable to ensure that the claims being made in linguistic publication can be traced back to well-managed and accessible data sets. Of course, the rights of the language users supersede the desire for open data, and some data cannot be recorded for various reasons (Arnold 2021; Gribanova 2021), but wherever possible, researchers should record interactions with language users.

Start small if data management for large projects feels overwhelming. Many repositories and archives are happy to accept small deposits, especially if the deposits are relevant for the repositories stated areas of interest and if the data themselves are well organized. Chapter 7 will help to locate such a repository. The size of a deposit can always grow over time. Elicitation sessions that focus on grammatical judgments on specific constructions may only last for a few hours and create only a few dozen or so data points. A smaller project like this is a perfect place to familiarize oneself with recording, use of relevant software, data management, data preparation for archiving, and citation best practices. See chapters 44 and 45 for some examples of elicitation based data management.

Finally, a basic familiarity with coding and programming can go a long way. Although such skills are not required for data management, a beginner's level of skill in these areas will make data management easier. Data management can sometimes involve large data sets and processes which are tedious for humans but perfect for a computer. There is much automation that can be done in data management if one has the basic tools. Another advantage of having these skills, even just a beginner's level, is that it can help researchers understand how to make data easily accessible for as wide a range of use cases as possible. Even without using programming directly in data management, an understanding of some basics can make it easier for others to interact with your data. Chapter 6 lists some useful resources, and chapter 28 appropriately demonstrates how proper data organization can make data available for computational analysis even if the depositor is less knowledgeable on such methods.

Our field is undergoing a shift towards open data and along with that shift comes new pressures on linguists to update and improve upon their data management practices. The Handbook provides much needed guidance for not only linguists to improve their own methods, but for administrators to better appreciate the value of data sets as a product of research. The Handbook excels in describing the culture of data management in linguistics, identifying areas for improvement, and providing practical guidance in fundamental aspects of data management. Future work may better narrow that guidance for specific sub-disciplines in a way that is maximally useful for individual scholars. As the editors state, The Handbook is not a manual for data management and one should not rely on The Handbook alone, but it can serve as a useful resource along with other supplementary resources based on the reader's specific needs.

Acknowledgements

Thanks to Michael Yoshitaka Erlewine for providing helpful comments on an earlier draft of this review and to the editors of The Handbook.

Competing interests

The author has no competing interests to declare.

References

Alperin, Juan Pablo & Schimanski, Lesley A. & La, Michelle & Niles, Meredith T. & McKiernan, Erin C. 2021. The value of data and other non-traditional scholarly outputs in academic review, promotion, and tenure in Canada and the United States. In Berez-Kroeker, Andrea L. & McDonnell, Bradley & Koller, Eve & Collister, Lauren B. (eds.), *The open handbook of linguistic data management*, 171–184. Cambridge: MIT Press. DOI: https://doi.org/10.7551/mitpress/12200.003.0017

Arnold, Lynnette. 2021. Data management practices in an ethnographic study of language and migration. In Berez-Kroeker, Andrea L. & McDonnell, Bradley & Koller, Eve & Collister, Lauren B. (eds.), *The open handbook of linguistic data management*, 249–256. Cambridge: MIT Press.

Austin, Peter K. 2021. Language documentation and language revitalization. In Olko, Justyna & Sallabank, Julia (eds.), *Revitalizing endangered languages: a practical guide*, 199–212. Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/9781108641142

Berez-kroeker, Andrea L. & Gawne, Lauren & Kung, Susan & Kelly, Barbara F. & Heston, Tyler & Holton, Gary & Pulsifer, Peter. 2018. Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics* 56. 1–18. DOI: https://doi.org/10.1515/ling-2017-0032

Champieux, Robin & Coates, Heather L. 2021. Metrics for evaluating the impact of data sets. In Berez-Kroeker, Andrea L. & McDonnell, Bradley & Koller, Eve & Collister, Lauren B. (eds.), *The open handbook of linguistic data management*, 157–170. Cambridge: MIT Press. DOI: https://doi.org/10.7551/mitpress/12200.003.0016

Gawne, Lauren & Styles, Suzy. 2021. Situating linguistics in the social science data movement. In Berez-Kroeker, Andrea L. & McDonnell, Bradley & Koller, Eve & Collister, Lauren B. (eds.), *The open handbook of linguistic data management*, 9–25. Cambridge: MIT Press.

Gribanova, Vera. 2021. Managing data in a formal syntactic stuydy of an underinvestigated language (Uzbek). In Berez-Kroeker, Andrea L. & McDonnell, Bradley & Koller, Eve & Collister, Lauren B. (eds.), *The open handbook of linguistic data management*, 513–522. Cambridge: MIT Press. DOI: https://doi.org/10.7551/mitpress/12200.003.0049

Kaipuleohone. n.d. Kaipuleohone, University of Hawai‘i Digital Language Archive. https://hdl.handle.net/10125/4250. Accessed: 16 July, 2022.

Kendall, Tyler & Farrington, Charlie. 2021. The corpus of regional african american language. http://oraal.uoregon.edu/coraal. Version 2021.07. Eugene, OR: The Online Resources for African American Language Project.

OLAC (Open Language Archives Community). n.d. http://www.language-archives.org/archives. Accessed: 13 July, 2022.

re3data (Registry of Research Data Repositories). n.d. https://www.re3data.org/. Accessed: 13 July, 2022.

Thomason, Sarah G. 1994. The editor's department. *Language* 70(2). 409–413.