The Open Handbook of Linguistic Data Management

Edited By: Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller, Lauren B. Collister

Citation:

The Open Handbook of Linguistic Data Management
Edited By: Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller, Lauren B. Collister
DOI: 10.7551/mitpress/12200.001.0001
ISBN (electronic): 9780262366076
Publisher: The MIT Press
Published: 2022

The MIT Press

10 Linguistic Data in the Long View

Laura Buszard-Welcher

1 Introduction

How do we move human knowledge into the future?[1] It seems like this should be a fundamental question for any archival effort, because the intention of transmitting knowledge to future stakeholders is presumably a primary reason to go to the trouble of archiving in the first place. It becomes an especially important question for the creation and archiving of language data, because many of the world's languages are endangered, and this is an aspect of our shared humanity where we are at great risk of losing an expansive amount of human knowledge—of the languages themselves and any knowledge that is dependent on being communicated through them and the culture they are part of and express. Collected and archived endangered language data may be the only record of its kind available to the future,[2] and in the case of critically endangered languages, may be the only record of the language that remains at all.

For those in the trenches working to document against the ticking clock of language endangerment, there are pressing tasks of data collection, analysis, and presentation. For the archivist, there are tasks of ingesting collections, organizing and making them discoverable, migrating them as formats and storage practices change, all while managing the resources needed to maintain the archive and the accessibility of collections indefinitely. So, given the critical path to just getting everything done, perhaps it is not surprising that the question of how we move human knowledge into the future (and whether we are actually accomplishing that) usually remains unasked and unaddressed.

I came to the realization that this is a fundamental problem in archiving as a result of my own work, where I interact with a number of projects that are developing materials and methods—as well as curating content—for very long-term archiving, that is, on the scale of hundreds to even thousands of years. Indeed, I work on one such project myself.[3] If anyone would be working on the problem of how to move knowledge into the future, it seems like it should be this cohort of very long-term archivists. But we generally don't address the question either. We are pursuing emerging technologies, such as storing data in nano-manipulated quartz crystals (SPIE 2016) or the nucleotides of DNA (Church, Gao, & Kosuri 2012) or discovering methods we didn't even know existed such as quantum information storage in the orbital angular momentum of photons (Erhard et al. 2017). We are also exploring new archival environments such as storing data in salt mines (Memory of Mankind, n.d.), or on the moon (Arch Mission Foundation, n.d.), or in geosynchronous orbit (Quast 2018), or transmitting data across interstellar space (Interstellar Beacon, n.d.). These are methods of moving data into the future, but don't specifically address how knowledge will be transmitted to the future.

Partly the problem of how to move knowledge into the future relies on making sure it lasts and remains accessible. Any of these explorations into new archival materials and methods could lead to ways of reliably storing or transmitting data in the very long term, and they could help solve the problem of how to store large amounts of data reliably in the here and now. Yet none of these projects is explicitly focused on the problem of how future archival users will be able to make any sense or gather any meaning out of the data they retrieve, if indeed they can discover and access it.

I suspect that as linguists this problem may sit squarely in our bailiwick because humans encode, express, and transmit knowledge over vast lengths of time through our languages and cultures. I will say at the outset that I don't have a solution to this problem and apparently neither does anyone else. However, I think as linguists—especially linguists who create and archive endangered language documentation data—we should be thinking about this. In what follows, I'll look at a variety of linguistic data archives (broadly construed) and for each, ask what we can learn from them. Where have we succeeded in moving knowledge into the future? Where have our efforts fallen short? What will help our data last and be meaningful in the future?

2 Archives of the past

How long do you think the linguistic data you create will last? Do you expect to be able to access and use it throughout your career? Will other researchers? How about the next generation of scholars? What about in one hundred years? Five hundred? One thousand or more?[4]

I came of age as a researcher at the end of the late paper-record era, before data practices were digital or even digitized. Also, for many people creating language documentation as I was, there weren't obvious destination archives for the data once collected. The main venue for sharing the data was through analysis and publication, and those products tended to include only illustrative examples that proved or disproved a particular theoretical point. The vast majority of collected data for many researchers remained unpublished and inaccessible to the research community.

Therefore, I could tell you how long my data would probably last—it was however long it would take for the paper to molder on my bookshelves. Or less time than that, in the case of my tape cassette recordings. Some of those had become unusable within a matter of a few years. Thankfully we have made great strides toward addressing this problem as a discipline. We are creating a culture of data management and archiving that is built into our data creation practices; there are now many available language archives, and we are expected to identify an archive as the destination for our data before it is even created. There is no longer any good excuse for allowing your data to molder on your computer hard drive or office bookshelves.

However, it is another question how long your data *should* last. The data I generate are primarily endangered language documentation, and I want that data to last as long as possible—for whomever might have need of it, but especially for the language communities themselves and their descendants, because they have the strongest connection to the information contained within that data, and the greatest stake in its future.

But again, how long is long? My organization, the Long Now Foundation, likes to take what is for most people an absurdly long frame of reference—the last ten thousand years and the next ten thousand—and think about it as a human-actionable time frame. We call this the "Long Now." We even build "artifacts for the future" that are meant to last and be meaningful to humans for millennia.[5] If we adopt this as our frame of agency, can it help us think about problems in the here and now? What would it mean to take responsibility for our data, ensuring it could last and remain meaningful for the next ten thousand years?

Probably few of us would imagine that the data we archive could last or be as important in the future as famous examples that enabled the decipherment and discovery of ancient languages and cultures, such as the Rosetta Stone. At the same time, the archives we are creating could collectively be seen as just as important because they may be the only archival data available about these languages in the future. What can we learn from "accidental" linguistic archives that have held so much value for the future? If the archival data we are creating today were to be viewed from an equally distant future, would any of it remain, and what meaning if any could be derived from it?

The Rosetta Stone was not created as an archival object. It was not even created as a unique object, as copies were housed in temples across Egypt (British Museum, n.d.). It is just the one copy that chanced to survive. "Lots of Copies Keeps Stuff Safe," also known as LOCKSS, turns out to be a useful strategy for long-term archiving and one used in modern digital preservation systems (Stanford University, n.d.).

The Rosetta Stone is made of granodiorite, an igneous rock that today is either crushed and made into roads, or used for ornamental building materials. Indeed, when the Rosetta Stone was found by Napoleon's soldiers, it had been reused as building material in Fort Julien near Rosetta (Rashid) in Egypt. Having reuse value can, surprisingly, occasionally work in the favor of long-term preservation of information. Another noteworthy example of this is the Archimedes Palimpsest, a thirteenth-century prayer book with text that overwrote at least seven Archimedes treatises written in the tenth century, two of which exist nowhere else (Archimedes Palimpsest, n.d.).

Despite the advantages of modern digital data creation and archiving, I should point out there is potential value for data stored in physical formats and in it being analog. If I had a gold coin for every time someone has suggested to me that inscribing information into stone would be the best means of preserving it for the long-term, I'd be a very rich lady. Of course, these suggestions have a valid point, as inscriptions in stone can be very robust and can withstand quite a bit of abuse or neglect as did the Rosetta Stone. Do we have anything in the digital realm that can compare? Or even anything in the digital realm that can compete with information on paper kept in an acid-free environment, which could potentially last five hundred years?[6] Inscribing your field notes onto large slabs of granodiorite isn't very practical or cost-effective, nor are copies, unless you happen to have the resources of Ptolemy V. Related to the longevity of analog formats, note the information on the Rosetta Stone degraded gracefully, rather than catastrophically (except for the part that broke off). It is much more likely that digital data will fail catastrophically, as with a corrupted file.

The Rosetta Stone is also relatively unencumbered by encoding. One has the primary encoding of the message in human language and then the secondary encoding of that language into writing. But the additional layers of encoding required by a digital file or by translating a digital file into other formats such as the nucleotides of DNA could serve as serious barriers to decoding the information in the future. In comparison, all we needed to do to figure out the Rosetta Stone was look at it (the writing was small, but human-eye visible) and then learn how to read it. We should at least be as kind to the future.

Moving from aspects of archival format to content, it is instructive for our primary question here of "how do we move human knowledge into the future" that the textual content of the Rosetta Stone isn't of particular importance today. The value of the artifact isn't in its message (which was a decree) but rather how that message was presented. The nearly same content was written in three different languages and writing systems: Ancient Egyptian hieroglyphs, Demotic (a script used for writing a later stage of Egyptian), and Ancient Greek. The translations served a symbolic as well as practical purpose at the time the stones were inscribed and erected: hieroglyphs were appropriate for a religious text, Demotic for a decree, and Greek as the language of the people. It was the parallel format of the multilingual text that was key to enabling the decipherment of hieroglyphs.

It turns out that we may be creating some future Rosetta Stone–like data in our modern documentation practices today. An example of parallel data from modern fieldwork practices is interlinear glossed text, where text from the language being described is provided with word and morpheme translations as well as a free translation in another, usually more widespread, audience language. Because the creation of glossed texts is a part of the process of linguistic analysis and the development of lexical and grammatical resources, the practice of developing interlinear glossed text, particularly time-aligned with an audio or video recording, is a core activity of language documentation and description. Besides being very practical in the here and now, it could be that the parallel interlinear glossed text we collect will be key resources for the future, so long as either the source or translation languages remain accessible.

Another type of parallel data that linguists create comes from the practice of collecting a Swadesh vocabulary list. A Swadesh list (as it is commonly known) is a vocabulary elicitation tool created by the linguist Morris Swadesh in the mid-twentieth century. Its intended purpose was to generate data for the study of glottochronology, which aimed to determine the rate of lexical change in language. It was also a tool for lexical comparison between languages to develop hypotheses of language relatedness. Swadesh developed several versions of the list and finally settled on a list of one hundred basic concepts commonly expressed lexically across the world's languages. When the Swadesh list is used in fieldwork today, it is generally in early lexical elicitation. The Swadesh list fell out of use as a research tool for many decades when the theory of glottochronology was deprecated. Over half a century later, however, the theory and use of Swadesh data collections were revived for study using methods of computational analysis (Wichmann et al. 2010).

The Swadesh list is an example of a data type that derives its parallelism by virtue of being collected by many different researchers for many different languages using the same templated structure. For lack of a term, I'll call this *exocentric parallelism*. Data sets built this way are typically the result of a coordinated activity of a group, rather than the product of any one researcher. It represents the intellectual and cultural infrastructure of a field of study.

Another example of parallel data creation is a text translated into many different languages. Short parallel texts of this type are often created for practical purposes, such as ballots and drivers' tests (or in published decrees, like the Rosetta Stone). Religious texts are often translated across languages, and translations of the Bible alone probably constitute the single largest exocentric parallel text collection in the world (Wycliffe Bible Translators, n.d.). In the domain of non-religious texts, translations of the Universal Declaration of Human Rights (interestingly, a modern kind of decree) exist for several hundred languages and are even a showcase project for the Unicode Consortium (Unicode, n.d.).

The *Pear Film* is another example of an elicitation tool intended for translation into many different languages (Chafe, n.d.). It is a short film of about six minutes in length that doesn't have any conversation or narration, rather the characters act out a series of events. The viewer then paraphrases the action of the film in their own language. The *Pear Film* was used to study narrative structure across languages, and while the parallelism is based on a shared target for translation, the translations themselves may vary considerably in structure and lexical choice.

Another good example of parallel text collection is the practice by phoneticians of transcribing the fable "The North Wind and the Sun." This text has been translated and transcribed for many languages, and they are published as illustrations of some of the language descriptions in the *Journal of the International Phonetic Association*.

Aside from very long translations, parallel data sets do not typically contain a great deal of meaningful content in any given language, in and of themselves. A single Swadesh list, for example, will not provide much information about the people who used the language and how they experienced the world. Neither did the Rosetta Stone decree. Rather, it provided the means of decipherment of a much larger corpus of existing texts. And this, in turn, unlocked the experience of an ancient civilization as recorded by them in text form, some part of which is available to us today.

Another famous example of an "accidental" linguistic archive (or historical archive with linguistic import) is the very large corpus of Hittite texts discovered in 1906 by Hugo Winckler at an excavation in Boğazköy, Turkey, being the ancient archives of the Hittites at their capital Hattusas (Sturtevant & Hahn 1951). For linguists who are not experts in Hittite, its discovery represents less an example of epic decipherment (although that in and of itself is truly impressive) and more an example of a linguistic theory that was epically proven when the evidence of Hittite emerged, because it displays certain archaic features of reconstructed Proto-Indo-European that other extant languages had lost.[7]

The texts were written on clay tablets using a Babylonian cuneiform script. Like many forms of written language around the world, writing systems are far more commonly borrowed and adapted than uniquely created.[8] This adaptation both helped and hindered understanding of the spoken language. It helped that the cuneiform writing system was already well understood from its use with many other ancient languages well represented in the archaeological record. It was a hindrance in that the writing system used was (like Ancient Egyptian hieroglyphs and Maya glyphs) a combination of ideograms and a syllabary, and the ideograms often represented non-Hittite Sumerian or Akkadian word signs. The adaptation also illustrates a common problem when a writing system of an unrelated language is adopted to represent another: Hittite had consonant clusters that weren't well suited to a syllabary, and various strategies had to be employed by scribes to make everything work.[9]

The puzzles of decipherment left by these ancient artifacts may seem like a problem of the past, but consider that today our documentation and linguistic analysis is still very much text-dependent, and for glosses and translations, we typically use modern writing systems that represent many of these same issues: English writing is widespread as a language for translation, but already represents a wide gap between its alphabetic spelling and its pronunciation. Japanese is a major world language but represents many of the same compromises as Hittite in adapting Chinese ideographic writing to fit its non-Sinitic grammar and frequent use of loan word vocabulary.

These linguistic records from the deep past show us possible strategies for building long-lasting, long meaningful data collections. One of these strategies is an expansive and varied corpus, and the more in each of these dimensions, the better. As "accidental" linguistic archives, they represent the kind of information that the effort of writing was reserved for: bureaucratic records of trade and proclamations. But occasionally we are rewarded with treasured glimpses into wider culture: recorded rituals, prayers, recipes, poetry, legends, historical accounts, and even fascinating procedural texts.[10] Because we are purposeful creators of corpora, we can think about ways we think the material will be used in the future, and while constraints of time mean we must still pick and choose, we can pay special attention to those areas of culture and language that seem most valuable and unique.[11]

Another important strategy for data longevity is providing tools to decode whatever layers of encoding may exist. In this respect, modern digital archives are far more complex than any of these ancient artifacts. To see this, imagine a scenario from the not-too-distant future where language data are written for storage in DNA (this technology is available now and could be much more widespread in the near future). Say we wanted to encode a simple message such as "The quick brown fox jumps over the lazy dog" in DNA. First, we have the text as written, and/or transcribed in the International Phonetic Alphabet. This glottographic writing is the first layer of encoding we have introduced. Then we need to get it into digital text form. The relationship between analog writing and digital writing is highly complex, especially if you want to be thorough about it—witness the extensive Unicode Standard (printed out, it would be about 1,500 pages long) (Unicode 2017). So, in the transition from analog to digital writing we have introduced another layer of encoding. Then we have to go from the representation of text in binary 1s and 0s and map these onto the nucleotides of DNA (hopefully standard practices for how to do this will emerge by the time the technology becomes widespread).

Now, for fun, imagine you discover such an archive three thousand years from now. Perhaps you found it by sampling the DNA of a de-extincted passenger pigeon genetically modified to have a florescent pink tail feather as a marker of the archival data it contains.[12] Then you'd need to work back through all of the layers of mapped encoding. From the ACGT of DNA to binary 1s and 0s. Then you would need to know that the data string was encoded text, and that the Unicode text was the glottographic rendering of some form of language.[13] Given how bad humans are at producing, much less reading instruction manuals, the fact that you have gotten this far seems pretty far-fetched.[14] Nevertheless, you succeed and obtain the string "The quick brown fox jumps over the lazy dog." Now, what on postapocalyptic Neo-Earth does it mean?

Let's go further with our scenario and imagine this string is the Rosetta Stone decoding key that unlocks a corpus of ancient texts in the English language (a corpus that also provides attestation for the use of archaic letters X and Q). You even find a Basic English lexicon and grammatical sketch that help provide referents and uses for most of the words in your short text string. You might even use your linguistic sleuthing skills to figure out that the sentence is a pangram. From there, would you guess at its use in ancient typography or its cultural import in students trying to master typewriters and other keyboard text-entry tools? Would you suspect that the sentence had moral import? Or not having any context, would you take it more literally and think that it was just about a fox and a hound?

Without a great deal of other information or access to a native language user, you would be hard-pressed to know for certain. No archival data completely document a language, much less the experiences of a people. To be sure, some languages are much better documented than others, but for most endangered languages, the best of our efforts will still leave a thin, incomplete record, and if the history of these ancient linguistic artifacts is any guide, the record will only become more fragmentary with the passing of time.

3 Archives of the present

Compared with "accidental" archives of the deep past, there is reason to expect that data created by more recent language documentation projects—ones in the last 150 years or so—would be better equipped to move human knowledge as expressed in language into the future. After all, these were efforts to purposefully document and study language in all its variety; many of them done with the awareness that the languages being studied were in danger of falling out of use, and the linguistic record being created might be the only record of them available in the future. Also, archival records of the past 150 years are much closer to our own time and understanding. We have a sense of continuity with them unlike with the records of the deep past where greater gulfs of difference exist.

As a graduate student in linguistics at the University of California, Berkeley, I had the opportunity to work in a language archive, the Survey of California and Other Indian Languages (also known today as the California Language Archive, or "the Survey" by those affiliated with it). This was during the time that the Master-Apprentice and Breath of Life programs were first being developed, and the Survey was very much a part of both, as it provided access to critical source material for those working to revitalize critically endangered languages, or languages that are no longer in active use and the archival record is the only documentation descendants have for the purpose of bringing the languages back as lived languages once again (Advocates for Indigenous California Language Survival, n.d.a,b).

Like other archives of its generation, the Survey came into being first as a place to put the growing collection of field notes created by linguistic researchers working on language documentation. In the case of the Survey, these were the students of Mary Haas, who had set them to the task of documenting as many of the languages of California as they could.[15] They very much realized that the languages were passing out of use and that time for documenting them was critical. When I started graduate school in the 1990s, the Survey collection was housed in an office where several of us had our work desks. As former students retired, or passed away, boxes and boxes of field notes would arrive and we would stack them wherever we could (sometimes even on our work tables) until we could find the time and place to properly catalog and shelve them (a glimpse into our own futures, as now many of my cohort have our own collections of language documentation in the Survey). While archiving was fairly ad hoc for many years, the Survey has gradually been developed into an exemplar of a modern regional language archive, with a fully digitized collection and web-accessible finding aids.

Part of the collection in the Survey contains elicitations and other source material for grammatical sketches of languages across California, and these were often worked up and published as doctoral dissertations. However, a large part of the collection were stacks of shoeboxes full of notecards containing lexical data for the preparation of dictionaries or notebooks full of transcribed texts (some with accompanying audio recordings) intended to be eventually published as annotated collections of texts. Many of these were expertly collected and carefully kept, but never published. Thus, a tremendous amount of source data and language description in manuscript form remains to this day the primary documentation that exists for many, many languages. Other archives of this type created for regional language documentation are the Alaska Native Languages Center, the Archive of the Indigenous Languages of Latin America, the Pacific and Regional Archive for Digital Sources in Endangered Cultures, the Native American Languages collection at the Sam Noble Museum in Oklahoma, and most recently, the Kaipuleohone Language Archive at the University of Hawaiʻi.

While these archives all continue to build their collections with language documentation created by new faculty and student research, the bulk of their collections and a great deal of their value is in the collections they house that are now between a half century and a century old. This amount of time provides a good distance for us to evaluate these language documentation collections with our question in mind: Have they been able to move human knowledge forward, and if not, what gaps exist?

I expect it is common for anyone who works with manuscript or other historical language documentation to come away humbled by the experience. It is a stark reminder of how your own data may be viewed or experienced fifty or a hundred years hence. It is also a powerful reminder that when creating endangered language documentation you are providing an essential record for the future, as no other may exist. I provide a few examples from my own experience of being "humbled in the archives" that illustrate different scenarios of data use by future stakeholders: linguists, heritage language community members, and humanity as a whole.

Future linguists. Primary future stakeholders for archived language data are future linguists, and linguistic theories might be advanced or argued against based on archival language data. Having myself worked on the Potawatomi language (Neshnabémwen) for many years, I saw several iterations of morphological theories being worked and reworked based on complex but orderly Potawatomi verbal inflectional morphology. Doubtless, morphologists will continue to test their theories on it into the future, although the full paradigmatic record is fragmentary (Lockwood 2017). I myself was never so persuaded by the explanatory power of an elegant theory applied to data as I was when trying to explain verb stem alternations found in Miwok languages of California to participants of the Breath of Life workshop being held by the Survey. The Miwok languages are some of the best documented languages of California, thanks in large part to the work of linguists Catherine Callaghan and Silvia Broadbent, both students of Mary Haas. However, their grammatical descriptions required complex statements about the relationship of stem class alternations, patterns of consonants and vowels that we today recognize as being non-prosodic templatic morphology, as found in Semitic languages such as Arabic (McCarthy & Prince 1990). Not only has later linguistic theory improved our understanding of Miwok languages, we now have another language group that exemplifies templatic morphology, strengthening the case for its explanatory power.

Heritage language community members. As I mentioned, I worked on the documentation of critically endangered Neshnabémwen for many years, both with native speakers, as well as with a large amount of historical language documentation created by a succession of Jesuit priests and later by the linguist Charles Hockett who worked with fluent speakers in the 1940s. Both, but especially the latter, provided the basis for eliciting complex verbal morphology that is attested for the most part, but not in its entirety, today. Whether this is the result of the extreme contraction of the language-using community within the last century or regular language change is unknown and perhaps at this point unknowable. Attestations of the verbal paradigms provided by fluent elders over the last three decades as part of modern language documentation efforts are being used today for language revitalization activities. Perhaps, in time, new generations of users will go back to the older records of the nineteenth and twentieth centuries to explore and possibly incorporate some of the broader paradigms into their own usage. It is an available option only because those archival records exist.

Another example from the Neshnabémwen archival record relates to the recording and passing on of extra-linguistic knowledge, where I fear a great deal is being lost when languages are no longer used. Linguists today often document ethnobotanical knowledge as part of larger language documentation projects. I never focused on this with Neshnabémwen, partly because I have no talent for it, and also a fluent elder I was working with was very knowledgeable and wrote and published on it himself (Thunder 1996). I did find an extensive ethnobotany collected by Huron Smith who worked in the 1920s with many of the native tribes of Wisconsin including the Potawatomi (Smith 1933). He would have been working at a time when the language had many more native speakers than there were at the time of my field research, when there were about fifty speakers in total. I showed the work to the fluent elder who was a knowledgeable herbalist and remarked that despite the otherwise copious detail, many of uses of the plants were simply labeled as “medicine” with no further information. We speculated why the Smith record was so vague about information that would seem so beneficial to future generations (at the time of our discussion, many of the plant names and uses recorded by Smith were no longer known). One possible reason, we thought, is that the person who identified the plants considered the knowledge to be too sensitive to commit to publication where readers who had no direct connection might inadvertently misuse the knowledge, potentially causing great harm to themselves or others. It would be irresponsible to disclose the information this way, even if it meant the knowledge would be lost.

This example illustrates the detailed encyclopedic ethnobotanical knowledge of the world’s ecosystems contained and communicated through human languages. The loss of this knowledge is only a part of the broader set of knowledge we are losing when languages cease to be used. In the case of Neshnabémwen and other language communities that are striving for language maintenance and revitalization, there is a conduit for the continuity of knowledge through lived communication and practice. In the next example, that conduit was largely severed, and while we may never know the magnitude of the loss to humanity, we have evidence that it was great.

Humanity. One day in the Survey, a group of well-trained linguists sat puzzling over a text. We were a small working group of professors and graduate students attempting to develop an annotated corpus of the texts told by Ishi in his Yahi language to the linguist Edward Sapir in 1915 (Ishi & Luthin 1955; Hinton et al. 2001–2002). The texts were expertly transcribed by Sapir and given running translations in English. We also had access to published resources in the closely related Yana languages including a dictionary (Sapir & Swadesh 1960) and grammatical sketch (Sapir 1922). We found we were able to provide word and morpheme glosses for most of the texts, but sometimes certainty about the meaning of a passage simply eluded us. The story was wonderful, about the original human quest for fire, and there is a similar story told by the Yana (Sapir 1910:23–34). In the passage, the grizzly bear ties his hair into a top knot, and (we think) wafts up in the smoke and ashes of the fire until he reaches a sky hole (we have the unanalyzed string glossed as "penetrated through hole in sky"). He pops through the sky hole and lands (presumably) on the floor of the sky—the next line literally reads "he sat down." There he sits and looks out to the four directions, finally spotting fire in the far South (Ishi & Luthin 1955:237–238). This was as close of a translation as we could get. There were many such passages in the corpus, although this one nearly twenty years later is one that really stands out in my memory. There was no way to learn more. We had all of the records of related Yana languages and used them wherever we could. There was no one to consult to learn more. Ishi was the last surviving speaker of Yahi, other Yana languages subsequently ceased to be spoken, and the records of Yahi made by Edward Sapir are the only ones that exist.

With respect to extralinguistic knowledge, Ishi was an expert archer. There is good evidence that he was a specialist in the making of bows and arrows, and in hunting with them, and one of his stories "Tale of Lizard" is embellished by a loving account of the craft of arrowmaking (Ishi & Luthin 1955:2–68). His skill was noted by Saxton Pope, who was Ishi's physician. They developed a friendship, and Pope learned arrow-making and hunting techniques from Ishi, and after Ishi's death became an expert archer himself, carrying on Ishi's legacy. Pope would later write *Hunting with a Bow and Arrow*, now a classic work on archery, and would go on to become a major popularizer of bow hunting in the twentieth century (Pope 2000). If you practice archery today, there is a strong chance you are practicing skills transmitted through a direct line of knowledge and practice that can trace its source to Ishi.[16] We only have a fleeting glimpse of Ishi's mastery of archery in the brief linguistic record we have of his time working with Sapir. What other encyclopedic knowledge have we lost forever? These examples hopefully serve to illustrate the depth of knowledge practiced in cultures around the world and communicated across generations through languages, many of which are highly endangered. And is not this, collectively, the knowledge we have attained as human beings about how to live—and hopefully thrive—in the myriad environments on planet Earth over the past millennia?

Turning to archives created more recently, it is worth taking a look at another major kind of language archive: those that were created to house the linguistic data from grant-funded endangered language documentation projects. There are only two major such archives in existence (would there were more), and these were developed at the beginning of the twenty-first century with substantial funding from philanthropic sources. This funding not only provided grants for endangered language documentation projects that were and are taking place around the world, but tools and infrastructure (such as archives) to support the research as well.[17] The efforts have been very important and influential in the field. The primary two are the Endangered Languages Archive, which was established in 2002 alongside the Endangered Language Documentation Program funded by the Arcadia Fund (SOAS, n.d.), and the Language Archive, which was created alongside the Dokumentation bedrohter Sprachen program that was funded by the Volkswagen Foundation starting in the year 1999, which has since moved to the Max Planck Institute for Psycholinguistics (Max Planck Institute for Psycholinguistics, n.d.).

Unlike the Survey and similar archives, the materials in these archives were all "born digital." Likewise, the archives were purpose-built to house and serve digital resources rather than physical ones, sidestepping the large task that most regional archives have had of digital conversion. Because they were developing digital infrastructure such as tools for language documentation and had close partnerships with their associated archives, they were able to develop project workflows and metadata schemes that structured digital resources from the point of data creation to eventual archival ingestion. They were drivers of innovation and were central to the broader initiatives that structure archival practices to this day.

One example of this is with the metadata schemes we use to describe language resources. The Dokumentation bedrohter Sprachen project required that its projects use the ISLE Meta Data Initiative scheme, a broad and detailed set of resource descriptors.[18] This metadata scheme is used today by both the Language Archive and Archive of the Indigenous Languages of Latin America. An alternative and simplified set of descriptors was developed by the Open Language Archives Community

(OLAC). As part of the Open Archive Initiative, OLAC sought to make language resources discoverable across otherwise siloed digital language archives. Because the ISLE Meta Data Initiative metadata set can be mapped to OLAC, the initiative was able to create a central metarepository of language resources (OLAC 2011).

The Digital Endangered Languages and Musics Archives Network (DELAMAN) was also established at this time and has created a community and network for the coordinated development of language archive infrastructure (DELAMAN, n.d.). All of the archives discussed here are represented in DELAMAN, and DELAMAN has adopted the OLAC metadata standard to represent all of its member archives. It is therefore likely that DELAMAN will play a key role in the future development of OLAC.

Other projects such as Electronic Metastructure for Endangered Languages Data sought to create tools and refine practices for digital data collection and the stewardship of electronic resources (EMELD 2010). One tool that was developed, the GOLD ontology (General Ontology for Linguistic Description), proposed a taxonomy of morphosyntactic descriptors that could be used to describe the morphosyntactic properties of any of the world's languages. If linguists mapped the morphosyntactic features of the languages they were documenting to the GOLD ontology, a GOLD-driven search function could find instances of the use of any particular feature across the linked data sets.[19] The GOLD ontology has not gained traction as a practice for language documentation research, but it still represents a tantalizing view into a future where linked language data are not siloed in archives but are discoverable and harvestable across them.

4 Archives of the future

If we extrapolate from archives of the present and their current development efforts, we can speculate about their near-term future, say the next ten to one hundred years, and have a reasonable expectation of being accurate, at least in part. Next, I offer a few prognostications centered on what I think will be areas of language archive focus in the future.

Digital focus. One area of speculation relates to the challenge of maintaining digital language resources. The physical archives of the recent past can withstand a bit of benign neglect. Paper can be left in boxes on bookshelves or in attics for decades. While this is not ideal, it certainly has frequently happened and the paper was later ingested into archives and data recovered from it.[20] We have wax cylinder language recordings from the early twentieth century that could not even be listened to until recently because any playing of them would further degrade the audio quality. Now thanks to new technology that reads them optically they can be remastered without damage (IRENE, n.d.).

However, this is not the case with digital resources, at least not so far in our experience. Digital resources are near-constantly being moved into the future—migrated onto new storage media, or new file formats, or new metadata formats or new content management systems. While the migration of any one digital resource may not be a significant task, it certainly is for an archive to manage the forward migration of all of its digital assets, and funding for most archives is not assured indefinitely. Some of the archives discussed here have experienced significant funding disruptions in the past two decades of their existence. Fortunately, these were of short enough duration that the digital records were preserved. We should all bear in mind though as we commit precious language documentation to a digital future that there is no good "lack-of-funding-model" for digital resources. Language archives must be supported into the future to ensure the digital longevity and future digital access of the data we are creating today.

Community focus. Another area of speculation relates to who the primary users of archived endangered language data will be in the future. I believe that the next stage of archival development will come from a strengthened focus on heritage language communities, and the use of archived language data for the purpose of language revitalization. The regional archives have a head start on this by virtue of their history, and regional archives already typically have a close working relationship with the language communities represented in their region. Activities like the Breath of Life workshop that started at the Survey are now taking place in many regional language archives, and there is even a National Breath of Life in the United States so that participants can access and learn about resources housed in government archives. These kinds of activities foster community across language boundaries and create new centers of shared innovation.

As languages are revitalized and reawakened, we should expect that the digital resources they create should be archived as language resources as well. If an archive has a strong relationship with a language community, the community could have a reasonable expectation that it would be able to archive its resources alongside the historical ones, should that be desired. Otherwise, the existing archives ought to play a supporting role in helping those communities establish their own as part of the larger language archive community, should that be desired. Alternatively, one could imagine a parallel network of revitalized language archives, hopefully that are not siloed from existing language archives or from each other.

Archives such as Endangered Languages Archive and the Language Archive, which were developed as part of grant-funded collection efforts, don't themselves have this same direct relationship with language communities, although their individual language documentation projects do. Grant-created archives, to the extent that they wish to lead in this area, will need to find other ways to build community. One way you could imagine them doing this is by becoming training centers for community-based linguists, who would in turn lead community-based language revitalization efforts.

Legacy data focus. If current trends continue, we expect many more languages will cease to be used in the coming decades. Also, the ability to continue documentation work on them is heavily dependent on funding. As opportunities to document endangered languages wane, I expect there will be a renewed focus on existing legacy collections of language data housed in regional archives. Not just digitizing them, as this has largely already been accomplished, but rather going back to all of those manuscript collections of data and working them up into computationally tractable data sets: rekeying handwritten notes, annotating them, providing them with detailed metadata, and hopefully incorporating them into new research and publications.

Computational focus. As many of the world's languages disappear, archives will become the only repository of a great deal of language data, and theoretical claims as well as hypotheses as to what is possible cross-linguistically will have to be tested against all of it. This means we will need more application programming interface (API) access into collections of data,[21] and we will want to prepare and expose primary data to these APIs so that we can search across archives and collections and conduct research from any location while accessing the entire worldwide corpus of archived language data. This may not happen in the next decade, but I hope to see good progress on it in my remaining lifetime. I expect that most of the work involved will be in making computationally tractable data sets, and this will require the development and accommodation to shared standards—if not the GOLD ontology, then in resources like it.

Related to the creation of computationally tractable data sets from legacy resources, I expect that language archives will become leaders in the creation of corpora and other natural language processing resources to better enable the world's languages to be used in electronically mediated communication. As I have argued elsewhere, there is economic motivation for enabling only a fraction of the world's largest languages in this important new domain of language use in the modern world. To participate, smaller language communities will have to bootstrap themselves by creating corpora and tools for their language using natural language processing (Buszard-Welcher 2018). They will need support to do all of this, and I can't imagine a better partnership to accomplish it than with the language archives of the world.

Training focus. Besides the training and support activities discussed, I believe that language archives will become central for the training of linguists, who will be needed in all of these activities. This requires retooling for many traditional linguistic departments so that linguists will have access to applied specializations alongside theoretical ones. To a certain extent this has already happened with the renewed focus on documenting endangered languages, but more specializations are needed—in archiving and information science, in natural language processing, in corpus linguistics, in programming and building APIs as well as archival software and tools, in language revitalization and community-based linguistics—as well as ways of applying all of these skills to the continued effort of endangered language documentation.

It is worth thinking about the role of linguistic archives, and who we expect to be primary communities of use for linguistic data archives in the future. The answers to these questions will undoubtedly be essential to the continued operation of language archives, and as we have argued, language archives need continued support for the linguistic data we create to continue to exist.

5 Conclusion

Returning to the question posed at the beginning of this chapter, How do we move human knowledge into the future?, I think there are a few aspects to its answer. One aspect is in how thoroughly we will be able to document endangered languages while there is still time to do so, and what that documentation contains by way of representing the knowledge and culture of its users. Another aspect lies in our practices of moving that information forward in time: these are our data creation and management practices, our archival practices, and our practices as a discipline and society in committing to the development and forward migration of archival records in the long term. The third aspect is whether in the future, and perhaps distant future, the users of the data we create will be able to obtain any meaning from them, because the passing on of knowledge is heavily dependent on this.

If a primary goal is to preserve human knowledge as expressed in language, then lived language is the primary "mode" of that knowledge—it is also situated, embodied, and encultured. Any recording of it strips some, or much, of this away to an audio or audio visual signal discontinuous from any lived linguistic event. With the passing of time, it increasingly becomes unsituated, disembodied, and un-encultured. Ideally, we enrich archived information both with annotation and metadata that help ground it in its historical context. But still, even the best of it is a simulacrum.

Both modes—lived language, and extracted and archived language—are potentially archival in the long term. Lived language would seem to be the best way to preserve meaning, but it is also precarious and ephemeral as evidenced by the vast majority of threatened and critically endangered languages around the world today. Archiving the audiovisual signal and annotations has its own precariousness but we are getting much better at it. It does not preserve the richness of meaning that lived language does, but it is an important record of it. And in some cases, with considerable effort, records of a language have enabled languages to be lived and meaningful again (Leonard 2007).

So, if a primary goal of our long-term archiving efforts is preserving human knowledge, and that is best ensured by preserving lived languages, then what is the role of archived linguistic data? Can we enrich our data, or our archives, so they are better repositories of knowledge? Could archives be critical infrastructure for lived languages, so that languages are enriched by their archives, bucking them up, or providing a kind of insurance policy against the forces of attrition and obsolescence? Could we make archives part of lived culture itself? If these ideas seem audacious, bordering on the incredible, the first step needn't be, although I expect it is still controversial: we could start by reframing archived endangered language data as archived human knowledge—knowledge that is part of all our shared heritage and needed for our common future.

Notes

1. The phrase "Long View" is a reference to a classic work of scenario planning by Peter Schwartz (1991).

2. For example, documenting aspects of language use that are changing or falling out of use due to attrition.

3. See the Rosetta Project (n.d.) and its "future artifact" the Rosetta Disk.

4. To help think about this question, see Mattern (chapter 5, this volume) on the linguistic data life cycle and Kung (chapter 8, this volume) on how to develop a data management plan.

5. The Rosetta Disk is one such future artifact (Rosetta Project, n.d.). Another is the 10,000 Year Clock, currently being built inside a mountain in the desert of West Texas (Long Now Foundation, n.d.).

6. For an expert conversation on this subject see the Time and Bits Workshop, held in the year 2000 at the Getty Institute (MacLean & Davis 2000).

7. This theory was originally postulated by de Saussure (1879) as a set of reconstructed "coefficients sonantiques" whose presence accounted for certain alternations found in Proto-Indo-European roots. Later, deciphered Hittite data was shown to have reflexes that correspond to two of these abstract elements (see Kuryłowicz 1927 and also Sturtevant & Hahn 1951:47–49 for a discussion). It has become a canonical example to demonstrate the value of internal historical reconstruction (for example, Hock 1991).

8. See Sampson (1985) for a discussion of the origins of many of the world's writing systems.

9. See Sturtevant and Hahn (1951:14) for a discussion of these strategies.

10. See for example the Hittite texts by Kikkulis of Matanni on the training of race horses, with methods still apparently employed today (Sturtevant 1951).

11. A note about this selection principle—in the mid-twentieth century, language documentation projects frequently focused on "high cultural-value" texts such as myths and legends, or narration, to the exclusion of other forms of language use. See Buszard-Welcher (2003) for an example of this where this practice led to obscuring the basic grammatical patterns in use in everyday language. Linguists are probably not the best curators, and language users and language communities can provide guidance for what they are most interested in documenting.

12. More likely archival data in DNA would be stored in vitro rather than in vivo, but we are imagining here, and so need not be prosaic. Also, there are published experiments of data written in DNA in vivo, into the *Escherichia coli* bacterium (Shipman et al. 2017).

13. Given modern language documentation practice, this could be a variety of media file types for recorded audio or video and their accompanying transcriptions. Still, the de-encoding problems exist no matter the file type.

14. I say this somewhat in jest, but there is a real challenge in the long-term archiving of contextualizing information such as metadata alongside the archiving of data themselves.

15. Mary Haas's own field notes have now been accessioned by the American Philosophical Society (n.d.a).

16. There is evidence today that Ishi's archery techniques were not solely Yahi, and it may be that he learned from a Nomlaki or Wintu relative (Kell 1996).

17. Two other programs worth mentioning here are the Endangered Language Fund (Endangered Language Fund, n.d.) and the Phillips Fund of the American Philosophical Society (American Philosophical Society, n.d.b) with associated archives that house the research products of the language documentation and revitalization activities they support.

18. The ISLE acronym within ISLE Meta Data Initiative stands for the International Standard for Language Engineering (EAGLES 2003).

19. For a demonstration of this see ODIN (2016).

20. J. P. Harrington was notorious for leaving his field notes in the attics of the people he worked with (Laird 1993).

21. For our purposes, an API allows data within a data set to be accessed remotely by using structured queries and would return structured data in response to those queries.

References

Advocates for Indigenous California Language Survival. n.d.a. About the advocates. https://aicls.org/about-the-advocates/. Accessed March 31, 2019.

Advocates for Indigenous California Language Survival. n.d.b. Breath of Life Institute. https://aicls.org/breath-of-life-institute/. Accessed March 31, 2019.

American Philosophical Society. n.d.a. Mary Rosamund Haas papers. https://search.amphilsoc.org/collections/view?docId=ead/Mss.Ms.Coll.94-ead.xml;query=haas;brand=default. Accessed March 31, 2019.

American Philosophical Society. n.d.b. Phillips Fund for Native American Research. https://www.amphilsoc.org/grants/phillips-fund-native-american-research. Accessed on March 31, 2019.

Archimedes Palimpsest. n.d. The Archimedes Palimpsest: About. http://archimedespalimpsest.org/about/. Accessed March 31, 2019.

Arch Mission Foundation. n.d. Humanity's backup plan. http://www.archmission.org. Accessed March 31, 2019.

British Museum. n.d. Everything you always wanted to know about the Rosetta Stone. https://blog.britishmuseum.org/everything-you-ever-wanted-to-know-about-the-rosetta-stone/. Accessed March 31, 2019.

Buszard-Welcher, Laura. 2003. Constructional polysemy and mental spaces in Potawatomi discourse. PhD dissertation, University of California, Berkeley.

Buszard-Welcher, Laura. 2018. New media for endangered languages. In *The Oxford Handbook of Endangered Languages*, ed. Kenneth L. Rehg and Lyle Campbell. New York: Oxford University Press.

Chafe, Wallace. n.d. The *Pear Film*. http://www.linguistics.ucsb.edu/faculty/chafe/pearfilm.htm. Accessed March 31, 2019.

Church, George, Yuan Gao, and Sriram Kosuri. 2012. Next generation digital information storage in DNA. *Science* 337 (610): 1628. doi:10.1126/science.1226355.

DELAMAN (Digital Endangered Languages and Musics Archives Network). n.d. http:www.delaman.org. Accessed December 30, 2018.

EAGLES. 2003. The ISLE metadata standard. https://www.mpi.nl/ISLE/. Accessed March 31, 2019.

EMELD. 2010. Electronic Metastructure for Endangered Languages Data. http://emeld.org/. Accessed March 31, 2019.

Endangered Language Fund. n.d. http://www.endangeredlanguagefund.org/. Accessed March 31, 2019.

Erhard, Manuel, Robert Fickler, Mario Krenn, and Anton Zeilinger. 2017. Twisted photons: New quantum perspectives in high dimensions. *Light: Science and Applications* 7 (3): 17146. doi:10.1038/lsa.2017.146.

Hinton, Leanne, Herb Luthin, Jean Perry, and Kenneth W. Whistler. 2001–2002. Yahi texts, Hinton.015.002. In *Leanne Hinton Papers on Indigenous Languages of the Americas*. Berkeley: Survey of California and Other Indian Languages, University of California, Berkeley. http://cla.berkeley.edu/item/2494.

Hock, Hans Heinrich. 1991. *Principles of Historical Linguistics*. Berlin: Walter de Gruyter.

Interstellar Beacon. n.d. The Interstellar Beacon: Backup humanity. https://www.interstellarbeacon.org/. Accessed March 31, 2019.

IRENE. n.d. Sound reproduction R&D home page. http://irene.lbl.gov/. Accessed March 31, 2019.

Ishi and Herb Luthin. 1955. Yahi texts with interlinear glossing, Luthin.002.001. In *Miscellaneous Papers from the Survey of California and Other Indian Languages*. Berkeley: Survey of California and Other Indian Languages, University of California, Berkeley. http://cla.berkeley.edu/item/1410.

Kell, Gretchen. 1996. Ishi apparently wasn't the last Yahi, according to new evidence from UC Berkeley research archaeologist. Press release, University of California, Berkeley. https://www.berkeley.edu/news/media/releases/96legacy/releases.96/14310.html. Accessed March 31, 2019.

Kuryłowicz, Jerzy. 1927. ə indo-européen et ḫ hittite. In *Symbolae grammaticae in honorem Ioannis Rozwadowski*, ed. W. Taszycki and W. Doroszewski, 95–104. Kraków: Gebethner and Wolff.

Laird, Carobeth. 1993. *Encounters with an Angry God*. Albuquerque: University of New Mexico Press.

Leonard, Wesley. 2007. Miami language reclamation in the home: A case study. PhD dissertation, University of California, Berkeley.

Lockwood, Hunter Thompson. 2017. How the Potawatomi language lives: A grammar of Potawatomi. PhD dissertation, University of Wisconsin, Madison.

Long Now Foundation. n.d. The 10,000 year clock. http://longnow.org/clock/. Accessed March 31, 2019.

MacLean, Margaret, and Ben H. Davis, eds. 2000. *Time and Bits: Managing Digital Continuity*. Los Angeles: Getty Research Institute.

Max Planck Institute for Psycholinguistics. n.d. The language archive. https://tla.mpi.nl/home/history/. Accessed March 31, 2019.

McCarthy, John J., and Alan Prince. 1990. Prosodic morphology and templatic morphology. In *Perspectives on Arabic Linguistics II: Papers from the Second Annual Symposium on Arabic Linguistics* 16. https://scholarworks.umass.edu/linguist_faculty_pubs/16.

Memory of Mankind. n.d. Memory of mankind. https://www.memory-of-mankind.com/. Accessed March 31, 2019.

ODIN. 2016. The ODIN data. http://depts.washington.edu/uwcl/odin/. Accessed March 31, 2019.

OLAC. 2011. OLAC: Open Language Archives Community. http://www.language-archives.org/. Accessed March 31, 2019.

Pope, Saxton. 2000. *Hunting with the Bow and Arrow*. Billings, MT: Sylvan Toxophilite Classics.

Quast, Paul. 2018. Beyond the Earth: Schematics for "Companion Guide for Earth" archival elements residing within geosynchronous orbit. https://www.researchgate.net/publication/327473491_Beyond_the_Earth_Schematics_for_'Companion_Guide_for_Earth'_archival_elements_residing_within_Geosynchronous_Orbit. Accessed March 31, 2019.

Rosetta Project. n.d. The Rosetta Project: Building an archive of all documented human languages. http://www.rosettaproject.org. Accessed March 31, 2019.

Sampson, Geoffrey. 1985. *Writing Systems*. Stanford, CA: Stanford University Press.

Sapir, Edward. 1910. Yana texts. *University of California Publications in American Archaeology and Ethnology* 9 (1): 1–235.

Sapir, Edward. 1922. The fundamental elements of Northern Yana. *University of California Publications in American Archaeology and Ethnology* 13:215–334.

Sapir, Edward, and Morris Swadesh. 1960. *Yana Dictionary*. Berkeley: University of California Press.

Schwartz, Peter. 1991. *The Art of the Long View: Planning for the Future in an Uncertain World.* New York: Currency Doubleday.

Shipman, Seth L., Jeff Nivala, Jeffrey D. Macklis, and George M. Church. 2017. CRISPR–Cas encoding of a digital movie into the genomes of a population of living bacteria. *Nature* 547 (7663): 345–349.

Smith, Huron. 1933. Ethnobotany of the Forest Potawatomi Indians. *Bulletin of the Public Museum of the City of Milwaukee* 7 (1): 1–130.

SOAS. n.d. Endangered languages archive. https://www.soas.ac.uk/elar/about-elar/. Accessed March 31, 2019.

SPIE. 2016. Peter Kazansky: Nanostructures in glass will store data for billions of years. *SPIE Newsroom*. doi:10.1117/2.3201603.02.

Stanford University. n.d. LOCKSS homepage. https://www.lockss.org/. Accessed March 31, 2019.

Sturtevant, Edgar, and E. Adelaide Hahn. 1951. *A Comparative Grammar of the Hittite Language*, rev ed. New Haven, CT: Yale University Press.

Thunder, Jim. 1996. *Medicines of the Potawatomi*. N.p., WI: Self-published.

Unicode. 2017. The Unicode standard. http://unicode.org/standard/standard.html. Accessed March 31, 2019.

Unicode. n.d. UDHR in Unicode. http://unicode.org/udhr/. Accessed March 31, 2019.

Wichmann, Søren, Eric W. Holman, André Müller, Viveka Velupillai, Johann-Mattis List, Oleg Belyaev, Matthias Urban, and Dik Bakker. 2010. Glottochronology as a heuristic for genealogical language relationships. *Journal of Quantitative Linguistics* 17 (4): 303–316. doi:10.1080/09296174.2010.512166.

Wycliffe Bible Translators. n.d. Wycliffe Bible translators. https://www.wycliffe.org/. Accessed March 31, 2019.