2022. *Proc Ling Soc Amer* 7(1). 5219. https://doi.org/10.3765/plsa.v7i1.55219.

How usable are digital collections for endangered languages? A review

Sarah Babinski, Jeremiah Jewell, Kassandra Haakman, Juhyae Kim, Amelia Lake, Irene Yi & Claire Bowern*

Abstract. Here, we report on pilot research on the extent to which language collections in digital linguistic archives are discoverable, accessible, and usable for linguistic research. Using a test case of common tasks in phonetic and phonological documentation, we evaluate a small random sample of collections and find substantial, striking problems in all domains. Of the original 20 collections, only six had digitized audio files with associated transcripts (preferably phrase-aligned). That is, only 30% of the collections in our sample were even potentially suitable for any type of phonetic work (regardless of quality of recording). Information about the contents of the collection was usually discoverable, though there was variation in the types of information that could be easily searched for in the collection. Though eventually three collections were aligned, only one collection was successfully force-aligned from the archival materials without substantial intervention. We close with recommendations for archive depositors to facilitate discoverability, accessibility, and functionality of language collections. Consistency and accuracy in file naming practices, data descriptions, and transcription practices is imperative. Providing a collection guide also helps. Including useful search terms about collection contents makes the materials more findable. Researchers need to be aware of the changes to collection structure that may result from archival uploads. Depositors need to consider how their metadata is included in collections and how items in the collection may be matched to each other and to metadata categories. Finally, if our random sample is indicative, linguistic documentation practices for future phonetic work need to change rapidly, if such work from archival collections is to be done in future.

Keywords. language reclamation; documentation; digital language archives; archival collections; usability and reusability; workflow; forced alignment

1. Introduction. Documentation of endangered languages sits at the nexus of tangible and intangible cultural heritage. Digital documentation is now pervasive, and has allowed the collection of larger and more varied corpora than could have been previously anticipated. Many older collections have also been converted to digital formats. This same revolution has affected research. Linguists can now run on their laptops – in the field – analyses that would have required rooms of specialized equipment 20 or 30 years ago. It is now possible to work with linguistic data in ways that were unthought of when many documentation collections were originally created, both in terms of research and in terms of distributing and use of linguistic

* The authors would like to thank the Yale Linguistics Fieldwork Group, the UC Berkeley Language Revitalization Working Group, the audience at PARADISEC@100, and audiences at the Annual Meeting of the LSA in Washington, DC (2022) and online for their feedback on this work at its various stages. We would also like to thank the depositors and language communities whose work forms the basis of this paper, and without whom this work (and much else) would not be possible. Authors: Sarah Babinski, Yale University; Jeremiah Jewell, Yale University; Kassandra Haakman, Yale University; Juhyae Kim, Cornell University; Amelia Lake, Yale University; Irene Yi, Yale University; & Claire Bowern, Yale University. Corresponding authors: Irene Yi (irene.yi@yale.edu) and Claire Bowern (claire.bowern@yale.edu).

resources in language reclamation, revitalization, and renewal. However, archives and individual collections are very heterogeneous in their usability and reusability as a starting point for linguistic research and/or language reclamation work. The aim of this paper, then, is to investigate the implications of these new possibilities for existing documentary corpora and the application of digital methods to linguistic research and community-oriented linguistic dissemination.

This paper examines the processes for continued linguistic work from digital archival sources, that is, the archive ⟹ documentation/description/analysis data pipeline. We report on the results of an investigation of a sample of individual archival collections. The review of individual collections is part of a broader project: a review of online digital language archives listed in OLAC's list of participating archives, as well as DELAMAN members and associate members. The results of the archive "audit", where we examined issues of accessibility, discoverability, and functionality of 41 archives, is reported in a companion paper (Yi et al. 2022). While this paper focuses on individual collections and the companion paper focuses on archives at a higher level, we emphasize that neither depositors nor archives act in isolation; reducing heterogeneity of archive level protocols will create standards for depositors to follow in their collections, and using "best practices" at the collections depositor level will in turn help uphold the consistency of inter-archive and intra-archive usability. Further, this broader investigation helps clarify the roles and responsibilities between archivists who maintain archives and those who deposit language materials into an archive.

To identify the individual collections we focus on in the present paper, we took a top-down approach, starting at a digital archive and then randomly selecting language collections to focus on within that archive. The audited collections represented a diverse range of language families and geographies, though with a particular focus on under-resourced languages. We tested 20 collections from 6 popular archives in their ease of completing a set of standard investigations commonly used for phonetic research and other types of linguistic research (described further in §3). We followed a survey of questions to investigate for each collection, and we tracked findings and issues we ran into in this survey. From our findings, we present a set of recommendations geared toward depositors of archival collections that can help ensure their materials be usable and reusable for future work. With these recommendations come important caveats, as we do not wish to discourage individuals from depositing their collections, and we firmly believe that depositing incomplete or imperfect materials to be archived is better than not depositing at all. We recognize the innumerable tradeoffs present in all aspects of language documentation and depositing materials. However, by addressing issues that are present across a variety of collections, individuals can undertake these small but important practices in their depositing process, and the benefit of these "best practices" will thus be felt long thereafter. Better to introduce better documentation standards now than to discover in twenty years that we could have made our lives easier and language records better. Naturally, as more languages become endangered and fall out of use, linguists will be increasingly relying on data from archival sources. Crucially, communities looking to linguistic archives for reclamation and renewal in the future should have the best possible materials to work from.

The remainder of this paper proceeds as follows: §2 provides background on digital language archives and archiving. §3 gives information about the review procedures, choice of collections, and the criteria used for evaluation. §4 provides the results, while §5 gives additional discussion and recommendations.

2. Background. Linguists have been documenting languages with a view to continued language work after their endangerment and extinction for many years (cf. Dobrin, Austin & Nathan 2007; Hammarström et al. 2018; Broeder et al. 2011; Himmelmann 1998; McDonnell, Berez-Kroeker & Holton 2019; Henke & Berez-Kroeker 2016). Digital archives such as ELAR have been very concerned with models of data accessibility for academics and communities (cf. Nathan 2010; Thieberger et al. 2015a, b; Harris et al. 2015). But archives distribute the data that linguists contribute, and if linguists see archiving as the *endpoint* of a detailed and complex process of data gathering (cf. Bowern 2008/2015:47), not the *starting point*, issues relevant for future use could be overlooked. At the same time, digital documentation is now pervasive and has allowed the collection of larger and more varied corpora than could have been previously anticipated. However, no one, to our knowledge, has systematically investigated how depositors' choices in structuring their collections might affect the usability of language resources (though cf. Innes 2010; Burke et al. 2021; Dobrin et al. 2007). While archival collections are, to some extent, being used, the majority of current language work is done from either primary fieldnotes (that is, linguists working on their own collections) or from secondary, published materials (Bowern 2018). Bowern (2008/2015:47) illustrates this view, depicted in Figure 1, where archiving is an endpoint of the documentation process after a detailed and complex process of data gathering.[2] Some work does not consider archive use at all. Gonzalez et al. (2017), for example, describe three phases of a data life cycle for endangered languages: acquisition, manufacturing, and diffusion (i.e. publication), but not long term preservation. Rehm (2016) also makes this omission.

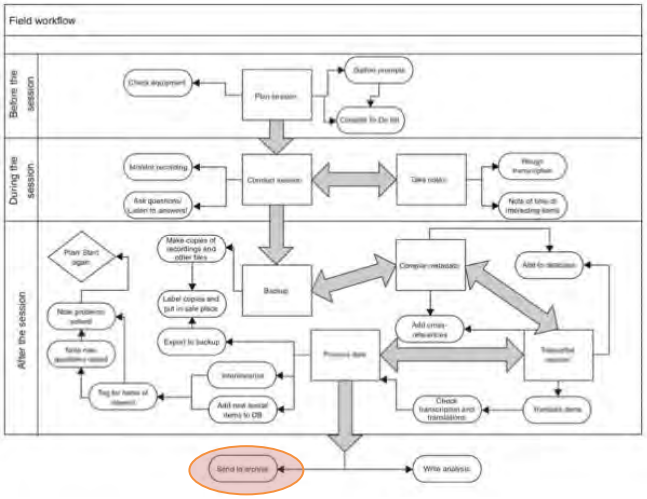

Figure 1. "Send to archive": the endpoint of a complex workflow

Finally, in discussions of digital archiving, digital access, and digital dissemination, prior work tends to focus either on the work of linguists who are probably outsiders to the language community (e.g. Solano, Nicholas & Wray 2018; Moeller & Hulden 2018; Schwartz et al. 2019; Michaud et al. 2018; Gessler 2019) or on resources for communities (Genee & Junker 2018; Bischoff & Jany 2018). This paper specifically looks for ways in which community materials can be integrated with research aimed at linguistic analysis, so that such materials can be either produced (or at least drafted) simultaneously with minimal additional overhead. That is, we

[2] The diagram in Figure 1 is not meant to be fully legible. It is included in miniature to show the degree of detail presented for the pre-archiving stage in comparison to the final archiving result (circled).

prioritize collection tests that are likely to lead to community outcomes as well as theoretical/documentary linguistic ones. We note, however, that community priorities vary.

Archived digital corpora vary extensively, even within the parameters set by individual archives. Some aspects of documentary standards were set early. For example, by the time Bowern (2008/2015) was writing recommendations for audio data formats (e.g. stereo, lossless formats, digitized at 44,100 Hz), these standards were already well established among fieldworkers (cf. Bird and Simons 2003, Barwick 2006). Another is metadata; indeed, most digital archives only accept deposits which are accompanied by appropriately structured information about the collection (cf. Sullivant 2020). Other aspects of digitization have no standard protocols, however. To take one simple example, if there are two speakers or signers on a recording, how is that represented? Do two speakers have their own tracks in a stereo recording (or video views for signers)? Are they transcribed on the same tier (with an annotation as to which is which), or on different tiers? Are the participants identified in the recording filename, or in unlinked metadata? For a human linguist reviewing recordings or working with the data directly, these choices are usually of little consequence. For working with materials computationally, however, they affect the way in which data should be extracted and processed.

A second example involves auxiliary files. Several software programs allow the user to customize settings, views, and other aspects of the program, and those choices are recorded in settings files. Linguists vary in whether they archive such files or not. Where they do not, usability of the deposit is affected, sometimes to the point of data loss. For example, ELAN (Wittenburg et al. 2006) settings files are typically only cosmetic, and losing them does not affect usability. The exception is information about audio-transcript offsets, which is crucial for being able to accurately replay the files. If this information is not present, file offsets must be manually recreated. Digital corpora may vary in other ways. The circumstances of recording will affect how usable the materials are for computational work (such as the presence of background noise, legibility, and accuracy). That is, the variation in collection setup is in addition to the variability that results from different recording circumstances, linguistic structures, and the like.

In summary, digital archives such as ELAR and Paradisec are building up substantial digital holdings with much variation in the composition and formats of individual collections. No one, to our knowledge, has systematically investigated how depositors' choices for archiving might affect the uses to which these corpora can be put in the future.[3] It is important to find this out now, while something can be done about it. It would be dreadful if we found in twenty years that a substantial fraction of our digital language holdings were not as usable as they could have been. This is the rationale for the current investigation.

As mentioned above, documentary linguists have tended to think of archiving as an end point after planning and conducting fieldwork, rather than the input to further work. Yet at the same time, there is a large amount of work on the outputs of memory documentation; that is, writing grammars, dictionaries, and analytical articles from text collections and other early archival sources (see also Baldwin & Olds 2007; Whalen, Moss & Baldwin 2016; Hinton 2003, 2018). Contemporary collections, however, are quite different from those early products. The

[3] There is, of course, a long literature on archiving, particularly since the advent of digital holdings. Hinton (2001) discussed archives in language reclamation. A small sample of this literature includes Khait et al.'s (2021) survey of archive users for their needs, and Wasson et al. (2016) on how to center users in archive web design. Sullivant (2020) describes archival metadata (and its problems). Barwick and Thieberger (2006, 2018) describe sustainable data collection practices (a precursor to longterm archiving) and the FAIR principles for archiving. Harris et al. (2015) and Thieberger et al. (2015a, b), focusing on the Paradisec archive, discuss archives' responsibilities.

early products are simpler, in that they tend to contain texts, vocabularies, and relatively small amounts of audio recording. Such collections usually need processing before they can be used in linguistic reclamation work. That is, they usually cannot be used without further work on the collection.[4] Examples include making dictionaries with audio examples, conducting linguistic work on the collections to write a reference grammar or conduct other linguistic research (McAuliffe et al. 2017; DiCanio et al. 2015, Babinski et al. 2019); compiling narratives into a book, or; making pedagogical materials from elicitation sessions (cf. Cruz 2021).

While there are publications based on data collected by others, such publications typically gloss over the work required to make the collections usable for their analyses. Many linguistic publications mention something along the lines of "data were extensively cleaned" with no further explanations. Examples include Raha et al. (2020):

> Our method includes scraping of code-mixed English-Bengali tweets on Twitter and cleaning them…. The collected tweets contained multiple degrees of noise and hence, it needed to be cleaned before using it to develop our future systems. After cleaning the tweets….

A footnote in Dowlagar and Mamidi (2021) states "We pre-process the data to deal with variations in spelling and transliterations." In Barman et al. (2016), a footnote about their cleaned dataset says "We are preparing to release the data set," yet we were unable to find the released dataset. As these publications mention how crucial data cleaning and preprocessing were to their methods and results, it is important that these cleaning methods are transparently communicated. Not only does this aid in reproducibility and replicability of important work, good documentation of data processing is good practice. Stanley (2021) notes that the order of operations of cleaning and processing a dataset can affect the results; thus such information should be recorded and reported. Finally, if researchers are extensively cleaning datasets before use, those doing language documentation might be able to adjust methods to make the use of such collections more straightforward.

3. Methods and data.

3.1. COLLECTIONS. The general principles for choosing collections for the current study are as follows. We examined 20 collections from 6 popular archives, with a global focus to represent languages across the world. We concentrated on publicly accessible data which was obtainable without further restriction.[5] Secondly, collections may have had a variety of materials, but for us to have used them here, they must have included at least some sound files and transcripts.[6] Collections varied in size, and we did not explicitly track the size of the collection (. We did not place a minimum or maximum limit on the amount of data in the collection, and collections ranged in size from 6 narratives of 1-2 minutes each, to many hours of recordings. We focused on materials that were deposited at archives within the last 15 years, whether digitally recorded

[4] An exception to this might be cases where community members play unedited recordings. This is one use case, but even in this case, some "processing" is usually needed, such as copying files to an audio player or computer and working out what is in each track.

[5] We understand and respect that material may be restricted for cultural reasons.

[6] We tracked attempts to access corpora that do not have relevant material. Archives such as Paradisec, ELAR, and AILLA allow users to search by material type, which allowed us to target corpora that were suitable for our purpose.

originally or digitized from analog sources. Because digital language recording standards have been essentially unchanged during this time, this window gave us a good snapshot of current practices.[7] Corpora were downloaded from the following digital archives:

- AIATSIS (Australian Institute of Aboriginal and Torres Strait Islander Studies)
- AILLA (Archive of the Indigenous Languages of Latin America)
- Berkeley Survey of Californian and other Indian Languages
- ELAR (Endangered Language Archive housed by SOAS)
- LACITO's Pangloss Collection
- Paradisec (Pacific and Regional Archive for Digital Sources in Endangered Cultures)

Collections included in the archive audit were selected randomly, with some effort to include a range of archival sources and geographical spread of languages, as shown on the map in Fig. 2. Deposits were investigated using a top-down approach, without prior knowledge of the collections themselves and only using the information provided by the archive to identify potentially usable materials.[8]

Figure 2. Location of languages used in the review sample

The twenty collections used in this paper are not explicitly identified. This is deliberate. We did not wish to give the appearance of criticizing particular individuals, communities, or archives; as is clear from our results, these problems are most likely widespread. We also

[7] It was pointed out in the question period at the LSA, however, that while practices may have been fairly stable from the depositors' perspective, standards on the archive side have changed greatly over this time. We recognize that this is the case. We do not know how changing archival practices within digital archives might have affected the results here. As noted in Yi et al. (2022), crucial information about language archives (such as their backup procedures and content management systems) is not readily available.

[8] Note that only a small number of collections were investigated in detail. While we applied for NSF funding for a larger project, the grant was not successful. This publication, Yi et al. (2022), our LSA presentations and a webinar (March 11, 2022) are attempts to begin a conversation about these issues, with a set of pilot data. Note that although the number of collections examined was small, the number of problems was large and varied, and we have no reason to believe that we have come close to exhausting the issues that such work raises. Every collection raised new problems (see further §4 below).

appreciate that many different considerations go into the building of a documentary collection and our methods of review do not give access to knowledge of such considerations. "Archived but incomplete" is better than "not archived at all", as Yi et al. (2022) emphasize.

3.2. PARAMETERS OF THE REVIEW. We investigated three aspects of each collection. First, we examined what metadata information was available for each collection. An ideal archival collection would have easily findable metadata on the number and types of files, as well as summaries of file contents, including speaker information, genre (e.g. narrative, elicitation), and circumstances of recording. The second set of questions we had centered around availability and format of audio recordings.[9] We sought information about what type of audio was available (e.g. WAV, MP3, M4A), how many minutes/hours of content was available, and how easy it was to find this information. In some cases, this information was available in a general metadata file, but in other cases this information had to be intuited from file names or other sources.

Finally, we looked specifically at the usability of collection materials for forced alignment (cf. Evanini et al 2008; McAauliffe et al 2017; cf. Babinski 2022a, 2022b). We determined forced alignment to be a good stress test for collections: it requires good quality audio recording, accessibility of files and file formats, and completeness of transcripts. Going through the process of automatic alignment allows us to identify potential problems such as corrupted files, inaccurate or incompatible transcripts, and inconsistent file type designations. It is also a useful tool for those working with any large amount of audio data. That is, whether phonetic research and forced alignment is the direct, end goal of a research project is less relevant for evaluating collections, since the methods of forced alignment place certain strict requirements on collections.[10] Finally, forced aligned materials are useful for the creation of other materials, such as talking dictionaries, embedded examples in pdfs, and captioned recordings. They create opportunities for accessible materials for community-oriented language documentation.

3.3. SUMMARY OF AIMS. A portion of the questions we surveyed and answered for each collection are as follows:

- Where individual speaker attribution is appropriate, how easy is it to recover that information?
- How easy is it to retrieve metadata? Can we find out easily what is in the collection?
- Does the metadata match the deposit? If not, in what ways?
- How complete are transcript files?
- Does the corpus have digital alignments at the utterance level? If not, can we create them (using a forced aligner)?
- Can we align the corpora (or a substantial proportion of it) at the segment level?
- Can we extract data about the segments (formant data for vowels, f0 measurements, for example) to produce a basic acoustic description of the language?
- If the answer is no for any of the above, what problems were encountered? Why did alignment fail? What features were the result of earlier data processing choices, rather than being inherent to the corpus materials themselves? Is it possible to alter or adapt workflows so that usable data may be obtained?

[9] We did not investigate video files in the collections. Given the problems we had with downloading audio files (cf. Yi et al. 2022), the additional size of video downloads would have been an additional substantial hurdle.

[10] Note that we focused here on audio collections of spoken languages. This was because we used forced alignment. See §5 below for comments on how our results generalize to corpora with visual modality data.

We kept track of the answers to these questions in a shared google document, along with other annotations of problems that arose, their workarounds, and screenshots that illustrated problems and solutions. More regimented methods of tracking files proved infeasible. We originally attempted to use an issue tracking system, but it proved infeasible to track the issues that arose and their solutions. That is, issue track software is oriented towards addressing and solving issues, whereas we wanted to conduct analytics on the issues that arose (as well as addressing them).

4. Results.

4.1. FINDINGS: PROBLEMS AND SOLUTIONS. We encountered varied and substantial problems in completing any of the tasks we used as benchmarks. In presenting these findings, we class issues as those that made it impossible to work further with materials in the collections (for our purposes, i.e. towards forced alignment and phonetic work), and problems that added substantially to processing time but did not ultimately prevent alignment.

The issues that made it impossible to work further for our purposes included unavailable sound files, irreconcilable metadata, and unprocessable transcripts. Of the 20 collections surveyed, 14 had no available sound files, even though we began our search by filtering for collections which were flagged as containing sound files. There were a few sources of this problem. Some collections had audio files, but they were not part of the public collection. While we only surveyed freely accessible collections, some collections had mixed access status, with the audio recordings requiring additional permissions. In several other cases, the collection metadata was available but the materials had not yet been released by the archive and were still in processing.

Secondly, 11 of the 20 collections had irreconcilable metadata which made the recordings unusable for our purposes. These included problems such as inconsistent file naming conventions between the transcript files and the audio files, which meant that we could not match the transcripts to the audio. Four collections had unprocessable transcripts, meaning we could not work with the transcripts towards forced alignment and other phonetic work. In one case, the transcripts were built into a web audio player and could not be extracted for forced alignment. In another, the transcripts existed in an unparsable xml format (or required great expertise in working with xml to parse, which limits the accessibility to researchers who may not have such expertise). Finally, some transcripts were incomplete or contained only empty annotations. Figure 3 illustrates an ELAN file with empty annotations.

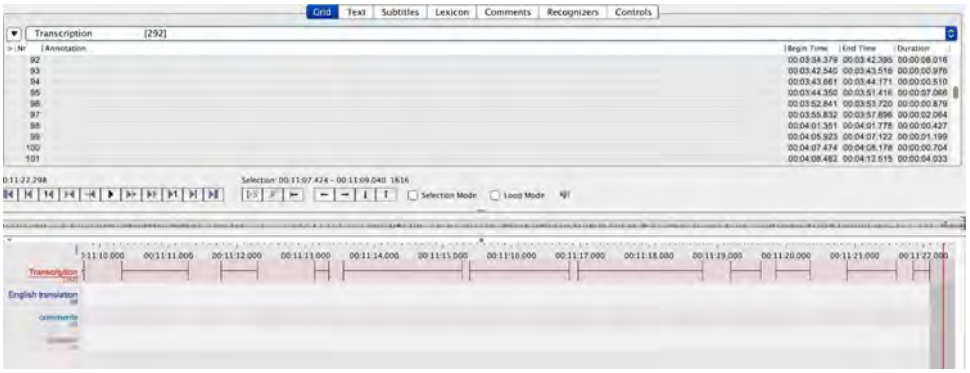

Figure 3. Empty transcripts in ELAN file

Next, we encountered issues that did not prevent further work with the collection, but added to the processing time of the materials. Out of the 11 collections that we were able to continue working with, 8 contained files without matching metadata. There were also inconsistencies in file names; for example spaces in between words in file names were at times represented with an underscore (_) and other times represented with a hyphen (-) or space ( ), which is not an issue for a human looking at the files, but makes running computational scripts on the files impossible until they are renamed to be consistent. Secondly, many of the collections had inconsistent transcription tier use, meaning that they had notes in transcription tiers, or the order of tiers and tier types within collections differed between files in the same collection. Out of the 5 collections that had ELAN transcriptions available, 2 had no preserved ELAN offsets in their settings file, so the transcripts had to be manually realigned, adding substantial preprocessing time to using the materials. Such settings files are crucial to the ability to reuse archival collections and materials.

Depositors' file naming practices cause problems. For example, it is common to use file names in version control, making copies of a file and changing the name. This is seen in Figure 4 where there are two ELAN transcript files: `CB0702.eaf` and `CB0702-rev.eaf`.[11] The second is presumably a revision of the first (as is implied by the later date of modification and the larger file size. Therefore, the "revised" file is what we would want to watch to the audio file for forced alignment other uses. However, if we were to run forced alignment on this directory with no further modification, the revised file would not be selected as the match for the audio file, as the original file has the same name (minus filetype suffix) as the audio file. A human would not make this error if working with files one-by-one. Ideally one would avoid using file naming as version control, but if it is truly necessary, the most up to date version should have the same name as the audio files, and the earlier versions should have the version suffix.

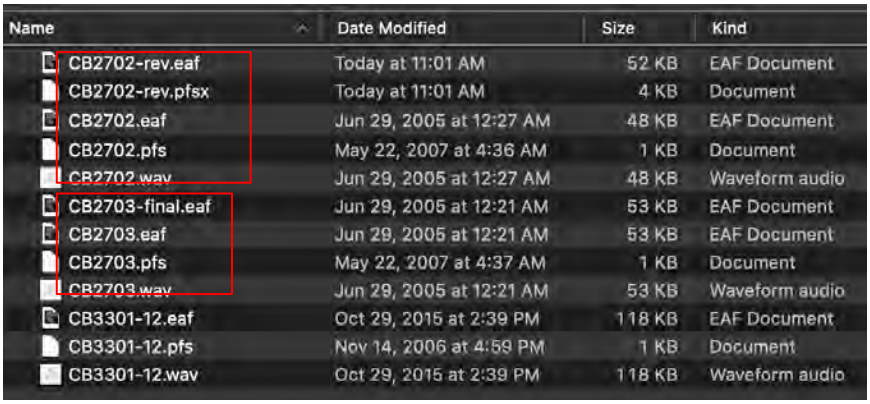

Figure 4. Version control inconsistencies in file naming

We also encountered font, encoding, and data structure issues. These did not prevent alignment but were time-consuming to diagnose and fix. These included text being broken up across lines (e.g. "nonbreaking spaces") in the middle of words or transcript chunks, and stray characters in the middle of other words (that could stem from file corruption); one document had single Chinese characters appearing in certain words (e.g. "chara糸cters"). Further, Legacy Mac

[11] Note that in the interests of preserving anonymity, we recreated this example using different filenames.

OS files were sometimes incorrectly saved as unicode, which resulted in in transcriptions like "ŋätji" being realized as the sequence of characters "≈ã√§tji".[12] Finally, there were occasionally delimiters in the wrong places (e.g. tabs or line breaks within records, or incorrect parsing of quotation marks), which do not prevent humans from interacting with the files, but cause would be workable when read by a human, but would hinder any automated or computational script run on such transcripts.

These findings came in addition to the usual amount of variation we expected as a natural result of languages being social objects that are different from one another (e.g. in differing orthographies or transcription standards). The results we presented above proved to be issues that were not foreseeable at the time of choosing collections to investigate. Rather, these obstacles emerged throughout the review or in retrospect, which further shows just how important such a review was on the accessibility and usability of language materials found in archival collections.

4.2. SUCCESSFULLY ALIGNED COLLECTIONS. Initially, only one collection was successfully force aligned without substantial intervention of cleaning or preprocessing. Even so, substantial preprocessing was required, as follows.

First, some of the files required renaming. For example, one collection had ELAN transcript files (.eaf), which are a type of XML file. At some point (either in the archive upload or download process), these files acquired an additional extension (e.g. XXX.eaf.xml). Other files e.g. from .xml formats to .eaf. We resolved and removed file naming inconsistencies that resulted from versioning (cf. Figure 4). In several cases, audio files were rematched to transcripts, where the titles of the transcripts were in one language but the audio files were translated into a different language. Additionally, we removed non-transcript items from the transcription lines and tiers of our files, and we corrected for inconsistencies or mistakes in font and character encoding issues. Finally, there were grapheme-to-phoneme conversion problems we needed to correct for, such as Unicode characters that were not handled by the g2p, and finding equivalents for a pronunciation dictionary. One collection had videos which were downloaded as .mp4 and the audio channel extracted. All WAV files were downsampled. One set of files had been upsampled (to 96,000 Hz from the original 44,100 Hz). While the mfa alignment initially threw errors that appeared to be due to audio file corruption, subsequent investigation revealed that the errors were due to extra spaces in the TextGrids produced by bulk export from ELAN, rather than being an error in the audio files.

5. Discussion and conclusions.

5.1. RECOMMENDATIONS. In conclusion, there is a considerable gap between the ideal language documentation archive and current practice. There are many circumstances that may lead to collections being the way they are, but many of the problems identified here are preventable. In summary, the following broad recommendations arise. The recommendations made in this paper are not so different from those made in Sullivant (2020), Chelliah (2021), Holton et al. (2017), Bird & Simons (2003), among others. They are worth repeating, however. While working on the collection, be as consistent as possible. Decide on a scheme and stick to it. Put time aside specifically for collection management, and address issues early. Document

[12] This seems to have happened at different points in the file history for different files. For some, they appear to have been archived as Unicode files (or converted to Unicode upon upload, with the archive introducing the encoding problem). In other cases, our own workflows introduced this error.

decisions, both for yourself and for users of the archival collection, and archive that document along with other parts of the collection.

Include rich metadata with your files. Important information such as date of recording and speaker information should be included in metadata when possible if the speakers and community members who contributed to the gathering of those materials consent to such information being included.[13] While preparing the collection for archiving, be clear about what type of files are necessarily part of the collection (and conversely, what should not be there). Test the collection before depositing it. That is, imagine you are working on the materials, can you find and use the files you need? Remember that settings (such as file locations and other document preferences) are often saved from session to session.

Given that the issues raised here apply to existing collections, it would be ideal if archives could allow depositors to update and revise their collections to address some of these problems. Finally, a broader review of collections needs to take place. While there were recurring problems, every collection we examined introduced new complexities for data processing. Given that we looked at a small number of collections, from a small number of archives, it is clearly important to address these questions at a larger scale.

5.2. GENERALIZATIONS BEYOND AUDIO. This pilot study only considered spoken language collections. This was because we wanted to use a single test and benchmark for all corpora (audio forced alignment). We lack the domain-specific knowledge to evaluate signed language and other video-documented collections, and we lack the digital infrastructure to handle the terabytes of video footage that would result. As it stood, we ran into problems with audio data. That said, some of these results generalize, while others do not. First, many problems stemmed from incomplete or inaccurate metadata. Such issues will apply regardless of language modality. Metadata is important, it needs to be there, and it needs to be accurate for collections to be usable. Secondly, problems arose from file naming conventions, insufficient files archived – that is, problems arose from the way that items within the collection were (or were not) linked to one another. That also applies regardless of language modality. Thirdly, we had issues working with the collections (file storage and management); such issues are likely to be more acute with video data. Fourthly, there were problems that arose from the circumstances of recording: high signal to noise ratios, for example. The details are different for signed language data but presumably also affect research: insufficient frame rates, signers not within the full video frame, poor lighting, for example. Finally, we had issues with backwards compatibility of file formats, and issues of backwards compatibility and file longevity are ubiquitous. Therefore our tests raise issues for documentary collections regardless of modality. However, there are clearly additional considerations for video data and signed languages which are not raised here.

5.3. FURTHER DISCUSSION. We recognize that this is not a new problem; rather, people who work with language collections have been writing about these issues for almost 20 years. It is striking, then, that these problems are still pervasive. A quote from Bird and Simons (2003) concretely timestamps just how long these problems have been discussed (emphasis ours):

> The issue is acute for endangered languages. ***In the very generation when the rate of language death is at its peak, we have chosen to use moribund technologies, and to***

[13] In our 'in case you missed it' LSA webinar on March 11, 2022, we asked audience members what their highest priority needs were in language documentation and more than 50% mentioned a tool to curate consistent metadata. We suggest lameta (Hatton et al 2021) for this purpose.

> ***create endangered data.*** When the technologies die, unique heritage is either lost or encrypted. Fortunately, linguists can follow BEST PRACTICES in digital language documentation and description, greatly increasing the likelihood that their work will survive in the long term.
>
> An archival format is useless unless there are tools for creating, managing, and browsing the content stored in that format…. The technological solutions must be coupled with a sociological innovation, one that produces broad consensus about the design and operation of common digital infrastructure for the archiving of language documentation and description.

Finally, we reiterate that archives often do not receive the funding they need, and they rely on a great deal of volunteer energy, labor, and expertise to function. Likewise, language collections are often produced incidentally, and we believe that a collection that is "incomplete but archived" is clearly preferable to "not archived at all". We do not wish to discount any of the work that depositors of language collections put into preserving these materials. We recognize that, as academic linguists, we do not represent all opinions and views around archive use and reuse, nor do we encompass the unique set of questions and obstacles each individual depositor experiences in gathering and depositing materials. Nonetheless, it is important to ask these questions and examine the current state of archiving options, given just how important and irreplaceable the language materials are that archives hold for communities and researchers.

References

Babinski, Sarah. 2022a. Best practices in the collection and analysis of "noisy" audio in phonetics. Presentation at the Annual Meeting of the Linguistic Society of America, Washington, DC.

Babinski, Sarah. 2022b. *Archival phonetics and prosodic typology in 16 Australian languages*. New Haven, CT: Yale University dissertation.

Babinski, Sarah, Rikker Dockum, J. Hunter Craft, Anelisa Fergus, Dolly Goldenberg, & Claire Bowern. 2019. A Robin Hood approach to forced alignment: English-trained algorithms and their use on Australian languages. *Proceedings of the Linguistic Society of America* 4(1). 3. https://doi.org/10.3765/plsa.v4i1.4468.

Baldwin, Daryl & Julie Olds. 2007. Miami Indian language and cultural research at Miami University. In Daniel Cobb & Loretta Fowler (eds.), *Beyond red power: American Indian politics and activism since 1900*, 280–90. Albuquerque: University of New Mexico Press.

Barman, Ustab, Joachim Wagner, & Jennifer Foster. 2016. Part-of-speech Tagging of Code-mixed Social Media Content: Pipeline, Stacking and Joint Modeling. *Proceedings of the Second Workshop on Computational Approaches to Code Switching* 5804. 30–39. https://doi.org/10.18653/v1/W16-5804.

Barwick, Linda & Nicholas Thieberger (eds.). 2006. *Sustainable data from digital fieldwork*. Sydney: Sydney University Press.

Barwick, Linda & Nicholas Thieberger. 2018. Unlocking the archives. In Vera Ferreira & Nick Ostler (eds.), *Communities in Control: Learning tools and strategies for multilingual endangered language communities. Proceedings of the 2017 XXI FEL conference*. 135–139.

Bird, Steven & Gary Simons. 2003. Seven dimensions of portability for language documentation and description. *Language* 79(3). 557–582. https://doi.org/110.1353/lan.2003.0149.

Bischoff, Shannon T. & Carmen Jany (eds.). 2018. *Insights from practices in community-based research: From theory to practice around the globe*. Berlin: De Gruyter Mouton.

Bowern, Claire. 2008/2015. *Linguistic fieldwork: A practical guide*. Basingstoke: Palgrave Macmillan.

Bowern, Claire. 2018. Reflections on linguistic fieldwork. In Bradley McDonnell, Andrea L. Berez-Kroeker & Gary Holton (eds.), *Reflections on language documentation 20 years after Himmelmann 1998* (Language Documentation & Conservation Special Publication no. 15). 202–209. Honolulu: University of Hawai'i Press. http://scholarspace.manoa.hawaii.edu/handle/10125/24822.

Broeder, Daan, Han Sloetjes, Paul Trilsbeek, Dieter Van Uytvanck, Menzo Windhouwer & Peter Wittenburg. 2011. Evolving challenges in archiving and data infrastructures. In Geoffrey L.J. Haig, Nicole Nau, Stefan Schnell & Claudia Wegener (eds.) *Documenting endangered languages: Achievements and perspectives*, 33–54. Berlin: De Gruyter Mouton. https://doi.org/10.1515/9783110260021.33.

Burke, Mary, Oksana L. Zavalina, Mark Edward Phillips & Shobhana Chelliah. 2021. Organization of knowledge and information in digital archives of language materials. *Journal of Library Metadata* 20(4). 185–217. https://doi.org/10/gjr93v.

Chelliah, Shobhana. 2021. *Why language documentation matters.* Berlin: Springer.

Cruz, Hilaria. 2021. *KwanC laE ngyanJanI siK tykwenqEenE ktyiC chaqF tnyaJ* (Chatino Tonal Books Project 7). https://ir.library.louisville.edu/chatino/7.

DiCanio, Christian, Hosung Nam, Jonathan D. Amith, Rey Castillo García & Douglas H. Whalen. 2015. Vowel variability in elicited versus spontaneous speech: Evidence from Mixtec. *Journal of Phonetics* 48. 45–59. https://doi.org/10.1016/j.wocn.2014.10.003.

Dobrin, Lise, Peter Austin & David Nathan. 2007. Dying to be counted: Commodification of endangered languages in documentary linguistics. In Peter Austin, Oliver Bond & David Nathan (eds.), *Proceedings of the Conference on Language Documentation and Linguistic Theory*, 59–68. London: HRELP, SOAS.

Dowlagar, Suman & Radhika Mamidi. 2021. A pre-trained transformer and CNN model with joint language ID and part-of-speech tagging for code-mixed social-media text. In Ruslan Mitkov & Galia Angelova (eds.), *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, 367–374. INCOMA Ltd.

Genee, Inge & Marie-Odile Junker. 2018. The Blackfoot Language Resources and Digital Dictionary project: Creating integrated web resources for language documentation and revitalization. *Language Documentation and Conservation* 12. 298–338..

Gessler, Luke. 2019. Developing without developers: choosing labor-saving tools for language documentation apps. In Antti Arppe, Jeff Good, Mans Hulden, Jordan Lachler, Alexis Palmer, Lane Schwartz & Miikka Silfverberg (eds.), *Proceedings of the 3rd Workshop on the Use of Computational Methods in the Study of Endangered Languages* (Volume 1, Papers), 6–13. Honolulu: Association for Computational Linguistics. https://www.aclweb.org/anthology/W19-6002.

Gonzalez, J. L., Anuschka van't Hooft, Jesus Carretero & Victor J. Sosa-Sosa. 2017. Nenek: A cloud-based collaboration platform for the management of Amerindian language resources. *Language Resources and Evaluation* 51(4). 897–925. https://doi.org/10.1007/s10579-016-9361-8.

Hammarström, Harald, Thom Castermans, Robert Forkel, Kevin Verbeek, Michel A. Westenberg & Bettina Speckman. 2018. Simultaneous visualization of language endangerment and language description. *Language Documentation and Conservation* 12. 359–392. http://hdl.handle.net/10125/24792.

Harris, Amanda, Nick Thieberger & Linda Barwick. 2015. *Research, records and responsibility: Ten years of the Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC)*. Sydney: Sydney University Press.
Hatton, John, Gary Holton, Mandana Seyfeddinipur, Nick Thieberger. 2021. Lameta [software] https://github.com/onset/laMETA/releases.
Henke, Ryan & Andrea L. Berez-Kroeker. 2016. A brief history of archiving in language documentation, with an annotated bibliography. *Language Documentation & Conservation* 10. 411–457. http://hdl.handle.net/10125/24714.
Himmelmann, Niklaus. 1998. Documentary and descriptive linguistics. *Linguistics* 36(1). 161–196. https://doi.org/10.1515/ling.1998.36.1.161.
Hinton, Leanne. 2001. Language revitalization: An overview. In Leanne Hinton & Kenneth Hale (eds.), *The green book of language revitalization in practice*, 3–18. Boston: Brill.
Hinton, Leanne. 2003. How to teach when the teacher isn't fluent. In Jon Reyhner, Octaviana V. Trujillo, Roberto Luis Carrasco & Louise Lockard (eds.). *Nurturing native languages,* 79–92. Flagstaff, AZ: Northern Arizona University.
Hinton, Leanne. 2018. Approaches to and Strategies for Language Revitalization. In Kenneth L. Rehg & Lyle Campbell (eds.), *The Oxford handbook of endangered languages*, 442–465. Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190610029.013.22.
Holton, Gary, Kavon Hooshiar & Nicholas Thieberger. 2017. Developing collection management tools to create more robust and reliable linguistic data. In Antti Arppe, Jeff Good, Mans Hulden, Jordan Lachler, Alexis Palmer & Lane Schwartz (eds.), *Proceedings of the 2nd Workshop on Computational Methods for Endangered Languages*, 33–38, Honolulu. Association for Computational Linguistics
Innes, Pamela. 2010. Ethical problems in archival research. *Language & Communication* 30(3). 198–203. https://doi.org/10.1016/j.langcom.2009.11.006.
Khait, Ilya, Leonore Lukschy, & Mandana Seyfeddinipur. 2021. Linguistic archives and language communities questionnaire: Establishing (re-)use criteria. *Proceedings of the 1st International Workshop on Digital Language Archives*. https://doi.org/10.12794/langarc1851179.
McAuliffe, Michael, Michaela Socolof, Sarah Mihuc, Michael Wagner, & Morgan Sonderegger. 2017. Montreal Forced Aligner: Trainable text-speech alignment using Kaldi. *Interspeech* 2017. 498–502. https://doi.org/10.21437/Interspeech.2017-1386.
McDonnell, Bradley, Andrea L Berez-Kroeker & Gary Holton (eds.). 2019. *Reflections on language documentation 20 Years after Himmelmann 1998*. Honolulu: University of Hawai'i Press.
Michaud, Alexis, Oliver Adams, Trevor Anthony Cohn, Graham Neubig & Séverine Guillaume. 2018. Integrating automatic transcription into the language documentation workflow: Experiments with Na data and the Persephone toolkit. 481–513. *Language Documentation & Conservation* 12. 393–429.
Moeller, Sarah & Mans Hulden. 2018. Automatic glossing in a low-resource setting for language documentation. In Judith L. Klavans (ed.) *Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages*, 84–93. Santa Fe, NM: Association for Computational Linguistics. https://www.aclweb.org/anthology/W18-4809.
Nathan, David. 2010. Archives 2.0 for endangered languages: From disk space to MySpace. *International Journal of Humanities and Arts Computing* 4(1–2). 111–124.

Raha, Tathagata, Sainik Kumar Mahata, Dipankar Das, & Sivaji Bandyopadhyay. 2020. Development of POS tagger for English-Bengali Code-Mixed data. *16th International Conference on Natural Language Processing (ICON-2019)*. https://doi.org/10.48550/arXiv.2007.14576.

Rehm, Georg. 2016. The language resource life cycle: Towards a generic model for creating, maintaining, using and distributing language resources. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)* 2450–2454. https://www.aclweb.org/anthology/L16-1388.

Schwartz, Lane, Emily Chen, Benjamin Hunt & Sylvia L.R. Schreiner. 2019. Bootstrapping a neural morphological analyzer for St. Lawrence Island Yupik from a finite-state transducer. In Antti Arppe, Jeff Good, Mans Hulden, Jordan Lachler, Alexis Palmer, Lane Schwartz, Miikka Silfverberg (eds.), *Proceedings of the 3rd Workshop on the Use of Computational Methods in the Study of Endangered Languages* (Volume 1, Papers), 87–96. Honolulu: Association for Computational Linguistics. https://www.aclweb.org/anthology/W19-6012.

Solano, Rolando Coto, Sally Akevai Nicholas & Samantha Wray. 2018. Development of natural language processing tools for Cook Islands Māori. *Proceedings of the Australasian Language Technology Association Workshop* 2018. 26–33. https://www.aclweb.org/anthology/U18-1003.

Stanley, Joey. 2021. Order of operations in sociophonetic data processing. Presentation at NWAV49 (online). https://youtu.be/8TEip-Fixyw.

Sullivant, Ryan. 2020. Archival description for language documentation collections. *Language Documentation & Conservation* 14. 520–578. http://hdl.handle.net/10125/24949

Thieberger, Nicholas, Amanda Harris, & Linda Barwick. 2015a. PARADISEC: Its history and future. In Amanda Harris, Nick Thieberger & Linda Barwick (eds.)*, Research, records and responsibility: Ten years of the Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC)*, 1–16. Sydney: Sydney University Press.

Thieberger, Nicholas, Anna Margetts, Stephen Morey & Simon Musgrave. 2015b. Assessing annotated corpora as research output. *Australian Journal of Linguistic*s 36(1). 1–21. https://doi.org/10/ghk659.

Wasson, Christina, Gary Holton & Heather S. Roth. 2016. Bringing user-centered design to the field of language archives. *Language Documentation & Conservation* 10. 641–681. http://hdl.handle.net/10125/24721.

Wittenburg, Peter, Hennie Brugman, Albert Russel, Alex Klassmann & Han Sloetjes. 2006. ELAN: A professional framework for multimodality research. In Nicoletta Calzolari, Khalid Choukri, Aldo Gangemi, Bente Maegaard, Joseph Mariani, Jan Odijk, Daniel Tapias (eds.), *Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation*, 1556–1559. Genoa: European Language Resources Association.

Whalen, D. H., Margaret Moss & Daryl Baldwin. 2016. Healing through language: Positive physical health effects of indigenous language use. *F1000Research* 5. 852. https://doi.org/10.12688/f1000research.8656.1.

Yi, Irene, Amelia Lake, Juhyae Kim, Kassandra Haakman, Jeremiah Jewell, Sarah Babinski, & Claire Bowern. 2022. Accessibility, discoverability, and functionality: An audit of and recommendations for digital language archives. *Journal of Open Humanities Data* [In press].