Annual Review of Linguistics

Recent Advances in Technologies for Resource Creation and Mobilization in Language Documentation

Andrea L. Berez-Kroeker, Shirley Gabber, and Aliya Slayton

Department of Linguistics, University of Hawaiʻi at Mānoa, Honolulu, Hawaii, USA; email: andrea.berez@hawaii.edu

Annu. Rev. Linguist. 2023. 9:195–214

First published as a Review in Advance on October 19, 2022

The *Annual Review of Linguistics* is online at linguistics.annualreviews.org

https://doi.org/10.1146/annurev-linguistics-031220-120504

Keywords

language documentation, technology, recent advances, resource creation, resource mobilization

Abstract

Language documentation as a subfield of linguistics has arisen over the past roughly two and a half decades more or less simultaneously with the widespread availability of inexpensive hardware and software for creating, storing, and sharing digital objects. Thus, in some ways the history of advancements within the discipline is also a history of how technological tools have been developed, tested, adopted, and eventually abandoned as newer technologies appear. In this article we examine some recent technologies used both for creating documentary resources, usually considered to include recorded language events in a variety of genres and settings and enough annotation to make them decipherable, and for then mobilizing those resources so that they can be used and shared in language learning, reclamation, revitalization, and analysis.

1. INTRODUCTION

Annotation: contextualizing and explanatory information that helps a user understand data; in language documentation this includes metadata, transcriptions, translations, notes, etc.

Language documentation as a subfield of linguistics (e.g., Himmelmann 1998, 2006; Woodbury 2011) has arisen over the past roughly two and a half decades more or less simultaneously with the widespread availability of inexpensive hardware and software for creating, storing, and sharing digital objects. Thus, in some ways the history of advancements within the discipline is also a history of how technological tools have been developed, tested, adopted, and eventually abandoned as newer technologies appear. In this article we examine some recent technologies used both for creating documentary resources, usually considered to include recorded language events in a variety of genres and settings (Himmelmann 1998) and enough annotation to make them decipherable, and for then mobilizing those resources so that they can be used and shared in language learning, reclamation, revitalization, and analysis.

This article is laid out as follows. In Section 2, we give the reader an overview of the development of standards for technology in language documentation since the early 2000s. In Section 3, we turn to an overview of the newest tools that are part of a common language documentation workflow for resource creation: data management, making recordings, annotation, and archiving. We then turn our attention in Section 4 to tools for the mobilization of language documentation resources for revitalization, reclamation, and language learning, including tools for making transcriptions more widely accessible, recent advances in geospatial applications, and new directions in media creation. We close in Section 5 with a brief look at how digital tools have been used during the COVID-19 pandemic, and some thoughts about the future of tool development for documentation and resource mobilization.[1]

2. THE EVOLUTION OF STANDARDS

All technological advances occur with some awareness of expected standards that make explicit the priorities of users that software should embody. The software and tools described in this article have taken place against a backdrop of two decades of discussions in the linguistics technology community and of the ever-renewing technological advances in the world at large. In this section we discuss the evolution of standards for language documentation and revitalization tools since the early 2000s, a period of rapidly expanding access to technology.

2.1. Early Advances

We mark the beginning of the disciplinary conversation about standards for technology with Bird & Simons's (2003) article about the paramount need to ensure the portability of digital language products. This landmark article outlined seven parameters along which researchers should ensure the long-term preservation, sharing, and reuse of digital language documentation and description materials. These parameters are content, format, discovery, access, citation, preservation, and rights, and this article set in motion an international discussion about best practices for managing data in a newly digital world.

[1] Given the ubiquity of digital methods in general, and the word limit of this review in particular, it is not possible to include every corner of the field that is affected by advances in technology, and we have had to make some hard decisions about which areas to exclude. We have opted to leave out some obvious realms, including lexicon creation, which has been dominated for the last decade by FieldWorks Language Explorer (FLEx) (SIL International 2022, Beier & Michael 2022), and the staggering array of mobile applications for documentation and language learning (see, e.g., Surma & Truong forthcoming), as well as many areas in computational linguistics and natural language processing. We have instead chosen to focus on tools that are likely to be relevant to a greater number of practitioners or that have greater potential to change the way documentary linguists work.

Early efforts to develop these practices, and to educate the academic linguistic audience about them, include the Electronic Metastructure for Endangered Language Data (E-MELD) project (Boyton et al. 2006). E-MELD was a five-year, multimillion-dollar project that launched in 2001 and sought to develop methods for digitizing the ubiquitous analog collections of fieldwork data[2] from the previous century of linguistic investigation. Among the deliverables of E-MELD was the online School of Best Practices website (**http://emeld.org/school/index.html**). The School was aimed at individual researchers in possession of, for example, paper lexical slip files or boxes of audio tapes who wanted to learn to properly digitize them for posterity. The School contained virtual classrooms on media and technology types, case studies of digitization projects, a reading room, a tools room, and more. Arguably, E-MELD and the School educated a generation of linguists in digital standards and paved the way for much of the software discussed in the rest of this article.

Metadata: information about data; in language documentation this includes information such as language name, participants' names, date of recording, location

Another important development of this early era was the creation of the Open Language Archive Community (OLAC) (Simons & Bird 2003), a network of international archives whose administrators still cooperate on standards for the preservation and description of endangered language resources. The OLAC search engine (**http://dla.library.upenn.edu/dla/olac/index.html**) allows for the discovery of thousands of items in participating OLAC archives worldwide. Archiving has since become a regular part of documentary practice, and recent developments in this area are discussed further in Section 3.4.

The early 2000s also saw the rise of two prominent loci of software development. The Max Planck Institute (MPI) for Psycholinguistics housed the archives of the Documentation of Endangered Languages projects and developed a suite of tools known as Language Archiving Technology (LAT) (Koenig et al. 2009). The LAT suite included tools for metadata creation, resource uploading, lexicon development, and, in its most widely adopted tool ELAN (EUDICO Linguistic Annotator), transcription of multimedia resources (see Section 3.3). SIL International (2022) has also been a prominent developer of tools over the years; early projects include the no-longer-supported Shoebox and Field Linguist's Toolbox for transcription and lexicon development, which led to the more robust FieldWorks Language Explorer (FLEx). While the tools developed by MPI have had a primarily academic audience in mind, the tools from SIL International have been directed mostly at a more diverse user base, including language workers with no formal linguistics training.

At the same time, more general language-related developments were taking place that influenced the direction of software development. The 2007 publication of the ISO 639-3 standard for three-letter codes for representing language names (ISO 2007) allowed for the cross-referencing of materials in or on a particular language without the need for agreement on the language name itself or even on its status as a dialect, language, or macrolanguage. The rise of Unicode allowed for the interoperability of different scripts across platforms and fonts by assigning a font-independent code point to each character in a given script (for more on linguistic applications of Unicode, see, e.g., Moran & Cysouw 2018). Cheap data storage and inexpensive, high-quality solid-state digital recording devices put access to lossless audio formats such as WAV into everyone's hands, and

[2] A brief note about the use of the term data in this article: We acknowledge and respect that languages and samples thereof, whether occurring ephemerally in an oral or signed modality or fixed via writing or recording, represent people (individually) and peoples (collectively). Such samples are gifts, without which the endeavor of language documentation would have no foundation. Our use of the term data is not meant to distance language from the people who produce(d) it or to relegate language to a commodity; rather, it is used as a shorthand to refer to the fixed records of instances of language around which so much discussion of technology, as a forum for promoting, protecting, and sharing it, has developed. For more on the relationality between language, language workers, and language data, see, for example, Leonard (2017, 2021) and Gaby & Woods (2020).

knowledge of basic programming, including for presentation formats such as HTML, became more common among language workers.

Open-Source software: software for which the original code is freely available and can be modified and distributed by others

2.2. The Second Wave of Standards Development

Once digital standards for portability were more or less in place, attention turned to other kinds of standards relating to language documentation technology. The first of these pertains to new ethical considerations that digital technology brings to the field of language documentation. Nathan (2010) discusses the ethics of audio recording: Linguists must not forget that language consultants are agentive and social, and because audio recording preserves the identities of people who participate in a recording session, linguists are obligated to provide as rich and faithful a recording of the interactional qualities of the recording session as possible. A 2010 special issue of *Language & Communication* on ethics and linguistic fieldwork aimed to examine the ethical implications of digital methods for language documentation. Among the papers found in this issue are O'Meara & Good (2010), which examines the appropriate use of the legacy materials that are often found in archives; Robinson (2010), which discusses informed consent in communities with little first-hand experience in the implications of putting language data on the Internet; and Debenport (2010), which explores community-specific notions of language ownership as they contrast with non-Indigenous understanding of "universal ownership" (e.g., Hill 2002) of language information in the digital world.[3]

A second recent discussion pertains to the role of colonial practices in the development of technology for language reclamation in Indigenous communities, with an aim toward advocating for digital tools that keep decolonization goals at the forefront. For example, Brinklow et al. (2019) pose three guiding principles for language technology in Indigenous contexts in Canada. The first guiding principle is that language revitalization technologies are a means to an end, not an end in themselves. "To properly scope a given technology, it is necessary to have a firm understanding that it is **people** that revitalize a language and that technology can merely multiply their efforts" (Brinklow et al. 2019, p. 403; emphasis in the original). This approach centers people over tools, even as tools can motivate and assist people in reaching language goals. The second guiding principle put forth by the authors is attention to the Indigenous user experience, especially culturally salient contexts such as the orality of language and protocols over how, by whom, and to whom knowledge can be transferred. Their third principle pertains to data sovereignty and Open Source. Regarding data sovereignty, a growing issue of concern in Canada (among other places, including Australia and Aotearoa), Indigenous communities often claim ownership of language data, a belief with which many of the data policies found in software user agreements are out of step. "[L]anguage technology must. . .ensure that communities are able to protect themselves in digital spaces" (Brinklow et al. 2019, p. 404). Similarly, Open-Source software is a countermeasure to digital colonization: Inaccessible software "reinforce[s] the power of a small group of experts at the expense of the larger language revitalization community" (Brinklow et al. 2019, p. 405). First Peoples' Cultural Council (2020) has recently published *Check Before You Tech!*, a guide to selecting software to be used in Indigenous language programs. The guide walks people selecting technology for use in revitalization through a series of questions such as How will the technology help us? Who is providing the technology? Who owns, controls, and accesses the data? and What are the plans for security, privacy, and disaster recovery?

[3] A reviewer rightly points out that university-based research in language documentation intersects with ethical considerations vis-à-vis protocols needing approval by institutional review boards or research ethics committees (see, e.g., Bowern 2010, Warner 2014, Good 2018).

In recent years, discussion of standards has been influenced by both the open scholarship movement and the push toward reproducible research in the social sciences. In both of these movements, increased broad access to data is considered assets for FAIR (findable, accessible, interoperable, reusable; see Wilkinson et al. 2016) research practices. In the former, removing financial barriers to data and publication is seen as a benefit for the common good; in the latter, researchers are encouraged to make research data shareable and transparent in service of better science (e.g., Berez-Kroeker et al. 2018, 2022; but see also Dobrin 2021). While these movements seem to be of obvious benefit, recordings of natural language use can be sensitive or private, or they may contain information that could be of financial benefit to parties outside the language community; in these cases, open data may be unethical (Holton et al. 2022). Seyfeddinipur et al. (2019) outline the challenges of public access, copyright, and other legal issues and provide strategies for reaching solutions in various settings, including language communities, universities, and archives.

3. TECHNOLOGY FOR THE LANGUAGE DOCUMENTATION WORKFLOW

The typical workflow of language documentation involves a more or less widely accepted series of steps, including making recordings, creating metadata, annotating media files, backing up files, and depositing materials into an archive for safekeeping (e.g., Cox 2022 or Thieberger & Berez 2011; for more on the nontechnological aspects of the documentary workflow, see also, e.g., Bowern 2015 or Meakins et al. 2018). In this section we examine the more common tools used in this workflow.

3.1. Workflow Management

Let us first look at recent software designed to manage the documentary workflow. Thieberger (2016) presents the results of a survey of field linguists about their habits for tracking metadata and archiving materials. The survey found that most respondents do not use dedicated software, relying instead on a spreadsheet, pen and paper, or word processing software for creating and tracking metadata. Those who do use dedicated software [i.e., linguistic metadata creation tools such as Arbil (TLA 2020) or CMDI Maker (**https://cmdi-maker.uni-koeln.de/**)] found them difficult to use. As for archiving, while half of the respondents use metadata tools to upload metadata with their materials,[4] many still enter metadata by typing them from paper notes (and many respondents admitted to not archiving their materials at all). This points to the need for simple, user-friendly tools to take the drudgery out of the process of metadata creation.

One such tool that is under development is Lameta (Hatton et al. 2021), an outgrowth of the earlier SayMore tool (Moeller 2014, Lee 2022). Lameta is designed to help users track projects (e.g., sponsors, access protocols), people [e.g., names, birth year, language(s), education, ethnicity, occupation], and session metadata (e.g., media files and associated transcripts, descriptions, topics, participants, subject and working languages, dates, location) and then export all files and metadata into formats required by some of the most widely used archives.

3.2. Recording Media

In general, for audio and video recording, documentary linguists rely on advances in industry technologies paired with guidance from colleagues with particular experience in making

[4] It is not clear from the survey results whether this number includes those who were allowed to submit a spreadsheet for ingestion or whether it includes only dedicated tools.

quality recordings for the purposes of language documentation. Guidance on audio standards and equipment selection for language documentation is easy to obtain (e.g., Bowern 2015).

Virtual reality (VR): a computer-generated interactive and immersive environment, usually experienced through a headset

In recent years, the importance of video documentation in revealing the aspects of language use that are not interpretable from audio alone, including signs, gestures, reference, deixis, interaction, and more, has come to the forefront (for a recent discussion, see Pentangelo 2020). However, the shift to video brings with it a host of choices, as the available selection of video cameras, settings, codecs, and formats is dizzying. As Seyfeddinipur & Rau (2020) note, most of the information found online is usually directed not toward linguists or the needs of language documentation but rather toward documentary filmmakers and YouTube content creators. To further complicate matters, the development of new technologies moves faster than the typical time needed to distill information and publish it to make it useful for language documenters. Nonetheless, Seyfeddinipur & Rau provide up-to-date guidance on current recommendations for cameras, settings, encodings, and formats while still acknowledging that their advice is likely to become increasingly moot over the next five years. They also stress the need for linguists to seek specialized training in video creation, because "[a] video camera can **see** but it cannot **look** and it can **hear** but it cannot **listen**" (p. 507; emphasis in the original), and the creation of language documentation requires more skills than an amateur home video of a birthday party.

Advances in virtual reality (VR) and 360° video mean that the field is on the precipice of some potentially exciting new modalities for language documentation. Pentangelo (2020) describes his 10-plus-hour corpus of Kanien'kéha spontaneous conversation created using 360° video and ambisonic audio, which allows for a more fully immersive experience of the corpus. Among other advantages, this setup allows for an expanded frame in which participant configurations that are too complex to capture with traditional video can be recorded simultaneously. **Figure 1*a*** is a diagram of a conversation between six participants in a boardroom with multiple six-foot conference tables. **Figure 1*b*** is a still from the 360° video, which easily captures the faces of all six participants, a feat that would be possible only with multiple traditional cameras.

3.3. Annotation

Annotation, a central activity in the documentation workflow, transforms the raw data of a recording into primary data that can then be derived via many analytical or presentational methods and turned into linguistic analyses, teaching materials, or useful language products (Himmelmann 2006; for a discussion of raw versus primary data in language documentation, see Himmelmann 2012). The manual transcription software landscape has long been dominated by ELAN (Sloetjes & Wittenburg 2008).[5] The main function of ELAN is straightforward: to time-align annotations to media, that is, to link typed annotations to media files via offsets that reference the starting and ending points of the annotation so that the exact point in the media that an annotation references can be easily located. ELAN has remained the most widely used annotation software in the field in part because of its portability: ELAN is Open Source, Unicode compliant, and XML based, such that the information encoded will not be lost should ELAN cease to be supported in the future (Berez 2007). Tier structure is highly customizable, allowing for multiple participants and multiple layers of information, including morphemic parsing, glossing, free translations, gestures, metacommentary, and even phone-level alignment. Over the years, the developers have increased functionality and user-friendliness by including specialized interfaces for different

[5] Other software, including FLEx (SIL International 2022) and Praat (Boersma & Weenink 2022), also occupy the annotation software space.

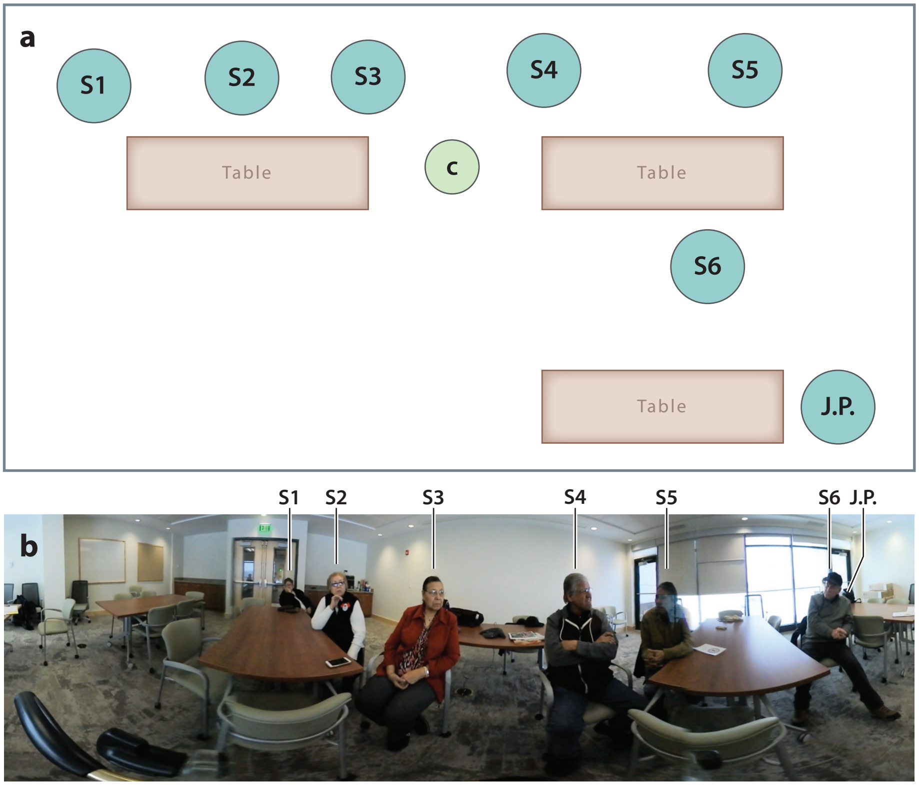

Figure 1

(*a*) A diagram of the conversational situation. Lowercase c refers to the camera; S1–S6 refer to the six participants; and the initials J.P. refer to the researcher, Joseph Pentangelo. (*b*, *left to right*) S1, S2, S3, S4, S5, S6; the researcher's body is obstructed by S6. Figure adapted with permission from Pentangelo (2020).

tasks; enhancing automaticity of some aspects of the workflow, including audio recognizers and linkage to a lexicon; and adding corpus-level query options.

3.3.1. Challenges for annotation technology. Several challenges to annotation exist. The most pressing of these is the so-called transcription bottleneck (Seifart et al. 2018, Himmelmann 2018), wherein "language documentation projects typically manifest a yawning gap between the amount of material recorded and archived and the amount of data that is minimally annotated (transcribed and translated), let alone more thoroughly analyzed (e.g., morphologically segmented and glossed)" (Seifart et al. 2018, p. e335). This means that linguistics is sitting on a massive body of raw data that is currently unusable for even the most basic analysis.

Human transcription has obvious limitations on speed and efficiency: Spoken language is "underdetermined by the acoustic signal" (Himmelmann 2018, p. 35), consultants may edit the original phrasing when assisting with transcription, and everyone involved can get fatigued. Basic oral language documentation (BOLD) (e.g., Reiman 2010) was piloted as an effort to overcome these challenges and to take advantage of fluent consultants now by circumventing keyboarded transcription. BOLD uses oral methods to create verbal annotations, whereby consultants carefully "respeak" recordings (as the equivalent of a transcript) and also speak translations. The result is a multitrack augmented recording containing the original recording interspersed with careful

Automatic speech recognition (ASR): the automated conversion of an audio signal into a string of textual transcription

Neural network: network of processing elements (artificial neurons) that each convert input into output toward a specific computational goal

Pronunciation lexicon: a set of grapheme-to-phoneme mappings that is specified manually for use in ASR

respeaking and interpretation. Later, the theory goes, someone can come back to keyboard the annotations when there is more time. BOLD showed some initial methodological success (Boerger 2011), although it has not had broad uptake, perhaps because it kicks the can down the road, nor has it been broadly demonstrated that BOLD eventually results in widely repurposable textual transcriptions. At the moment, the most promising technologies for speech-to-text exist in the realm of natural language processing (NLP) (discussed in Section 3.3.2).

Beyond the transcription bottleneck, other authors have discussed technological challenges to annotation pertaining to the usefulness of collections of transcripts via, for example, corpus methods and other kinds of computational processing (for discussions on integrating corpus linguistics with language documentation, see Cox 2011 and Lüdeling 2012). Orthography is one such challenge. Idiosyncratic or dialectal variations in spelling can hinder the usefulness of a collection of transcriptions by masking patterns in the data. Rytting & Yelle (2017) propose a solution based on an extant method for the Detection of Errors and Correction in Corpus Annotation (DECCA) that does not make reference to a lexicon and is therefore potentially more useful for an under-resourced language. Another orthographic challenge is the use of multiple scripts; for example, Cyrillic and Latin scripts and International Phonetic Alphabet/Americanist phonetic notation are all in use by various stakeholders in the documentation of St. Lawrence Island/Central Siberian Yupik, as described by Schwartz & Chen (2017). They developed a web-based utility to convert between these scripts that takes into account language-specific orthographic conventions for representing morphophonological processes.

3.3.2. NLP for annotation.

NLP offers many opportunities for language science, but those technologies are available only for a small fraction of the world's languages (Joshi et al. 2020). Even NLP technology that claims to be language independent is often not so on closer examination (Bender 2019). In recent years, the intersection of computational linguistics and language documentation has been growing. NLP technologies, such as those described in this section and other models such as finite-state transducers, have the potential to form a virtuous circle with language documentation, wherein each guides the other to richer language records. In this section we examine a few growing NLP technologies that can aid in the documentary workflow.

The transcription bottleneck has motivated large strides in applying automatic speech recognition (ASR) technology to smaller documentary datasets. The established ASR systems for major world languages are trained on data so big as to be out of reach for most documentation and revitalization contexts. Making useful ASR for smaller data is a computational challenge but can greatly aid in speeding up transcription by creating an imperfect first-pass transcript that can later be corrected by hand, including potentially through crowdsourcing, as described by Anastasopoulos & Chiang (2017) for Griko.

ELPIS (the CoEDL Endangered Language Pipeline and Inference System) (Foley et al. 2018) allows users to train an ASR system on a limited body of transcribed data with either the Kaldi (**https://github.com/kaldi-asr**) Open-Source toolkit or Hugging Face Transformers (**https://huggingface.co**). Phonemic transcription is also now possible in ELPIS using ESPnet (Watanabe et al. 2018), a neural network speech-processing toolkit that allows ELPIS to be used without a pronunciation lexicon or other language-specific preprocessing. Less preprocessing and an emphasis on user-friendly interfaces are intended to make ELPIS more easily accessible to language workers without computational training (Adams et al. 2021).

Persephone, a precursor to ELPIS, is another Open-Source phonemic transcription tool that uses neural networks to predict phonemes and tones directly from audio (Adams et al. 2018). Some of the data preprocessing needed to use Persephone can be automated, as described by Wisniewski et al. (2020) for materials in the open-access Pangloss Collection (Michailovsky et al. 2014), further lowering the barrier for interdisciplinary work between linguists and computer scientists.

In low-resource contexts, it can be helpful to utilize a multilingual acoustic model, where various languages are used for training. However, such models tend to ignore differences between phonemes and phones. The performance of ASR in low-resource contexts can be improved with multilingual models that incorporate language-specific phonological knowledge about which phonemes in each language correspond to which allophones (Li et al. 2020).

Forced aligners automatically narrowly match existing transcriptions of utterances to audio recordings. For a low-resource language, there are two main ways to use a forced aligner. The first option is to apply an aligner that was trained on a high-resource language such as English (this method is often referred to as untrained forced alignment). For example, DiCanio et al. (2013) describe a project to use two English-trained forced aligners, HMALIGN and the Penn Phonetics Lab Forced Aligner toolkit, to segment Yoloxóchitl Mixtec data. Coto-Solano & Solórzano (2017) compared the English-trained Forced Alignment and Vowel Extraction (FAVE)-aligner (Rosenfelder et al. 2014) with the French model in Easy Align (Goldman 2011) in their efficacy with the Bribi language (Coto-Solano & Solórzano 2017); FAVE-align had smaller error rates.

Untrained forced alignment: forced alignment using a model trained on one language to align text and audio from another language

Forced alignment: automated alignment of extant textual transcriptions to audio at the phone level

Coto-Solano et al. (2022) describe a project to develop untrained forced alignment for six underrepresented languages: Cook Islands Māori; Denggan; Bribri, Cabécar, and Malecu; and Me'phaa Vátháá. As the authors describe, "the main challenge is to transform the data into a form that can be processed by an English aligner. . .and then to undo those transformations so that the data are not 'deformed' by having to fit into an English mold" (Coto-Solano et al. 2022, p. 425). This is accomplished through glyph management, in which the user develops a pronunciation lexicon mapping the phones of the target language as closely as possible onto the glyphs used by the aligner (in this case, the Arpabet) and then transforms the output of the aligner back into the desired graphemic representation.

The second option for forced alignment of low-resourced languages is to train an aligner on data from the target language or from across a family of underresourced languages (Strunk et al. 2014). The Montreal Forced Aligner (MFA) (McAuliffe et al. 2017) has been used successfully for Matukar Panau (González et al. 2018b), Kera (Kempton & Pearce 2020), 'Ōlelo Hawai'i (Kettig 2021), and Uspanteko (Bennett et al. 2022), among other languages. MFA uses the Kaldi ASR toolkit and can be trained on non-English language data or the English model can be used. The Munich Automatic Segmentation System (MAUS) (Schiel 1999) is an Open-Source forced aligner with an untrained mode that can be helpful for noisy conversational data, as Jones et al. (2019) show for North Australian Kriol. Although studies comparing different forced aligners are limited, conclusions drawn from major world language data may be useful for low-resource language work. González et al. (2018a) compared MFA with LaBB-CAT (Fromont & Hay 2012), FAVE, and MAUS on English data, and found that MFA gave the most accurate boundaries.

3.4. Archiving

Because the preservation of language materials is a central aim of language documentation, the scope of endangered language archiving is broad and could be described here from several standpoints: that of the archive user, who wants to find information about a language of interest; that of the depositor, who wants to ensure the long-term preservation of materials in their possession; and that of the archivist, who wants to create or maintain a collection in an institutional setting such as a tribal library, museum, or university. The history and role of archives in language documentation are discussed at length by Woodbury (2014) and Henke & Berez-Kroeker (2016); Kung et al. (2020) and Andreassen (2022) provide excellent guidance for potential future depositors. Instead, we focus here on a few recent developments in archiving thinking and technology that are likely to prove transformative in the coming years.

3.4.1. The usability of archives. The discussion of the usability of endangered language archives has widened over the past half-decade. On the basis of discussions with Indigenous language workers in Washington, Alaska, and California, Shepard (2015, 2016) critiques the usability of non-Indigenously administered archives, in that they often fail to honor Native American cultural beliefs and sovereignty. "Resources in archives are often underutilized by Native communities for a variety of reasons, including lack of management controls, limited discoverability and unclear intellectual property ownership" (Shepard 2016, p. 459). He continues:

Content management system (CMS): a software system or application used to manage (store, share, describe) digital content

> A particularly challenging issue for many tribal groups, is the idea that archives "grant easy access" (Trilsbeek & Wittenburg 2006, p. 313) to cultural documentation. Since documentation may contain sensitive information, inappropriate access may complicate capacity to affirm and maintain self-determination. At its core, self-determination is simply the idea that Native communities can effectively interpret their past and have a right to make decisions about the trajectory of their present and future. Self-determination is as much an issue of practical concern for tribes, as an ideological one. (Shepard 2016, pp. 460–61)

In addition to archive policies that may erode tribal self-determination, available technologies such as content management systems (CMSs) may also be problematic in that they cannot easily accommodate Indigenous traditions for knowledge access and transfer. Shepard advocates for "value-added" language archive platforms with features intended to address cultural and political concerns of the tribe. One example of such a platform is Mukurtu CMS (**https://mukurtu.org/**) (Shepard 2014, Christen 2019).

Mukurtu, which began around 2007 as a community-led archiving project prompted by the repatriation of digital photographs to the Warumungu Aboriginal community in Central Australia, now serves a much wider audience and is an alternative to more rigidly fixed CMSs that are commonly used in libraries and museums. Crucially, Mukurtu allows communities to implement culturally informed access protocols that may run counter to intellectual property or copyright law, which may instead be based on community-specific "relationships, obligations, and ongoing intergenerational knowledge sharing based on systems and networks of reciprocity and respect recognized by community members" (Christen 2019, p. 158). Mukurtu also integrates Traditional Knowledge labels (Anderson et al. 2008), which allow archive managers to label materials, for example, as restricted to members of a certain clan, as seasonal knowledge, or as sacred (see **Figure 2**). Finally, Mukurtu also allows for flexible metadata entry so that multiple people can contribute their knowledge about a resource.

Babinski & Bowern (2021) approach usefulness of materials in archives in terms of how well they can be understood and repurposed by a user other than the original depositor. They examined 23 audio-plus-transcription collections from three archives [Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC) in Australia, Endangered Language Archive (ELAR) in the United Kingdom, and Archive of the Indigenous Languages of Latin America (AILLA) in the United States] in terms of analyzing variability in the structure and contents of corpora. They discovered that barriers to usability include missing information, some of which is not recoverable (e.g., missing or empty transcription files, missing metadata); extraneous information (e.g., multiple copies of the same files); and errors (e.g., stray nonprinting characters, Unicode conversion errors, delimiting characters in the wrong places). All of these errors, while understandable, greatly reduce the usability of the materials in the archives.

The language archiving community is just beginning to grapple with possible solutions to the usability question. Wasson et al. (2016, p. 647) survey language archives in terms of user-centered design (UCD), an interdisciplinary research field that strives to ensure "that technologies. . .are designed to meet the needs and constraints of their users." They present a typology of archives,

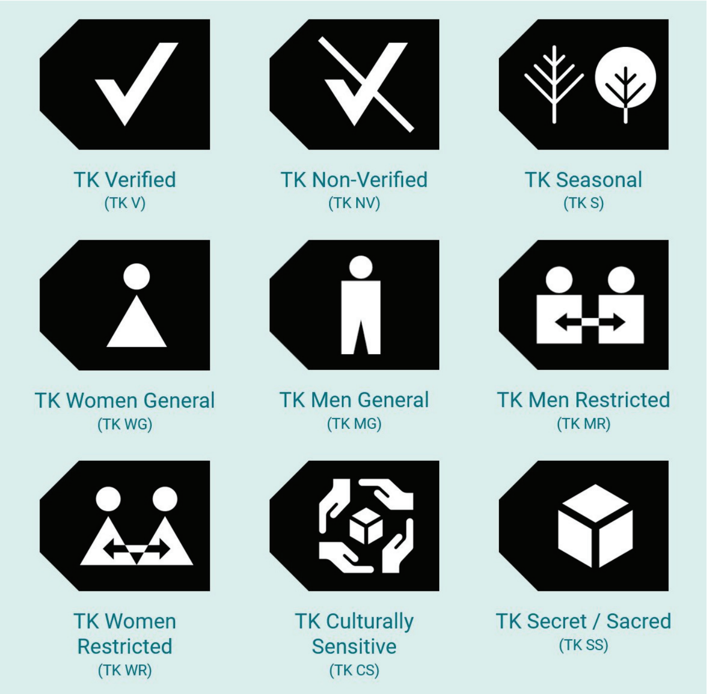

Figure 2

Traditional Knowledge (TK) labels. A screenshot of labels taken from **https://localcontexts.org/labels/traditional-knowledge-labels/** on August 26, 2022, and figure with permission from Jane Anderson, Kim Christen, Māui Hudson, and James Francis Sr.

and of stakeholder groups and their perspectives, based on a 2016 workshop at the University of North Texas. While a UCD approach to archiving is still in its infancy, the authors emphasized the continued importance of collaboration between stakeholder groups, and especially of finding ways to facilitate access to collections to facilitate novel and useful ways of visualizing the materials contained within.

3.4.2. Emerging technologies for archives. Creative application of emerging technologies can facilitate access to collections. One recent example is Glossopticon (**https://glossopticon.com/**) (Burrell et al. 2019, Hendery & Burrell 2020), a VR interface to audio recordings of language held in PARADISEC. As Hendery & Burrell note, exploring the contents of a language archive through its catalog can be daunting and unsatisfying, even when one is searching for something in particular. The immersive nature of VR can turn resource discovery into an embodied experience that "approximates. . .the environment from which the data originate" (Hendery & Burrell 2020, p. 485). PARADISEC contains approximately 9,700 hours of audio from more than 1,200 languages (Burrell et al. 2019), far too much to explore meaningfully for most users. In Glossopticon, the user experiences the materials in the collection through a three-dimensional map with pillars of light emanating from geographic points representing the location where the language is spoken. The pillars are surrounded by domes of different sizes depending on the number of speakers for each language. Users can move through the map using a VR headset (cardboard-style with smartphone, or dedicated hardware) with or without hand-gesture tracking (e.g., Leap) or via their desktop or mobile device. As users approach the language pillars, spatially dynamic snippets of audio from each language collection can be heard, and a heads-up display with collection metadata can be seen. Users can save collections of interest to return to later.

Exporting and sharing subcollections of archived materials with language communities have long been considered part of ethical documentary practice, but separating resources from the archive metadata catalogs risks shared materials becoming undiscoverable or indecipherable. Protocols such as the Oxford Common File Layout (OCFL; **https://ocfl.io/**) and Research Object Crate (RO-Crate; **https://www.researchobject.org/ro-crate/**) are gaining traction with language archives as a way to maintain file structure and metadata (Trilsbeek 2021, Thieberger & La Rosa 2021). OCFL is a protocol that specifies an expected structure for files and folders that is transparent, friendly to versioning, and application independent. RO-Crate is a protocol for packaging lightweight metadata alongside research data—as opposed to keeping metadata in a separate database—using JSON for Linking Data (**https://json-ld.org/**). In this way, metadata integrity can be maintained when, for example, one wants to share a subcollection via a lightweight computer such as Raspberry Pi (**https://www.raspberrypi.org/**) or a portable drive (e.g., Thieberger & La Rosa 2021; ARC Centre of Excellence for the Dynamics of Language 2021a,b).

4. MOBILIZING RESOURCES

Beyond archiving, once documentary materials have been created, those materials usually need to be mobilized for purposes such as revitalization, reclamation, and research. In this section we present recent advances in three subareas in the mobilization of documentary materials: accessibility of transcripts and corpora, applications of geospatial technology, and the creation of media such as books and animated films.

4.1. Mobilizing Transcripts

A number of tools are available to mobilize XML transcripts in the form of online corpora or display of interlinearized glossed text (IGT). Here, we cover some of the more recent tools in this sphere [for a longer history (e.g., TROVA, ANNEX, EOPAS, Pangloss), see Kaufman & Finkel 2018]. Kratylos (Kaufman & Finkel 2018) transforms IGT formatted in XML from software such as FLEx and ELAN and makes it viewable and searchable online while making no alterations to the source database. Its features include automatic orthography conversion, simultaneous search across multiple corpora, a comments and corrections feature, and data export with proper citations. Kratylos also supports searches for complex regular expressions, making IGT collected by field linguists more usable for theoretical linguistics research. Kratylos greatly facilitates IGT sharing among academic linguists, though it is not targeted specifically for use by speech communities.

Kwaras (Caballero et al. 2019) is a tool for managing ELAN corpora and linking them with WAV files. It creates a searchable interface that integrates transcriptions, annotations, audio clips, and metadata. Kwaras also generates citations for individual pieces of data and hyperlinks to provided audio files, which can facilitate data transparency in publications. Users can customize the structure of the Kwaras display by choosing which ELAN tiers to import, although the program does not preserve tier hierarchy. This customization allows Kwaras to be used for a variety of audiences; for example, a user could choose to display only fields relevant to practical use by native speakers. Kwaras can be used both on- and offline, further increasing its usability in the field.

Namuti (Caballero et al. 2019) is a sister tool to Kwaras that is specifically focused on community members' needs. Like Kwaras, Namuti links audio files, annotations, and metadata. It retains Kwaras's search capabilities and creation of unique identifiers for individual pieces of data. Namuti adds several different ways to visualize data for different purposes, such as a story view that displays texts with only audio, practical orthography, and translation. It lists texts by title and contributing author, the information most likely to be of use for community members.

Shifting from online corpora to tools for simple display, which continue in the tradition of the now-outdated CuPED (Berez & Cox 2009), the IATH ELAN Text-Sync Tool (ETST) (Dobrin & Ross 2017) plays ELAN files and their corresponding audiovisual media in a format more accessible to nonacademic audiences. It distinguishes speakers in a transcript by their initials and color-coding and highlights lines of the transcript that correspond to the audio as it plays. The tool can be used on- or offline and files play in a web browser. Coded in HTML, ETST is designed to function on any type of device, including smartphones and tablets. To run ETST, a user downloads three small files from the ETST website that create a playable interface when combined with an ELAN file and associated media formatted correctly (ETST provides a template for this format). When stored online, this derived file can be generated when a user clicks on it.

Augmented reality (AR): an interactive experience in which digital images are overlaid onto images from the real world creating a composite image

LingView (Pride et al. 2020), another display tool, takes input from both ELAN and FLEx. The tool's data pipeline merges ELAN and FLEx files into a JSON format and then converts them to HTML so they can be viewed together, along with associated audio and video, in a web interface. LingView prioritizes customizability in its design, such that end users can choose which fields to display at a given time, while the rich range of data in ELAN and FLEx is underlyingly still there. This makes LingView well suited for use by individuals in speech communities, as well as academic audiences, all with differing needs.

4.2. Applications of Geospatial Technology

The easy application of global positioning system (GPS) and geographic information system (GIS) technology to language documentation has led to some recent creative applications of spatial experiences of language. Larsson et al. (2021) describe three case studies integrating geospatial data into ELAN transcriptions of video-recorded linguistic events. By fitting language consultants with chest-strapped cameras equipped with GPS receivers, researchers can import the latitude and longitude coordinates into an ELAN tier, which synchronizes the audiovisual media and transcription with the locations where the recording took place. In one case study, the authors asked Eastern Penan speakers to walk through the Kelabit Highlands of Malaysia while discussing the places they passed. Exporting the geodata into KML format allows for easy visualization in Google Earth, where the paths they traversed can be viewed along with a point-of-view video of the location from the chest-strapped cameras, an audible recording of the conversation, and a transcript. Other case studies described by the authors include placenames documentation and visualization of locomotive verbs in Jahai.

Possemato et al. (2021) use GPS and GIS technology to analyze locational pointing gestures in the discourse of speakers of four Australian languages. The method involves using two cameras and precise GPS data about the location of the recording event and the positioning of the cameras. The vectors of pointing gestures made by the participants can be extrapolated and then plotted in visualization software such as Google Earth. The vectors can then be compared with ideal vectors between the speaker and the intended targets—sometimes thousands of kilometers away—and analyzed for accuracy. In one example, a Murrinhpatha speaker points toward a location called Thuykem several times, some 16 kilometers away, and is shown to be within 10 degrees of accuracy every time.

On a smaller scale, augmented reality (AR) and VR allow for other creative spatial explorations of language. Kelly & Cowell (forthcoming) describe a method they call virtual reality elicitation, in which Northern Arapaho elder speakers can immerse themselves in VR environments of, for example, the interior of an Arapaho tipi and sites in the traditional Arapaho homelands. The participants are then encouraged by other Arapaho elders to talk about what they see. Importantly, Kelly & Cowell show that samples of language collected through VR elicitation are comparable to samples collected using more traditional methods and in some ways are as structurally and cognitively rich as Arapaho narrative genres. Similarly, Running Wolf & Running Wolf (2017) use both AR and VR to enhance the user experience of Madison Buffalo Jump, a limestone cliff in Montana that was used for centuries by Indigenous peoples for harvesting bison. Trail markers in the state park allow users to access AR narratives, oral histories, and information about plants and animals. The VR experience (**http://runningwolf.io/entrance**) allows users who are not in the park to virtually visit the site.

4.3. Media Creation

Media such as books, games, and film can be major tools for language reclamation and revitalization. However, such media are rarely published on a large scale in minoritized and endangered languages. Bloom software (Hatton & Hatton 2019) allows the easy creation of books in underresourced languages. It is free for language communities and does not require an Internet connection to use. Users can create books from scratch or translate premade shell books to quickly generate a library of titles in a given language. Creators can easily insert images and multilingual text. Further, users can create audiobooks in Bloom by recording one sentence at a time, and viewers can follow along with highlighted text when listening. Bloom formats books with front and back matter, with the user filling in appropriate fields. A dialogue box guides users in establishing copyright.

Other projects mobilize existing media creation technologies for reclamation and revitalization. A high-profile example is the 2014 video game *Kisima Inŋitchuŋa*, or *Never Alone*, a collaboration between Alaska Native community members and knowledge holders and professional game developers (E-Line Media 2014). Based on a traditional Iñupiat story, the game's narration and dialogue are in spoken Iñupiaq. While *Kisima Inŋitchuŋa* was widely distributed and received international media attention, smaller-scale video game projects also have significant potential in revitalization and reclamation. For example, *Edànì Nǫgèe Dǫne Gok'eı̨dì*, or *How Fox Saved the People*, a game in Tłı̨cho that launched in 2018, was developed by self-taught volunteers using off-the-shelf resources on a small budget (West et al. 2019). Games such as this one can function as motivational teaching tools and build language skills through repetition, with in-game goals reliant on mastery of language skills, while incorporating cultural content.

Animation is another way to bring traditional stories to life as learning tools. Silva (2016) describes a project to animate traditional narratives in Desano. Yarbrough & Solomon (2019) describe the use of animated short films of Hawaiian traditional stories in placed-based pedagogy for immersion schools on O'ahu.

5. LOOKING AHEAD: THE PANDEMIC AND BEYOND

Since we were first invited to write this article in March 2020, methods for remote language documentation have accelerated quickly, catalyzed by the COVID-19 pandemic, which is still ongoing as we submit the present article in March 2022. The pandemic quickly focused language documentation in ways that may end up enduring for quite some time. The past two years have seen language workers worldwide engaging more with technologies that are not necessarily new but that are older and used in novel ways.

Remote data collection was not new during the pandemic, but it became much more widespread and necessary with COVID-19. Pandey (2021) discusses the convenience of using mobile phones for linguistic interviews, although potential problems include connectivity issues and background noise. Remote collaboration over the Internet has also become more common but can get complicated when collaborators can only access the Internet via shared computers in Internet cafes. Rice & Dagua Toquetón (2021) discuss a remote collaborative workflow using YouTube rather than ELAN for transcription and translation, which is well suited for when collaborators can only access public computers. Linguistics in the Pub, an informal gathering of language workers in the Melbourne area, held two virtual panel discussions in May and June 2020 (Linguistics in the Pub 2020a,b) to discuss nascent challenges, such as privacy concerns when recording over videoconferencing services such as Zoom, and opportunities, such as time to work with previously collected data or for complex projects such as dictionaries, brought about by the pandemic.

During this global crisis, linguists also became involved in the dissemination of accurate COVID-19 information in underresourced languages. VirALLanguages (**http://virallanguages.org/**) is a collaborative project between linguists, health specialists, and community members to distribute reliable COVID-19 information in 45 languages of Cameroon, Indonesia, and Pakistan in culturally appropriate ways, utilizing social media to circulate informational videos (Di Carlo et al. 2022). Anastasopoulos et al. (2020) discuss the Translation Initiative for COvid-19 (TICO-19), in which machine learning was used to translate COVID-19 information into 26 underresourced languages selected according to the priorities of Translators without Borders.

The reliance on remote language work that arose during the pandemic is not likely to disappear once the pandemic is over. Williams et al. (2021) argue that some of these changes should become permanent. The advantages of shifting to a model with increased virtual interaction between a language community and outsider linguists include considerable financial savings and a decentering of the researcher as the principal investigator. "Ultimately, this model of research could serve to flip the script from 'linguists' working with 'language consultants' to the language community working with a 'linguist consultant', more on their own terms" (Williams et al. 2021, p. 362). The authors advocate for tools for team communication and data sharing that are free and widely used around the world (e.g., WhatsApp and Google Drive), although they acknowledge that language-documentation-specific apps do have a place in the landscape of remote work.

We conclude this article echoing some of the same sentiments that Williams et al. (2021) share. While remote work will never replace the relationship-building that comes with face-to-face interaction (e.g., Leonard 2021), the pandemic has jolted all of us out of previous routines and into new ways of thinking about the tools we use, which in many arenas has turned out to be an improvement on older ways of working. "Although we have used some of the tools and software for being in contact with our collaborators, many of us only gave serious consideration to remote fieldwork when forced to by a global pandemic, while we ought to have been thinking outside the box in this way all along" (Williams et al. 2021, p. 360). Indeed, thinking outside the box has been an ethos in language documentation for the past three decades, since Hale et al. (1992) directed attention to the crisis of language endangerment, giving impetus to the development of a new subfield of linguistics. The technologies that have since arisen and continue to develop will see increased emphasis on equalizing participation by all parties, taking fuller advantage of computational methods being developed for large data, and increased mobilization of resources for purposes already known and for those only imagined today.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

We are grateful to Bradley McDonnell and Peter Austin for their feedback on and assistance with identifying topics to include in this article. Errors of omission and commission belong to the authors alone.

LITERATURE CITED

Adams O, Cohn T, Neubig G, Cruz H, Bird S, Michaud A. 2018. Evaluating phonemic transcription of low-resource tonal languages for language documentation. In *Proceedings of the 11th International Conference on Language Resources and Evaluation*, pp. 3356–65. Paris: Eur. Lang. Resour. Assoc.

Adams O, Galliot B, Wisniewski G, Lambourne N, Foley B, et al. 2021. User-friendly automatic transcription of low-resource languages: plugging ESPnet into Elpis. arXiv:2101.03027v2 [cs.CL]

Anastasopoulos A, Cattelan A, Dou Z-Y, Federico M, Federman C, et al. 2020. TICO-19: the Translation Initiative for COvid-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, pp. 1–25. Stroudsburg, PA: Assoc. Comput. Linguist.

Anastasopoulos A, Chiang D. 2017. A case study on using speech-to-translation alignments for language documentation. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pp. 170–78. Honolulu: Assoc. Comput. Linguist.

Anderson J, Christen K, Hudson M, Francis J Sr. 2008. TK labels. *Local Contexts*. **https://localcontexts.org/labels/traditional-knowledge-labels/**

Andreassen HN. 2022. Archiving research data. See Berez-Kroeker et al. 2022, pp. 89–100

ARC Centre of Excellence for the Dynamics of Language. 2021a. Data loader. **https://language-archives.services/about/data-loader/**

ARC Centre of Excellence for the Dynamics of Language. 2021b. Raspberry Pi as a repatriation device. **https://language-archives.services/about/pi/**

Babinski S, Bowern C. 2021. *Contemporary digital linguistics and the archive: an urgent review*. Paper presented at the 7th International Conference on Language Documentation & Conservation, Honolulu, Mar. 4–7

Beier C, Michael L. 2022. Managing lexicography data: a practical, principled approach using FLEx (Fieldworks Language Explorer). See Berez-Kroeker et al. 2022, pp. 301–14

Bender EM. 2019. *English isn't generic for language, despite what NLP papers might lead you to believe*. Paper presented at the Symposium on Data Science & Statistics, Bellevue, WA, May 30. **http://faculty.washington.edu/ebender/papers/Bender-SDSS-2019.pdf**

Bennett R, Henderson R, Harvey M. 2022. Tonal variability and marginal contrast: lexical pitch accent in Uspanteko. In *Prosody and Prosodic Interfaces*, ed. H Kobuzono, J Ito, A Mester, pp. 589–645. Oxford, UK: Oxford Univ. Press

Berez A, Cox C. 2009. *CuPED - Software demonstration*. Paper presented at the 1st International Conference on Language Documentation and Conservation, Honolulu, Mar. 12–14

Berez AL. 2007. Review of EUDICO Linguistic Annotator (ELAN). *Lang. Doc. Conserv.* 1(2):283–89

Berez-Kroeker AL, Gawne L, Kung SS, Kelly BF, Heston T, et al. 2018. Reproducible research in linguistics: a position statement on data citation and attribution in our field. *Linguistics* 56(1):1–18

Berez-Kroeker AL, McDonnell B, Koller E, Collister LB. 2022. *The Open Handbook of Linguistic Data Management*. Cambridge, MA: MIT Press

Bird S, Simons G. 2003. Seven dimensions of portability for language documentation and description. *Language* 79(3):557–82

Boerger BH. 2011. To BOLDly go where no one has gone before. *Lang. Doc. Conserv.* 5:208–33

Boersma P, Weenink D. 2022. Praat: doing phonetics by computer [computer program]. Version 6.2.09. **http://www.praat.org/**

Bowern C. 2010. Fieldwork and the IRB: a snapshot. *Language* 86(4):897–905

Bowern C. 2015. *Linguistic Fieldwork: A Practical Guide*. London: Palgrave Macmillan. 2nd ed.

Boyton J, Moran S, Aristar A, Aristar-Dry H. 2006. E-MELD and the School of Best Practices: an ongoing community effort. In *Sustainable Data from Digital Fieldwork*, ed. L Barwick, N Thieberger, pp. 87–97. Sydney: Univ. Sydney Press

Brinklow NT, Littell P, Lothian D, Pine A, Souter H. 2019. Indigenous language technologies and language reclamation in Canada. In *Language Technology for All (LT4All): Enabling Language Diversity and Multilingualism Worldwide*, ed. European Language Resources Association, pp. 402–6. Paris: UNESCO

Burrell A, Hendery R, Thieberger N. 2019. Glossopticon: visualising archival data. In *Proceedings of the 2019 23rd International Conference in Information Visualization - Part II IV-2 2019*, pp. 100–3. New York: IEEE

Caballero G, Carroll L, Mach K. 2019. Accessing, managing, and mobilizing an ELAN-based language documentation corpus: the Kwaras and Namuti tools. *Lang. Doc. Conserv.* 13:63–82

Christen K. 2019. "The songline is alive in Mukurtu": return, reuse, and respect. In *Language Documentation & Conservation Special Publication 18*, *Archival Returns: Central Australia and Beyond*, ed. L Barwick, J Green, P Vaarzon-Morel, pp. 153–72. Honolulu: Univ. Hawai'i Press

Coto-Solano R, Nicholas SA, Hoback B, Cano GT. 2022. Managing data workflows for untrained forced alignment: examples from Costa Rica, Mexico, the Cook Islands, and Vanuatu. See Berez-Kroeker et al. 2022, pp. 423–36

Coto-Solano R, Solórzano SF. 2017. Comparison of two forced alignment systems for aligning Bribri speech. *CLEI Electron. J.* 20(1):13

Cox C. 2011. Corpus linguistics and language documentation: challenges for collaboration. In *Corpus-Based Studies in Language Use, Language Learning, and Language Documentation*, ed. J Newman, H Baayen, S Rice, pp. 239–64. Leiden: Brill

Cox C. 2022. Managing data in a language documentation corpus. See Berez-Kroeker et al. 2022, pp. 277–86

Debenport E. 2010. The potential complexity of "universal ownership": cultural property, textual circulation, and linguistic fieldwork. *Lang. Commun.* 30(3):204–10

Di Carlo P, McDonnell B, Vahapoglu L, Good J, Seyfeddinipur M, Kordas K. 2022. Public health information for minority linguistic communities. *Bull World Health Organ*. 100(1):78–80

DiCanio C, Nam H, Whalen DH, Bunnell HT, Amith JD, García RC. 2013. Using automatic alignment to analyze endangered language data: testing the viability of untrained alignment. *J. Acoust. Soc. Am.* 134(3):2235–46

Dobrin LM. 2021. The Arapesh "suitcase miracle": the interpretive value of reproducible research. *Lang. Doc. Descr.* 21:37–69

Dobrin LM, Ross D. 2017. The IATH ELAN text-sync tool: a simple system for mobilizing ELAN transcripts on- or off-line. *Lang. Doc. Conserv.* 11:94–102

E-Line Media. 2014. *Never Alone*, *Kisima Iŋŋitchuŋa* [game]. 2014. **https://elinemedia.com/never-alone/**

First People's Cultural Council. 2020. *Check before you tech!* Brentwood Bay, Can.: First People's Cultural Council. **https://fpcc.ca/wp-content/uploads/2020/09/FPCC-Check-Before-You-Tech.pdf**

Foley B, Arnold J, Coto-Solano R, Durantin G, Ellison TM, et al. 2018. Building speech recognition systems for language documentation: the CoEDL Endangered Language Pipeline and Inference System (ELPIS). In *Proceedings of the 6th Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU 2018)*, pp. 200–4. Grenoble, France: Int. Speech. Commun. Assoc.

Fromont R, Hay J. 2012. LaBB-CAT: an annotation store. In *Proceedings of the Australasian Language Technology Association Workshop 2012*, pp. 113–17. Stroudsburg, PA: Assoc. Comput. Linguist.

Gaby A, Woods L. 2020. Toward linguistic justice for Indigenous people: a response to Charity Hudley, Mallinson, and Bucholtz. *Language* 96(4):e268–80

Goldman JP. 2011. *EasyAlign: an automatic phonetic alignment tool under Praat*. Paper presented at Interspeech, Florence, Italy, Aug. 28–31

González S, Grama J, Travis C. 2018a. *Forced-alignment comparison for linguistics.* Poster presented at CoEDL Fest, Melbourne, Feb. 8. **https://cloudstor.aarnet.edu.au/plus/s/gyC6vuX5uvc5soG**

González S, Travis CE, Grama J, Barth D, Ananthanarayan S. 2018b. Recursive forced alignment: a test on a minority language. In *Proceedings of the 17th Australasian International Conference on Speech Science and Technology*, pp. 145–48. Canberra, Aust.: Australas. Speech Sci. Technol. Assoc.

Good J. 2018. Ethics in language documentation and revitalization. In *The Oxford Handbook of Endangered Languages*, ed. K Rehg, L Campbell, pp. 419–42. Oxford, UK: Oxford Univ. Press

Hale K, Krauss M, Watahomigie LJ, Yamamoto AY, Craig C, et al. 1992. Endangered languages. *Language* 68(1):1–42

Hatton J, Hatton S. 2019. Bloom: now communities can create their own books. Paper presented at the *6th International Conference on Language Documentation & Conservation*, Honolulu, Feb. 28–Mar. 3. **http://scholarspace.manoa.hawaii.edu/handle/10125/44889**

Hatton J, Holton G, Seyfeddinipur M, Thieberger N. 2021. Lameta [software]. **https://sites.google.com/site/metadatatooldiscussion/home**

Hendery R, Burrell A. 2020. Playful interfaces to the archive and the embodied experience of data. *J. Doc.* 76(2):484–501

Henke RE, Berez-Kroeker AL. 2016. A brief history of archiving in language documentation, with an annotated bibliography. *Lang. Doc. Conserv.* 10:411–57

Hill JH. 2002. "Expert rhetorics" in advocacy for endangered languages: Who is listening, and what do they hear? *J. Linguist. Anthropol.* 12(2):119–33

Himmelmann NP. 1998. Documentary and descriptive linguistics. *Linguistics* 36:161–95

Himmelmann NP. 2006. Language documentation: What is it and what is it good for? In *Essentials of Language Documentation*, ed. J Gippert, NP Himmelmann, U Mosel, pp. 1–30. Berlin: De Gruyter Mouton

Himmelmann NP. 2012. Linguistic data types and the interface between language documentation and description. *Lang. Doc. Conserv.* 6:187–207

Himmelmann NP. 2018. Meeting the transcription challenge. In *Language Documentation & Conservation Special Publication 15*, *Reflections on Language Documentation: 20 Years after Himmelmann 1998*, ed. B McDonnell, AL Berez-Kroeker, G Holton, pp. 33–40. Honolulu: Univ. Hawaiʻi Press

Holton G, Leonard WY, Pulsifer PL. 2022. Indigenous peoples, ethics, and linguistic data. See Berez-Kroeker et al. 2022, pp. 49–60

ISO (Int. Organ. Stand.). 2007. *ISO 639-3:2007. Codes for the representation of names of languages—Part 3: Alpha-3 code for comprehensive coverage of languages*. Geneva: ISO. **https://www.iso.org/standard/39534.html**

Jones C, Li W, Almeida A, German A. 2019. Evaluating cross-linguistic forced alignment of conversational data in north Australian Kriol, an under-resourced language. *Lang. Doc. Conserv.* 13:281–99

Joshi P, Santy S, Budhiraja A, Bali K, Choudhury M. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6282–93. Stroudsburg, PA: Assoc. Comput. Linguist.

Kaufman D, Finkel R. 2018. Kratylos: a tool for sharing interlinearized and lexical data in diverse formats. *Lang. Doc. Conserv.* 12:124–46

Kelly P, Cowell A. Forthcoming. Northern Arapaho virtual reality linguistic elicitation. *Lang. Doc. Descr.* 22

Kempton T, Pearce M. 2020. Corpus phonetics for under-documented languages: a vowel harmony example. In *Proceedings of the 2019 Annual Meeting on Phonology*. Amherst, MA: Annu. Meet. Phonol. **https://doi.org/10.3765/amp.v8i0.4682**

Kettig T. 2021. *Haʻina ʻia mai ana ka puana: the vowels of ʻŌlelo Hawaiʻi*. PhD Thesis, Univ. Hawaiʻi, Mānoa

Koenig A, Ringersma J, Trilsbeek P. 2009. The Language Archiving Technology domain. In *Human Language Technology as a Challenge for Computer Science and Linguistics*, ed. Z Vetulani, pp. 295–99. Nijmegen, Neth.: Max Planck Inst. Psycholinguist.

Kung SS, Sullivant R, Pojman E, Niwagaba A. 2020. *Archiving for the Future: Simple Steps for Archiving Language Documentation Collections*. **https://archivingforthefuture.teachable.com**

Larsson J, Burenhult N, Kruspe N, Purves RS, Rothstein M, Sercombe P. 2021. Integrating behavioral and geospatial data on the timeline: towards new dimensions of analysis. *Int. J. Soc. Res. Methodol.* 24(1):1–13

Lee NH. 2022. Managing data for writing a reference grammar. See Berez-Kroeker et al. 2022, pp. 287–300

Leonard WY. 2017. Producing language reclamation by decolonizing 'language.' *Lang. Doc. Descr.* 14:15–36

Leonard WY. 2021. *Language reclamation through relational language work*. Plenary address presented at the 7th International Conference on Language Documentation & Conservation, Honolulu, Mar. 4–7. **https://scholarspace.manoa.hawaii.edu/items/68e60bf6-f82e-4f59-83e5-444275f5cfdf**

Li X, Dalmia S, Li J, Lee M, Littell P, Yao J, et al. 2020. Universal phone recognition with a multilingual allophone system. arXiv:2002.11800 [cs.CL]

Linguistics in the Pub. 2020a. *Fieldwork at a distance*. Online panel discussion, May 20. **https://www.youtube.com/watch?v=K89TKXaJNjU**

Linguistics in the Pub. 2020b. *Innovating and adapting to Covid-19 restrictions*. Online panel discussion, June 23. **https://www.youtube.com/watch?v=A-E_hxcdaoQ**

Lüdeling A. 2012. A corpus linguistics perspective on language documentation, data, and the challenge of small corpora. In *Language Documentation & Conservation Special Publication 3: Potentials of Language Documentation: Methods, Analyses, and Utilization*, ed. F Siefart, G Haig, NP Himmelmann, D Jung, A Margetts, P Trillsbeek, pp. 32–38. Honolulu: Univ. Hawai'i Press

McAuliffe M, Socolof M, Mihuc S, Wagner M, Sonderegger M. 2017. *Montreal forced aligner: trainable text-speech alignment using Kaldi*. Paper presented at Interspeech 2017, pp. 498–502. Stockholm, Aug. 20–24. **https://www.isca-speech.org/archive/interspeech_2017/mcauliffe17_interspeech.html**

Meakins F, Green J, Turpin M. 2018. *Understanding Linguistic Fieldwork*. London: Routledge

Michailovsky B, Mazaudon M, Michaud A, Guillaume S, François A, Adamou E. 2014. Documenting and researching endangered languages: the Pangloss Collection. *Lang. Doc. Conserv.* 8:119–35

Moeller SR. 2014. Review of SayMore, a tool for Language Documentation Productivity. *Lang. Doc. Conserv.* 8:66–74

Moran S, Cysouw M. 2018. *The Unicode Cookbook for Linguists: Managing Writing Systems Using Orthography Profiles*. Berlin: Lang. Sci.

Nathan D. 2010. Sound and unsound practices in documentary linguistics: towards an epistemology for audio. *Lang. Doc. Descr.* 7:262–84

O'Meara C, Good J. 2010. Ethical issues in legacy language resources. *Lang. Commun.* 30:162–70

Pandey A. 2021. *Fieldwork via mobile interviewing during Covid-19: a case study of Kanauji.* Paper presented at the 7th International Conference on Language Documentation & Conservation, Honolulu, Mar. 4–7. **https://scholarspace.manoa.hawaii.edu/items/a7e7e732-7906-40a7-8e15-a277961f751e**

Pentangelo J. 2020. *360° video and language documentation: towards a corpus of Kanien'kéha (Mohawk)*. PhD Thesis, The Graduate Center, City University of New York

Possemato F, Blythe J, de Dear C, Dahmen J, Gardner R, Stirling L. 2021. Using a geospatial approach to document and analyse locational points in face-to-face conversation. *Lang. Doc. Descr.* 20:313–51

Pride K, Tomlin N, AnderBois S. 2020. LingView: a web interface for viewing FLEx and ELAN files. *Lang. Doc. Conserv.* 14:87–107

Reiman DW. 2010. Basic oral language documentation. *Lang. Doc. Conserv.* 4:254–68

Rice A, Dagua Toquetón B. 2021. Collaborative corpus work at a distance: building a remote workflow around YouTube. Paper presented at the *7th International Conference on Language Documentation & Conservation*, Honolulu, Mar. 4–7

Robinson LC. 2010. Informed consent among analog people in a digital world. *Lang. Commun.* 30(3):186–91

Rosenfelder I, Fruehwald J, Evanini K, Seyfarth S, Gorman K, et al. 2014. *FAVE (Forced Alignment and Vowel Extraction) Program Suite v1.2.2*. **https://doi.org/10.5281/zenodo.9846**

Running Wolf M, Running Wolf C. 2017. *Reigniting the many voices of a communal bison hunt in virtual reality*. Paper presented at the 4th International Conference on Language Documentation & Conservation, Honolulu, Feb. 26–Mar. 1. **http://scholarspace.manoa.hawaii.edu/handle/10125/42023**

Rytting CA, Yelle J. 2017. DECCA repurposed: detecting transcription inconsistencies without an orthographic standard. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pp. 116–21. Honolulu: Assoc. Comput. Linguist.

Schiel F. 1999. Automatic phonetic transcription of non-prompted speech. In *Proceedings of the ICPhS*, pp. 607–10. London: Int. Phon. Assoc. **https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS1999/papers/p14_0607.pdf**

Schwartz L, Chen E. 2017. Liinnaqumalghiit: a web-based tool for addressing orthographic transparency in St. Lawrence Island/Central Siberian Yupik. *Lang. Doc. Conserv.* 11:275–88

Seifart F, Evans N, Hammarström H, Levinson SC. 2018. Language documentation twenty-five years on. *Language* 94(4):e324–45

Seyfeddinipur M, Ameka F, Bolton L, Blumtritt J, Carpenter B, et al. 2019. Public access to research data in language documentation: challenges and possible strategies. *Lang. Doc. Conserv.* 13:545–63

Seyfeddinipur M, Rau F. 2020. Keeping it real: video data in language documentation and language archiving. *Lang. Doc. Conserv.* 14:503–19

Shepard M. 2014. Review of Mukurtu content management system. *Lang. Doc. Conserv.* 8:315–25

Shepard MA. 2015. "*The substance of self-determination:*" *language, culture, archives and sovereignty*. PhD Thesis, Univ. Br. Columbia

Shepard MA. 2016. The value-added language archive: increasing cultural compatibility for Native American communities. *Lang. Doc. Conserv.* 10:458–79

SIL International. 2022. FieldWorks Language Explorer 9.0 [software]. **https://software.sil.org/fieldworks/**

Silva W. 2016. Animating traditional Amazonian storytelling: new methods and lessons from the field. *Lang. Doc. Conserv.* 10:480–96

Simons G, Bird S. 2003. The open language archives community: an infrastructure for distributed archiving of language resources. *Lit. Linguist. Comp.* 18(2):117–28

Sloetjes H, Wittenburg P. 2008. Annotation by category - ELAN and ISO DCR. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, pp. 816–20. Paris: Eur. Lang. Resour. Assoc.

Strunk J, Schiel F, Seifart F. 2014. Untrained forced alignment of transcriptions and audio for language documentation corpora using WebMAUS. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*, pp. 3940–47. Paris: Eur. Lang. Resour. Assoc.

Surma A, Truong CL. Forthcoming. Digital tools for language revitalization. In *Handbook of Languages and Linguistics of North America*, ed. C Jany, K Rice, M Mithun. Berlin: De Gruyter Mouton

TLA (The Lang. Arch.). 2020. Arbil information, manuals & download. *TLA Forums*. **https://archive.mpi.nl/forums/t/arbil-information-manuals-download/1045**

Thieberger N. 2016. Results of the metadata survey. *Endangered Languages and Cultures Blog, June 6*. **https://www.paradisec.org.au/blog/2016/06/results-of-the-metadata-survey/**

Thieberger N, Berez AL. 2011. Linguistic data management. In *The Oxford Handbook of Linguistic Fieldwork*, ed. N Thieberger, pp. 90–118. Oxford, UK: Oxford Univ. Press

Thieberger N, La Rosa M. 2021. *Collection data management and repatriation of archival materials back to their source communities*. Paper presented at the 7th International Conference on Language Documentation & Conservation, Honolulu, Mar. 4–7

Trilsbeek P. 2021. *Potential directions in the data repository landscape*. Paper presented at the PARADISEC@100 conference, Melbourne, Feb. 17–19

Trilsbeek P, Wittenburg P. 2006. Archiving challenges. In *Essentials of Language Documentation*, ed. J Gippert, N Himmelmann, U Mosel, pp. 311–62. New York: De Gruyter Mouton

Warner N. 2014. Sharing of data as it relates to human subjects issues and data management plans. *Lang. Linguist. Compass.* 8:512–18

Wasson C, Holton G, Roth HS. 2016. Bringing user-centered design to the field of language archives. *Lang. Doc. Conserv.* 10:641–81

Watanabe S, Hori T, Karita S, Hayashi T, Nishitoba J, et al. 2018. ESPnet: End-to-End Speech Processing Toolkit. arXiv:1804.00015 [cs.CL]

West L, Hucklebridge S, Mantla R, Lafferty L, Steinwand T, Welch N. 2019. *Creating video games for language revitalization and pedagogy*. Paper presented at the 6th International Conference on Language Documentation & Conservation, Honolulu, Feb. 28–Mar. 3. **http://scholarspace.manoa.hawaii.edu/handle/10125/44812**

Wilkinson M, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3:160018

Williams N, Silva WDL, McPherson L, Good J. 2021. COVID-19 and documentary linguistics: some ways forward. *Lang. Doc. Descr.* 20:359–77

Wisniewski G, Michaud A, Guillaume S. 2020. Phonemic transcription of low-resource languages: To what extent can preprocessing be automated? In *Proceedings of the 1st Joint SLTU (Spoken Language Technologies for Under-resourced languages) and CCURL (Collaboration and Computing for Under-Resourced Languages) Workshop, 2020*, Marseille, France, pp. 306–15. Paris: Eur. Lang. Resour. Assoc.

Woodbury AC. 2011. Language documentation. In *The Cambridge Handbook of Endangered Languages*, ed. PK Austin, J Sallbank, pp. 159–86. Cambridge, UK: Cambridge Univ. Press

Woodbury AC. 2014. Archives and audiences: toward making endangered language documentations people can read, use, understand, and admire. *Lang. Doc. Descr.* 12:19–36

Yarbrough D, Solomon NH. 2019. Animated Haʻi Moʻolelo ʻŌiwi: place-based pedagogy in practice. In *Proceedings of the 23rd Annual Graduate Student Conference of the College of LLL*, ed. N Handley, RL Hughes, pp. 104–10. Honolulu: Univ. Hawai'i Press