52 Managing Phonological Inventory Data in the Development of PHOIBLE

Steven Moran

1 Introduction

This data management use case describes PHOIBLE Online, a typological database of phonological inventory data from spoken languages developed in Moran (2012) and made publicly available online by Moran, McCloy, and Wright (2014).[1] The aim of this chapter is twofold. First, for developers of typological databases, we provide an overview of our data management workflow. We illustrate how we collect, curate, and disseminate the phonological inventory data in light of the issues raised by the contributions in Berez-Kroeker et al. (chapter 1, this volume). Second, for users and data consumers, we describe how to access the data, and we provide a brief overview of use cases and research questions that have been asked with PHOIBLE.

2 For developers: Data management workflow

Kung (chapter 8, this volume) discusses how to develop a data management plan (DMP). DMPs are now a standard requirement for many grant applications. The requirements of the DMP will differ from funding agency to funding agency and the topics of the DMP will differ from research project to research project. However, DMPs generally revolve around data collection and handling, documentation and metadata, data storage and preservation, data sharing, and ethical considerations. The data management workflow for PHOIBLE involves three steps:

1. Data collection and documentation
2. Data storage and preservation
3. Data sharing and reuse

Moran (2012) provides a detailed overview of PHOIBLE and its initial development. Herein we discuss in detail our current data management workflow, which has been applied successfully to similar typological data sets, for example, a database of reconstructed phonological inventories (BDPROTO; Marsico et al. 2018; Moran, Grossman, and Verkerk 2020) and a database of borrowed speech sounds (SegBo; Grossman et al. 2020). Our hope is that the issues discussed here are beneficial to other projects developing typological databases.

2.1 Data collection and documentation

PHOIBLE is a repository of cross-linguistic phonological inventory data from spoken languages, or in other words, the description of consonants, vowels, and tones in a large sample of the world's languages. These data have been extracted from primary source documents, such as grammars, and from existing databases, such as the UCLA Phonological Segment Inventory Database (UPSID; Maddieson 1984; Maddieson & Precoda 1990). We have compiled these different sources into a single interoperable resource through a data aggregation pipeline illustrated in figure 52.1.

The basic setup includes two metadata files (InventoryID-Bibtex.csv and InventoryID-LanguageCodes.csv), which index each inventory's bibliographic reference(s) and its unique language name identifier (colloquially called language code). Along with the raw data sources (e.g., spreadsheets containing phonological inventory data, database dumps from other segment inventory databases in various digital formats), we aggregate the different data sets and their metadata into a single aggregated table. An example is given in table 52.1. We make these data available via a web application (discussed in section 3) and as raw text in CSV format. All index files, raw data sources, the aggregation script, and aggregated data are openly available in our GitHub repository online.[2]

Regarding our data collection and source documentation, we provide a bibliographic record for each source, that is, its *doculect* (Cysouw & Good 2013), from which we extract information about a language's phonological

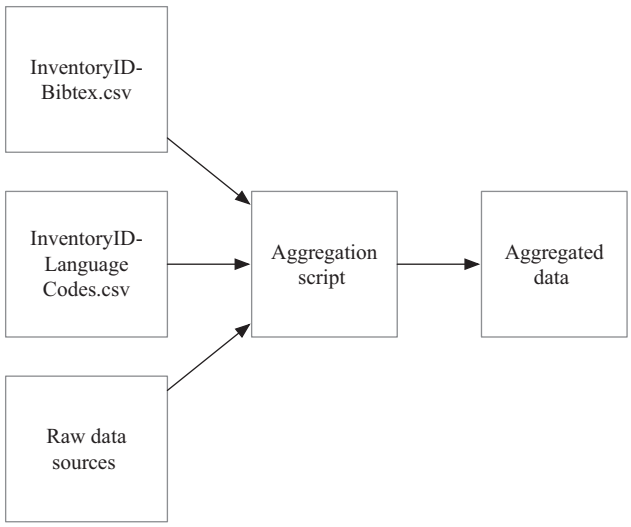


Figure 52.1
PHOIBLE aggregation pipeline.

inventory. We keep track of our references using BibTeX, a bibliography management software. BibTeX is stored in a plain text UTF-8 file.[3] An example of a BibTeX entry is:

```
@book{Cho1967,
Address={Uppsala},
Author={Cho, Seung-Bog},
Publisher={Almqvist and Wiksells},
Series={Acta Universitatis Upsaliensis, Studia
Uralica et Altaica Upsaliensia},
Title={A Phonological Study of Korean},
Volume={2},
Year={1967}}
```

Each bibliographic reference is given a unique identifier (here "Cho1967"), which we map to a particular PHOIBLE InventoryID in the CSV file InventoryID-Bibtex.csv that tracks which references belong to which data points, as illustrated in table 52.2.[4] We also keep track of the original source document from which the phonological inventory was extracted. Original sources are available online as PDF or as URLs to their original databases as listed in the Filename field.

The InventoryID-Bibtex.csv file is cross-referenced via the InventoryIDs to the file InventoryID-LanguageCodes.csv.[5] An example is given in table 52.3. This file contains inventory level metadata in terms of language name identifiers from the Glottolog (Hammarström, Forkel, & Haspelmath 2018) and from ISO 639-3.[6] This file also indicates the source from which the inventory came, for example, the Stanford Phonology Archive (SPA) or UPSID databases.

Mapping each inventory to a standard language code allows us to uniquely identify languages and dialects, and it also allows us to incorporate additional metadata from other linguistic and non-linguistic databases. For example, with Glottocodes we can merge into the PHOIBLE aggregated phonological inventory data information about a language's genealogical classification (e.g., language family, language genus), its geographic location (e.g., the macroarea in which its spoken, latitude and longitude coordinates), as provided by the Glottolog[7] and other resources. This additional information about languages and their speakers allows us to extend the PHOIBLE data to undertake quantitative analyses (see section 3), and it also provides additional information for displaying PHOIBLE's contents online in worldwide map formats (see figure 52.2).[8]

Good (chapter 3, this volume) discusses the scope of linguistic data in terms of the field's unusual position as a discipline that intersects humanities and science. In the development of PHOIBLE, data collection was initially a qualitative endeavor (independent of our database): a linguist went to the field to document and analyze the phonemic contrasts in a language. Given the different linguistic training and theoretical backgrounds of different linguists, phonological descriptions of the same language almost always differ in their

Table 52.1
Example of PHOIBLE aggregated data

InventoryID	Glottocode	LanguageCode	LanguageName	Phoneme	Source
1	kore1280	kor	Korean	m	spa
1	kore1280	kor	Korean	k	spa
1	kore1280	kor	Korean	i	spa
1	kore1280	kor	Korean	a	spa
1	kore1280	kor	Korean	p	spa

Table 52.2
Example from InventoryID-Bibtex.csv

InventoryID	BibtexKey	Filename
1	Cho1967	kor_SPA1979_phon.pdf
423	Martin1957	http://web.phonetik.uni-frankfurt.de/L/L2170.html

Table 52.3
Example from InventoryID-LanguageCodes.csv

InventoryID	Language Code	Glotto code	Language Name	Source
1	kor	kore1280	Korean	spa
423	kor	kore1280	Korean	upsid

number and composition of phonemes. To address this issue, we include multiple phonological inventories of the same language (or dialect) by different researchers and we leave it to users to decide which phonological description(s) they want to use in their analysis.[9]

The aggregation of the different sources in PHOIBLE is a quantitative product, which can be used for cross-linguistic comparison and statistical analysis. Different input sources come with different types of annotation (e.g., different transcription practices, different ways of describing phonological features). To make generalizations across all sources, we have typologized the data extracted from grammars into a uniform annotation using the International Phonetic Alphabet (IPA; International Phonetic Association 2015) encoded in the Unicode Standard (Unicode Consortium 2018). We discuss these decisions in detail in Moran and Cysouw (2018), a practical guide aimed at language scientists working in multilingual computational environments, in which we explicitly define a "strict" IPA that normalizes transcription practices across language descriptions from thousands of languages. For PHOIBLE users, we document our decisions regarding phonetic annotations in a set of notational conventions available online.[10]

2.2 Data storage and preservation

Mattern (chapter 5, this volume) discusses sustainability of linguistic data, its life cycle, and principles of data management in terms of sustainable data formats, version control, documentation, and archiving (see also Buszard-Welcher, chapter 10, this volume, regarding archiving and time-depths). For PHOIBLE, we collect the extracted data from grammars in easy-to-use working formats, such as Microsoft Excel and Google spreadsheets. These software tools allow us to quickly collect data and to share it with each other, so that we can compare and discuss our interpretations of the original source grammar. These working formats are not meant for long-term storage or archival preservation. Instead, we convert these spreadsheets into Unicode UTF-8 plain text CSV files, which we store in an online and publicly available GitHub repository.[11] GitHub lets us track changes over time and helps us manage our data and aggregation script through an issue tracker,[12] where each issue describes one clearly identified task (e.g., inventory ID 5 is missing a phoneme, correct it; update the aggregation script to include a new resource). Tasks can be assigned to the various contributors and we can set deadlines in order to manage the creation of new data releases, for example, version 2.0.

Han (chapter 6, this volume) gives an overview of data transformation in terms of data format and conversion, cleaning and reorganization, merging and processing. In our repository, we maintain our input data, metadata, and the scripts we wrote to aggregate the various sources together into a version that we release publicly. Once we are settled on a particular state of the data and feel they are ready for release, we create a version number that adheres to Semantic Versioning[13] and we release the data and our code together on GitHub.[14] This includes transforming the data into the Cross-Linguistic Data Format (CLDF) specification, which defines entities for languages, parameters (entities for comparative concepts), values (the measurements of these concepts), and sources (where the data come from). For details see Forkel et al. (2018) and the CLDF website.[15] Our CLDF data are encoded in four Unicode UTF-8 plain text CSV files that can be linked via primary keys into a relational database (Moran & McCloy 2019a). These tables provide the input format to update our web interface that is in the Cross-Linguistic Linked Data (CLLD) project,[16] which hosts many other typological data sets, such as the World Atlas of Language Structures (WALS; Dryer & Haspelmath 2013), the Automated Similarity Judgment Program (ASJP; Wichmann et al. 2019), and Tsammalex (Naumann et al. 2015).

Once we release a version of PHOIBLE on GitHub, we then archive that release on Zenodo (see Andreassen, chapter 7, this volume, who discusses issues of data archiving).[17] Zenodo adheres to FAIR (Findable, Accessible,

Interoperable, and Reusable) principles (Wilkinson 2016) and provides a new digital object identifier (DOI) for each new release along with a bibliographic citation that users can use when citing a particular release of the data in their research papers.

2.3 Data sharing and reuse

The data from PHOIBLE are shared via our GitHub repository, through Zenodo archived releases, and via the online web application. Ongoing development work on GitHub contains the most-up-to-date data and code.

It is not only pertinent to share data, but also the code that is used to transform and generate it (Barnes 2010). The necessity of reproducibility and replicability is particularly important in light of large quantitative data sets, including typological databases, on which statistical analyses are undertaken (Moran 2016). When these sources are made publicly available, it is pertinent that reproducibility and replicability are transparent, as noted by Gawne and Styles (chapter 2, this volume). This is not only important for the data publishers, but for data consumers. Consider, for example, how the reuse of data in linguistic studies leads to findings that may change from one analysis to another because of different language sample sizes (e.g., Everett, Blasi, & Roberts 2015 vs. Roberts 2018) or use of different statistical methods (e.g., Hay & Bauer 2007 vs. Moran, McCloy, & Wright 2012).[18]

Champieux and Coates (chapter 12, this volume) describe metrics for evaluating the impact of data sets. Metrics are obviously useful for relaying to funding agencies the impact of a data set, whether through scientific citations or use. For PHOIBLE, we track citations through Google Scholar, and to increase exposure and discovery, we list PHOIBLE in the Open Language Archives Community (OLAC).[19] OLAC tracks and evaluates metadata quality through data integrity checking (e.g., whether ISO 639-3 codes are valid in light of the fact that they are periodically updated). Due to our rigorous adherence to identifying data sources with language codes, OLAC gives PHOIBLE an overall five out of five star rating for metadata quality. This means users know which languages or dialects are present in our worldwide sample.

3 For users: Overview and use cases

The 2019 edition of PHOIBLE Online includes 3,020 inventories that contain 3,183 distinct phonemes found in 2,186 languages (Moran & McCloy 2019b). As such, PHOIBLE represents the largest database of sound systems about the world's languages as it builds on and brings together the SPA (Crothers et al. 1979), the UPSID (Maddieson 1984; Maddieson & Precoda 1990), the South American Phonological Inventory Database (SAPHON; Michael, Stark, & Chang 2012), Alphabets of Africa (Hartell 1993), Systèmes alphabétiques des langues africaines (Chanard 2006), the database of Eurasian phonologies (Nikolayev, Nikulin, & Kukhto 2015), Australian phonemic inventories (Round 2019), and hundreds of individual data points extracted from grammars by the PHOIBLE editors.

PHOIBLE Online is a web application that provides access to the phonological inventory data for browsing and searching by language, inventory, phoneme, and source. Languages are displayed on a worldwide map and each data point is color-coded for language family, as illustrated in figure 52.2. Each point can be clicked on and the user is taken to the phonological inventory—available both in a table format and as an IPA chart.

PHOIBLE Online allows users to explore the sound systems of the world's languages. We have received many reports that it is a useful tool in linguistic courses on phonology, typology, and languages of the world. We also make the PHOIBLE data available via one large aggregated CSV file (in tabular format; described in section 2.1). In linguistic courses with a quantitative focus, the phoneme inventory data and associated metadata (e.g., language family, geo-coordinates) are a powerful resource for learning descriptive statistics (e.g., what is the frequency distribution of sounds in the world's languages? how many data points belong to each language family?) and also for learning how to plot data with software such as R (e.g., plot the frequency distribution of sounds; plot the languages on a worldwide map).

There has also been a broad range of published research that uses the PHOIBLE data, including studies on linguistic and genetic diversity (Creanza et al. 2015), phonetics (Dediu & Moisik 2019), phonology (Cohen Priva 2017), typology (Nikolaev & Grossman 2018), historical linguistics (Barrack, McCloy, & Wright 2014), and computational linguistics (Johny, Gutkin, & Jansche 2019).[20]

Some of my own research has involved testing whether there is a correlation between the population size of speech communities and the number of sounds in their languages (Moran, McCloy, & Wright 2012), investigating whether there are compensations

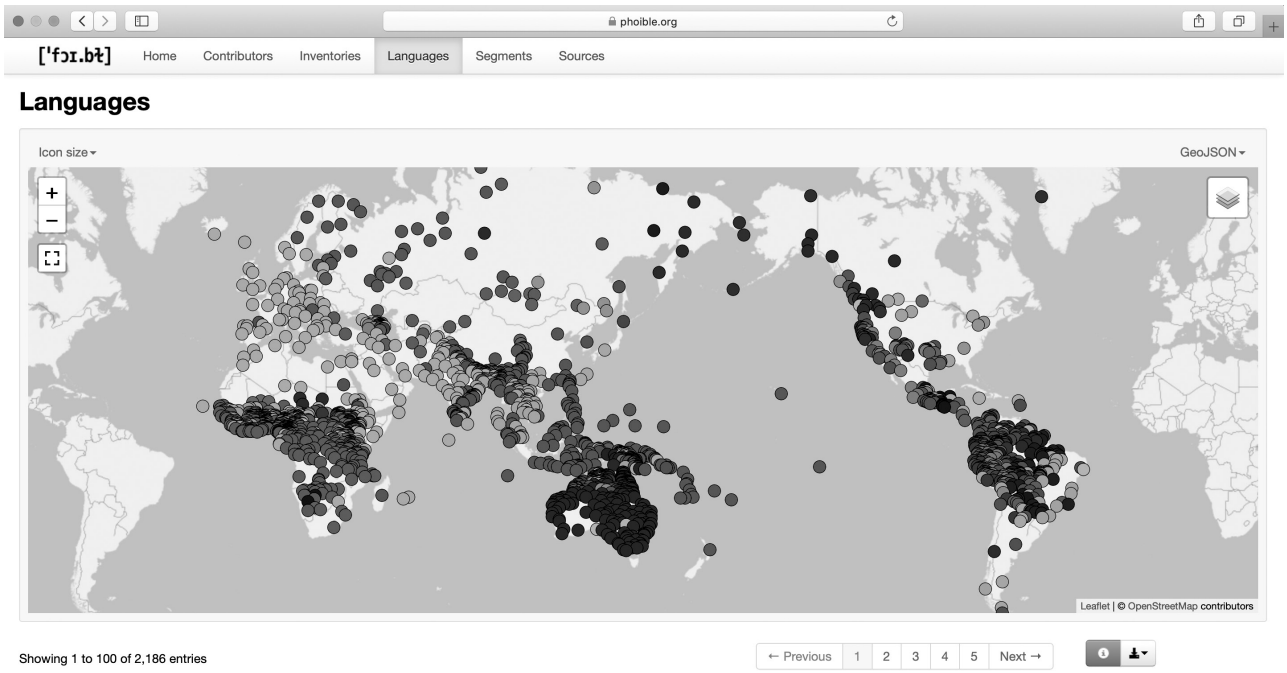


Figure 52.2
PHOIBLE Online.

in phonological system complexity cross-linguistically (Moran & Blasi 2014), and showing that labiodental sounds (such as "f" and "v") were innovated after the Neolithic period due to changes in food processing technology and its effect on human bite configuration (Blasi et al. 2019). In each of these studies, we were able to extend the PHOIBLE data with information about linguistic and non-linguistic variables, such as language family phylogenies, demography, and subsistence type, via our use of standardized language codes and the ability to use them to link together different databases, such as the Glottolog (Hammarström, Forkel, & Haspelmath 2019) and D-PLACE (Kirby et al. 2016). This illustrates the power of using linguistic metadata, beyond just uniquely identifying data points, so that we can undertake interdisciplinary research.

Notes

1. https://phoible.org/.

2. https://github.com/phoible/dev.

3. https://github.com/phoible/dev/blob/master/data/phoible-references.bib.

4. https://github.com/phoible/dev/blob/master/mappings/InventoryID-Bibtex.csv.

5. https://github.com/phoible/dev/blob/master/mappings/InventoryID-LanguageCodes.csv.

6. https://iso639-3.sil.org/.

7. https://glottolog.org/meta/downloads.

8. https://phoible.org/languages.

9. For more information, see our comprehensive FAQ: http://phoible.github.io/faq/.

10. http://phoible.github.io/conventions/.

11. https://github.com/phoible/dev/tree/master/raw-data.

12. https://github.com/phoible/dev/issues.

13. https://semver.org/.

14. https://github.com/phoible/dev/releases.

15. https://cldf.clld.org/.

16. https://clld.org/.

17. https://zenodo.org/.

18. See also enlightening discussions by Hatton (1997), Roberts and Winters (2013), and Silberzahn et al. (2015).

19. http://www.language-archives.org/.

20. For the full list and access to these papers, see: https://scholar.google.com/scholar?hl=en&as_sdt=2005&cites=5769811163093889 28&scipsc=.

References

Barnes, Nick. 2010. Publish your computer code: It is good enough. *Nature* 467 (7317): 753–753. doi:10.1038/467753a.

Barrack, Charles M., Daniel R. McCloy, and Richard A. Wright. 2014. Did murmur spread in Pre-Proto-Indo-European? *Indogermanische Forschungen* 119 (1): 149–158.

Blasi, Damián E., Steven Moran, Scott R. Moisik, Paul Widmer, Dan Dediu, and Balthasar Bickel. 2019. Human sound systems are shaped by post-Neolithic changes in bite configuration. *Science* 363 (6432). doi:10.1126/science.aav3218.

Chanard, C. 2006. Systèmes alphabétiques des langues africaines. http://sumale.vjf.cnrs.fr/phono/.

Cohen Priva, Uriel. 2017. Informativity and the actuation of lenition. *Language* 93 (3): 569–597.

Creanza, Nicole, Merritt Ruhlen, Trevor J. Pemberton, Noah A. Rosenberg, Marcus W. Feldman, and Sohini Ramachandran. 2015. A comparison of worldwide phonemic and genetic variation in human populations. *Proceedings of the National Academy of Sciences* 112 (5): 1265–1272.

Crothers, John H., James P. Lorentz, Donald A. Sherman, and Marilyn M. Vihman. 1979. Handbook of phonological data from a sample of the world's languages: A report of the Stanford Phonology Archive. Department of Linguistics, Stanford University.

Cysouw, Michael, and Jeff Good. 2013. Languoid, doculect and glossonym: Formalizing the notion "language." *Language Documentation and Conservation* 7:331–359.

Dediu, Dan, and Scott R Moisik. 2019. Pushes and pulls from below: Anatomical variation, articulation and sound change. *Glossa* 4 (1): 1–33. https://www.glossa-journal.org/articles/10.5334/gjgl.646/.

Dryer, Matthew S., and Martin Haspelmath, eds. 2013. *WALS Online*. Leipzig, Germany: Max Planck Institute for Evolutionary Anthropology. https://wals.info/.

Everett, Caleb, Damián E. Blasi, and Seán G. Roberts. 2015. Climate, vocal folds, and tonal languages: Connecting the physiological and geographic dots. *Proceedings of the National Academy of Sciences* 112 (5): 1322–1327. doi:10.1073/pnas.1417413112.

Forkel, Robert, Johann-Mattis List, Simon J. Greenhill, Christoph Rzymski, Sebastian Bank, Michael Cysouw, Harald Hammarström, Martin Haspelmath, Gereon A. Kaiping, and Russell D. Gray. 2018. Cross-linguistic data formats, advancing data sharing and re-use in comparative linguistics. *Scientific Data* 5:180205.

Grossman, Eitan, Elad Eisen, Dmitry Nikolaev, and Steven Moran. 2020. SegBo: A database of borrowed sounds in the world's languages. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020)*. http://www.lrec-conf.org/proceedings/lrec2020/pdf/2020.lrec-1.654.pdf.

Hammarström, Harald, Robert Forkel, and Martin Haspelmath. 2018. *Glottolog 3.3*. Jena, Germany: Max Planck Institute for the Science of Human History. https://glottolog.org/.

Hammarström, Harald, Robert Forkel, and Martin Haspelmath. 2019. *Glottolog 4.0*. Jena, Germany: Max Planck Institute for the Science of Human History. https://glottolog.org/. Accessed October 7, 2019.

Hartell, Rhonda L., ed. 1993. *Alphabets des langues africaines*. Dallas: UNESCO and Société Internationale de Linguistique.

Hatton, Les. 1997. The T experiments: Errors in scientific software. *IEEE Computational Science and Engineering* 4 (2): 27–38.

Hay, Jennifer, and Laurie Bauer. 2007. Phoneme inventory size and population size. *Language* 83 (2): 388–400. doi:10.1353/lan.2007.0071.

International Phonetic Association. 2015. *International Phonetic Alphabet*. International Phonetic Association. https://www.internationalphoneticassociation.org.

Johny, Cibu C., Alexander Gutkin, and Martin Jansche. 2019. Cross-lingual consistency of phonological features: An empirical study. *Interspeech* 2019 (September 15–19): 1741–1745.

Kirby, Kathryn R., Russell D. Gray, Simon J. Greenhill, Fiona M. Jordan, Stephanie Gomes-Ng, Hans-Jörg Bibiko, Damián E. Blasi, et al. 2016. D-PLACE: A global database of cultural, linguistic and environmental diversity. *PLoS One* 11 (7): e0158391.

Maddieson, Ian. 1984. *Pattern of Sounds*. Cambridge: Cambridge University Press.

Maddieson, Ian, and Kristin Precoda. 1990. Updating UPSID. *UCLA Working Papers in Phonetics* 74:104–111.

Marsico, Egidio, Sebastien Flavier, Annemarie Verkerk, and Steven Moran. 2018. BDPROTO: A database of phonological inventories from ancient and reconstructed languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, ed. Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, et al., 1654–1658. Paris: European Language Resources Association (ELRA).

Michael, Lev, Tammy Stark, and Will Chang. 2012. *South American Phonological Inventory Database*. http://linguistics.berkeley.edu/saphon/en/.

Moran, Steven. 2012. Phonetics information base and lexicon. PhD dissertation, University of Washington.

Moran, Steven. 2016. Commentary: Issues of time, tone, roots and replicability. *Journal of Language Evolution* 1 (1): 73. doi:10.1093/jole/lzv011.

Moran, Steven, and Damián Blasi. 2014. Cross-linguistic comparison of complexity measures in phonological systems. In *Measuring Grammatical Complexity*, ed. Frederick J. Newmeyer and Laurel Preston, 217–240. Oxford: Oxford University Press.

Moran, Steven, and Michael Cysouw. 2018. *The Unicode Cookbook for Linguists: Managing Writing Systems Using Orthography Profiles*. Berlin: Language Science Press. doi:10.5281/zenodo.1300528. http://langsci-press.org/catalog/view/176/889/1135-2.

Moran, Steven, Eitan Grossman, and Annemarie Verkerk. 2020. Investigating diachronic trends in phonological inventories using BDPROTO. *Language Resources and Evaluation*. https://doi.org/10.1007/s10579-019-09483-3.

Moran, Steven, and Daniel McCloy. 2019a. *cldf-datasets/phoible: PHOIBLE 2.0.1 as CLDF Dataset*. https://doi.org/10.5281/zenodo.2677911.

Moran, Steven, and Daniel McCloy, eds. 2019b. *PHOIBLE 2.0*. Jena: Max Planck Institute for the Science of Human History. https://phoible.org/.

Moran, Steven, Daniel McCloy, and Richard Wright. 2012. Revisiting population size vs. phoneme inventory size. *Language* 88 (4): 877–893. doi:10.1353/lan.2012.0087.

Moran, Steven, Daniel McCloy, and Richard Wright, eds. 2014. *PHOIBLE Online*. Leipzig, Germany: Max Planck Institute for Evolutionary Anthropology. http://phoible.org/.

Naumann, Christfried, Tom Güldemann, Steven Moran, Guillaume Segerer, and Robert Forkel, eds. 2015. *Tsammalex*. Leipzig, Germany: Max Planck Institute for Evolutionary Anthropology. https://tsammalex.clld.org.

Nikolaev, Dmitry, and Eitan Grossman. 2018. Areal sound change and the distributional typology of affricate richness in Eurasia. *Studies in Language* 42 (3): 562–599.

Nikolayev, Dmitry, Andrey Nikulin, and Anton Kukhto. 2015. The database of Eurasian phonological inventories. http://eurasianphonology.info.

Roberts, Seán G. 2018. Robust, causal, and incremental approaches to investigating linguistic adaptation. *Frontiers in Psychology* 9:1–22. doi:10.3389/fpsyg.2018.00166. https://www.frontiersin.org/article/10.3389/fpsyg.2018.00166.

Roberts, Seán, and James Winters. 2013. Linguistic diversity and traffic accidents: Lessons from statistical studies of cultural traits. *PLOS One* 8 (8): 1–13. https://doi.org/10.1371/journal.pone.0070902.

Round, Erich. 2019. Australian phonemic inventories contributed to PHOIBLE 2.0: Essential explanatory notes (Version 1.0). Zenodo. http://doi.org/10.5281/zenodo.3464333.

Silberzahn, Raphael, Eric L. Uhlmann, Daniel P. Martin, Pasquale Anselmi, Frederik Aust, Eli C. Awtrey, Štěpán Bahník, et al. 2015. Many analysts, one dataset: Making transparent how variations in analytical choices affect results. https://osf.io/gvm2z.

Unicode Consortium. 2018. *The Unicode Standard*, version 11.0.0. Technical report. Mountain View, CA: Unicode Consortium. http://www.unicode.org/versions/Unicode11.0.0/.

Wichmann, Søren, Eric W. Holman, and Cecil H. Brown. 2019. The ASJP database, version 18. https://asjp.clld.org.

Wilkinson, Mark D. 2016. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data* 3:160018. doi:10.1038/sdata.2016.18.

This is a section of doi:10.7551/mitpress/12200.001.0001

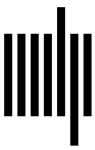
The Open Handbook of Linguistic Data Management

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 open access edition of this book was made possible by generous funding and support from the authors



The MIT Press

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