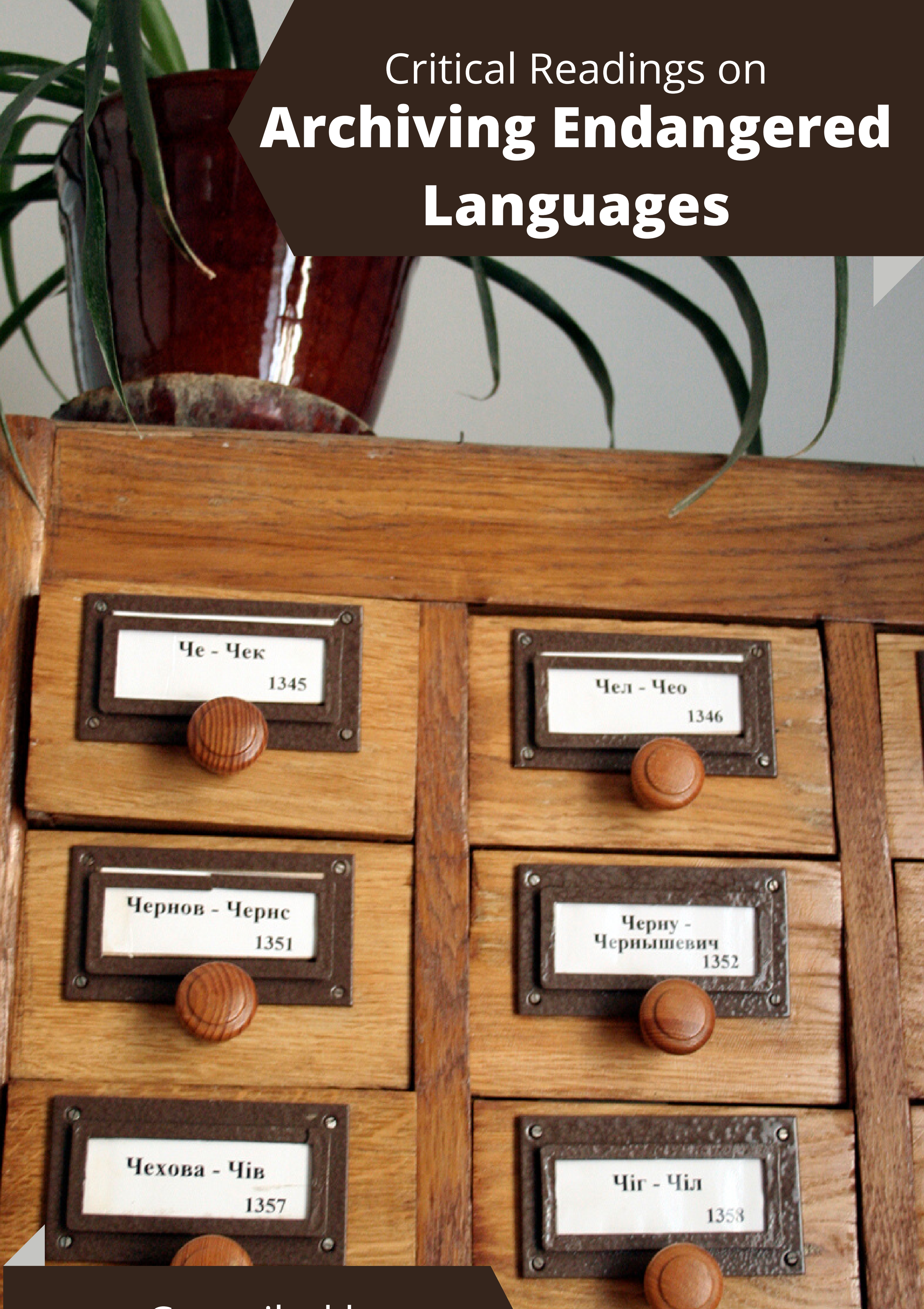
Critical Readings on

Archiving Endangered Languages



Compiled by
Susan Smythe Kung

Critical Readings on Archiving Endangered Languages:

Attributions

Cover attribution:
Cover graphics created by Susan Smythe Kung using Canva Pro Software, www.canva.com.
Cover image: Card Catalog by Serenity Ibsen, licensed under CC BY-NC 2.0.

Article attribution:

Henke, Ryan and 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
Licensed under CC BY-NC 4.0 International

Johnson, Heidi. 2004. Language documentation and archiving, or how to build a better corpus. In Peter K. Austin (ed.), *Language Documentation and Description, vol. 2*, pp. 140-153. London: SOAS. http://www.elpublishing.org/docs/1/02/ldd02_11.pdf
Licensed under CC BY-NC 4.0 International

Seyfeddinipur, Mandana, Felix Ameka, Lissant Bolton, Jonathan Blumtritt, Brian Carpenter, Hilaria Cruz, Sebastian Drude, Patience L. Epps, Vera Ferreira, Ana Vilacy Galucio, Brigit Hellwig, Oliver Hinte, Gary Holton, Dagmar Jung, Irmgarda Kasinskaite Buddeberg, Manfred Krifka, Susan Kung, Miyuki Monroig, Ayu'nwi Ngwabe Neba, Sabastian Nordhoff, Brigitte Pakendorf, Kilu von Prince, Felix Rau, Keren Rice, Michael Riessler, Vera Szoelloesi Brenig, Nick Thieberger, Paul Trilsbeek, Hein van der Voort, Tony Woodbury. 2019. Public access to research data in language documentation: Challenges and possible strategies. *Language Documentation & Conservation* 13: 545-563. http://hdl.handle.net/10125/24901
Licensed under CC BY-NC 4.0 International

Woodbury, Anthony C. 2014. Archives and audiences: Toward making endangered language documentations people can read, use, understand, and admire. In David Nathan & Peter K. Austin (eds) *Language Documentation and Description, vol 12: Special Issue on Language Documentation and Archiving,* pp. 19-36. London: SOAS. http://www.elpublishing.org/PID/135
Licensed under CC BY-NC 4.0 International

Critical Readings on Archiving Endangered Languages:

Table of Contents

Critical Readings on Archiving Endangered Languages: Introduction

Susan Smythe Kung

Welcome to *Critical Readings on Archiving Endangered Languages*. I have compiled this extremely limited set of articles as a way to introduce the reader to the field of endangered language archiving, a subdiscipline of the larger field of language documentation, which is a cross-disciplinary field made up of endangered language speakers, activists, teachers, administrators, linguists, anthropologists, archivists, curators, and Native American and Indigenous studies specialists. Rather than introduce the reader to the larger field of language documentation, about which much has been written, I have focused the scope of this compilation to four articles that help elucidate the context of the practice of archiving the digital and analog materials that have resulted from language documentation during the late 20th and early 21st centuries. The main source of the history contained in this introduction is my own memory, as I have been both a witness to and a participant in the field since the 1980s, first as a non-Indigenous member of a community experiencing language loss, then as a student of linguistics, next as a language documenter, and lately as a language archivist. While other people involved in this movement will have their own perspectives on this history, this one is uniquely mine.

I have chosen to begin this compilation with ***A brief history of archiving in language documentation with an annotated bibliography*** (2016) by Ryan Henke and Andrea Berez-Kroeker because the authors provide, as the title implies, a history of archiving in language documentation. A brief history of any subject is always a good place to start when diving into a new subject or just dipping a metaphorical toe into it. Though the other three articles are arranged in chronological order, that order reflects the chronology of emerging issues in language archiving, and the authors of each article seek to address a particular call to action or direction in which the field was moving during the time leading up to the article's original publication date.

Late in the 21st century, as many of the world's languages became critically endangered, the journal *Language* published Michael Krauss' call to action: "Obviously we must do some serious rethinking of our priorities lest linguistics go down in history as the only science that presided obliviously over the disappearance of 90% of the very field to which it is dedicated" (Krauss 1992: 10). This call to action was answered, and a new discipline emerged, *documentary and descriptive linguistics* (Himmelmann 1998); included in this new discipline was the idea that the products of language documentation must be preserved for and made available to academics for research and investigation as well as to the Origin Communities for language maintenance

and/or revitalization efforts. This language documentation revolution coincided with rapid advancements in digital recording technology and the beginning of the widespread use of and access to the World Wide Web. Digital advances and the Internet made creating and sharing language "data" (the primary and secondary materials that result from language documentation) easier than ever before, but these materials needed to be stewarded responsibly so that they could be preserved indefinitely (Bird & Simons 2003). The second article in this compilation, ***Language documentation and archiving, or how to build a better corpus*** (2004) by Heidi Johnson, was the first article to explain in simple and pragmatic terms just how to do this. Johnson laid out the who, where, why, what, when, and how of getting language documentation data into digital language archives, and the article became a handbook for the early digital language archives. As Johnson points out, there was considerable funding for language documentation at the time, but there was very little instruction on how to do it, and she sought to fill that gap.

Over the next ten years, there was an enormous push to get as much existing language documentation into archives as possible, but in many cases the results were large and poorly organized collections of language documentation data that were too dense to penetrate. Filenames were oblique and directories of related files (or "bundles" of files from "recording sessions") had incomprehensible titles that meant nothing to anyone except the research team that had collected the data; and descriptive metadata was almost non-existent, despite Johnson's (2004) instructions to the field to relentlessly and tirelessly record the metadata. The messy and nearly useless state of the language archives led to the publication of the third article in this collection, ***Archives and audiences: Toward making endangered language documentations people can read, use, understand, and admire*** (2014) by Anthony Woodbury, in which the author calls for a curated approach to archiving language documentation so that these collections would be, as the title says, something that people can read, use, understand, and admire. Woodbury promotes a stick instead of carrot approach to making orderly and admirable archival collections when he argues that academic language documenters would put more care into arranging their collections if they knew those archived collections would be peer-reviewed just like academic articles and books are. While the idea of peer-reviewing has been slow to take off, perhaps because the very people who might be qualified peer-reviewers are also guilty of having created sloppy and/or impenetrable archival collections, the reverse method—the carrot instead of the stick—has taken a foothold in the form of two separate award competitions held on a biennial and annual basis, respectively: The DELAMAN Award[1] and the SSILA Archiving Award.[2]

Some portion of the body of archived language documentation was collected during a time before most academic institutions had Ethics Review Boards (Internal Review Boards in the

[1] https://www.delaman.org/delaman-award/ (accessed July 19, 2020)
[2] https://www.ssila.org/awards/archiving/ (accessed July 19, 2020); note that the SSILA Archiving Award is in honor of Michael Krauss (1934-2019, https://en.wikipedia.org/wiki/Michael_E._Krauss), the herald of the call to action that gave rise to the field of language documentation.

US) and, thus, before informed consent was required for human participation in social science and humanities research. Another portion of archived language documentation includes linguistic and cultural materials for which cultural protocols exists (e.g., certain Native American stories are meant to be told only during the winter; ritual initiation is required to hear and/or sing certain shamanic songs). Yet another portion includes materials that were collected using colonial and extractive practices and archived without the involvement or knowledge of the Origin Community. In many cases, those bodies of archived materials overlap. For this reason, language archives have graded access to the holdings (described in Johnson 2004), which in most cases prevent anonymous archives users (i.e., anyone on the internet who finds their way to the archive's website) from gaining immediate access to the files.

During the ten years between the publication of Johnson 2004 and Woodbury 2014, a period of time during which there was (relatively speaking) a great deal of funding available for language documentation projects, the string attached to the funding was that the resulting research "data" must be archived. Since the language archives had ways to restrict access to the files (graded access, described in Johnson 2004), many researchers complied with the letter of the law—they archived the language data—but not with the spirit of the law—they restricted all or most of the data in their collections for a variety of reasons. Some of those reasons where quite legitimate and involved sensitivity of the data and/or cultural protocols to access them, but other reasons were quite self-serving, including unwillingness to share the data for fear of being scooped and lack of time needed to anonymize files where necessary, check files for sensitive content, carefully arrange the files and compile the detailed metadata required to give each file the context necessary for future reuse. It was simply easier to restrict access to the data than it was to properly curate them. At the same time, computational researchers and practitioners were making rapid advancements in *natural language processing* (NLP), and the technology that worked for English and other major world languages (e.g., screen readers, automated transcription, OCR, and the like) still did not work at all for under-resourced minority languages. To make NLP work for any give language, researchers need access to a large amount of data in that language; they turned to the language archives to get that data, but they found that much of the data was restricted or difficult to find, access, or download. The Open Data Movement arrived at the digital language archives and found that it could not access the data.

For the last decade, researchers from disciplines outside language documentation have called for more open access to the data in language archives. The final paper in this collection, ***Public access to research data in language documentation: Challenges and possible strategies*** (2019) by Mandana Seyfeddinipur et al., is a response to this call. This paper is the result of a workshop[3] on "Open Access and Open Data of Endangered Languages Collections" during which the participants sought to find a compromise between the legitimate need, on the one hand, to make data open both for funding compliance and in the interest of furthering scientific research and, on the other hand, the equally legitimate need to restrict data from general access,

[3] The workshop was held at the University of Cologne, Germany, on October 10–12, 2016. I was a participant in this workshop, and, as such, I am a co-author on this article.

sharing and reuse according to the cultural practices, needs, and wishes of the Origin Communities. Though the workshop took place in 2016, the paper was published three years later in 2019, shortly before the worldwide COVID-19 pandemic began. We have yet to see what impact it will have on keeping data in language archives as open as possible, and as closed as necessary.

I hope that you will learn something from reading or skimming the articles in this collection. If nothing else, perhaps they will provide a way to pass the time as you wait for the pandemic to end and the world to return to something similar to what it was before.

Susan Smythe Kung
July 19, 2020
Sheltering at home in Austin, Texas, USA, during the COVID-19 pandemic

References in this Introduction

Bird, Steven and Gary Simons. 2003. Seven dimensions of portability for language documentation & description. *Language* 79(3): 557-582.

Henke, Ryan and Andrea L. Berez-Kroeker. 2016. A brief history of archiving in language documentation with an annotated bibliography. *Language Documentation & Conservation* 10: 411-457.

Himmelmann, Nikolaus. 1998. Documentary and descriptive linguistics. *Linguistics* 36(1): 161-196.

Johnson, Heidi. 2004. Language documentation and archiving, or how to build a better corpus. In Peter K. Austin (ed.), *Language documentation and description, vol. 2*, pp. 140-153. London: SOAS.

Krauss, Michael E. 1992. The world's languages in crisis. *Language* 68(1): 4-10.

Seyfeddinipur, Mandana, Felix Ameka, Lissant Bolton, Jonathan Blumtritt, Brian Carpenter, Hilaria Cruz, Sebastian Drude, Patience L. Epps, Vera Ferreira, Ana Vilacy Galucio, Brigit Hellwig, Oliver Hinte, Gary Holton, Dagmar Jung, Irmgarda Kasinskaite Buddeberg, Manfred Krifka, Susan Kung, Miyuki Monroig, Ayu'nwi Ngwabe Neba, Sabastian Nordhoff, Brigitte Pakendorf, Kilu von Prince, Felix Rau, Keren Rice, Michael Riessler, Vera Szoelloesi Brenig, Nick Thieberger, Paul Trilsbeek, Hein van der Voort, Tony Woodbury. 2019. Public access to research data in language documentation: Challenges and possible strategies. *Language Documentation & Conservation* 13: 545-563.

Woodbury, Anthony C. 2003. Defining documentary linguistics. In P.K. Austin (ed.), *Language Documentation and Description, vol. 1,* pp. 85-101. London: SOAS.

Woodbury, Anthony C. 2014. Archives and audiences: Toward making endangered language documentations people can read, use, understand, and admire. In David Nathan & Peter K. Austin (eds) *Language Documentation and Description, vol 12: Special Issue on Language Documentation and Archiving,* pp. 19-36. London: SOAS.

LANGUAGE DOCUMENTATION & CONSERVATION

Vol. 10 (2016), pp. 411–457
http://nflrc.hawaii.edu/ldc
http://hdl.handle.net/10125/24714
Revised Version Received: 19 April 2016

Series: Emergent Use and Conceptualization of Language Archives
Michael Alvarez Shepard, Gary Holton & Ryan Henke (eds.)

A Brief History of Archiving in Language Documentation, with an Annotated Bibliography

Ryan E. Henke
University of Hawai'i at Mānoa

Andrea L. Berez-Kroeker
University of Hawai'i at Mānoa

We survey the history of practices, theories, and trends in archiving for the purposes of language documentation and endangered language conservation. We identify four major periods in the history of such archiving. First, a period from before the time of Boas and Sapir until the early 1990s, in which analog materials were collected and deposited into physical repositories that were not easily accessible to many researchers or speaker communities. A second period began in the 1990s, when increased attention to language endangerment and the development of modern documentary linguistics engendered a renewed and redefined focus on archiving and an embrace of digital technology. A third period took shape in the early twenty-first century, where technological advancements and efforts to develop standards of practice met with important critiques. Finally, in the current period, conversations have arisen toward participatory models for archiving, which break traditional boundaries to expand the audiences and uses for archives while involving speaker communities directly in the archival process. Following the article, we provide an annotated bibliography of 85 publications from the literature surrounding archiving in documentary linguistics. This bibliography contains cornerstone contributions to theory and practice, and it also includes pieces that embody conversations representative of particular historical periods.

1. Introduction It is difficult to imagine a contemporary practice of language documentation that does not consider among its top priorities the digital preservation of endangered language materials. Nearly all handbooks on documentation contain chapters on it; conferences hold panels on it; funding agencies provide money for it; and even this special issue evinces the central role of archiving in endangered language work. In fact, archiving language data now stands as a regular and normal part of the field linguistics workflow (e.g., Thieberger & Berez 2011).

This state of affairs has not always been the norm. Moreover, the idea of archiving as an ongoing process instead of something to be done at the end of one's career is a relatively new development. This paper is a historical exploration of the chain of

E-ISSN 1934-5275

events that have led us to this state, beginning in the late eighteenth century and continuing through to the present day.

Traditionally, archived resources consisted of physical objects (e.g., books, tools, photographs, artwork, and clay tablets), and because of the value of such objects, archives restricted access to them to varying degrees (Austin 2011, Nordhoff & Hammarström 2014, Trilsbeek & Wittenburg 2006). Typical homes for archived materials have long included museums, libraries, universities, and, of course, dedicated archival institutions (Linn 2014). In terms of access, this traditional model of archiving has entailed a 'one-way' street: Depositors put material into archives managed by archivists, and only people with the requisite permission and ability can find and access archived resources (Nathan 2014). In a nutshell, this was more or less the model for archiving from the beginning of modern linguistic work.

In order to provide a foundation for assessing how conceptualizations of archiving have changed dramatically, especially over the last twenty-five years, it is helpful to define what we mean by *endangered language archive*. We take *archive* to mean "a trusted repository created and maintained by an institution with a demonstrated commitment to permanence and the long-term preservation of archived resources" (Johnson 2004:143). Furthermore, this history is concerned primarily with archives designed to preserve materials related to small, endangered, and/or Indigenous languages.

We have identified four major periods in the development of endangered language archiving, each of which is discussed in the sections below:

- An early period, lasting from before the time of Boas and Sapir until the early 1990s, in which analog materials—everything from paper documents and wax cylinders to magnetic audio tapes—were collected and deposited by researchers into physical repositories that were not easily accessible to other researchers or speaker communities (§2);
- A second period, beginning in the 1990s, in which increased attention to language endangerment and language documentation brought about a redefined focus on the preservation of languages and language data (§3);
- A third period, starting in the early twenty-first century, in which technological advancements, concerted efforts to develop standards of practice, and large-scale financial support of language documentation projects made archiving a core component of the documentation workflow (§4);
- The current period, in which conversations have arisen toward expanding audiences for archives and breaking traditional boundaries between depositors, users, and archivists. (§5).

In §6, we present some critical review of the current state of archiving. This includes assessing how archiving has actually permeated the workflow of documentary linguists as well as how our field acknowledges and rewards scholarly and professional contributions in archiving.

2. Early linguistic archiving: Late 19th century–1991 For Americanist pioneers like Franz Boas and Edward Sapir in the late nineteenth and early twentieth centuries, archiving was an essential component of the work to document Indigenous languages (Johnson 2004). Documentation during this period consisted mostly of textual materials such as fieldnotes, translations, elicitation data, lexical compilations, and grammatical descriptions (Golla 2005, Johnson 2004). Throughout this period, linguists deposited their records in archives, universities, and museums; even monographs from such institutions as well as publications like the *International Journal of American Linguistics* served as "archiving mechanisms" for texts, grammars, and dictionaries from Indigenous languages, inasmuch as they became part of the published record (Woodbury 2011:163). However, with the exception of publications, such collections were available only to researchers with the inclination and capabilities to travel to archives and access the materials (Johnson 2004).

This conceptualization of archiving as the protection of physical items behind a brick-and-mortar wall remained relatively stable for many decades, and several notable archival institutions arose during this period. Among the most significant are the following:

1. Since its founding in 1743, the American Philosophical Society (APS)[1] collected Native American manuscripts, including a famous and extensive collection from Thomas Jefferson (Golla 1995). With its 1945 acquisition of the Franz Boas Collection of American Indian Linguistics from the American Council of Learned Societies, the APS became the "primary repository for the records of twentieth-century American Indian linguistics" (1995:148).

2. The University of California, Berkeley has been involved with archiving linguistic data since the early twentieth century,[2] beginning with the work of A. L. Kroeber, Pliny Earle Goddard, T. T. Waterman, Edward Sapir, and E. W. Gifford (Golla 1995). The Survey of California Indian Languages was officially founded at Berkeley in 1953 and renamed The Survey of California and Other Indian Languages in 1965. The leadership of Murray Emeneau and Mary Haas yielded a particularly important period: Under their direction, Berkeley housed "a veritable factory of graduate students who produced Boasian grammar-dictionary-text trilogies published by the *University of California Publications in Linguistics*. These texts were linked to audio-recordings which, along with field notes and slip-files, were archived with the Survey of California Indian Languages" (Woodbury 2011:166).

3. The National Anthropological Archives (NAA)[3] was created in 1965 from a merger between the Department of Anthropology at the Museum of Natural History of the Smithsonian Institution and the Bureau of American Ethnology (BAE). The latter was the "most active sponsor of linguistic research on American Indian languages" during the late nineteenth and early twentieth centuries

[1]https://www.amphilsoc.org/
[2]http://linguistics.berkeley.edu/~survey/about-us/history.php
[3]http://anthropology.si.edu/naa/index.htm

(Golla 1995:148). Along with many other linguists, the BAE employed John Peabody "J. P." Harrington from 1915 until 1954, and he produced a massive amount of documentary linguistic work (Golla 1995, Macri & Sarmento 2010).

4. In 1972, Michael Krauss founded the Alaska Native Language Center (ANLC), later renamed the Alaska Native Language Archive (ANLA), at the University of Alaska, Fairbanks.[4] ANLA's archival library contains an unparalleled collection of print and audio materials from and about Alaska's 20 Indigenous languages (Krauss 1974, Woodbury 2010, Holton 2012, Holton 2014).

From Boas' time onward, technological developments changed linguistic fieldwork as well as the types of materials stored in archives. As noted, text (whether handwritten or created via typewriter) had always served as a cornerstone of linguistic archives, but the beginning of the twentieth century also brought about the capacity to archive audio materials. Linguists captured and archived sound data using a progression of technology, employing wax cylinders (used to collect, for example, recordings of Native American music and language for the BAE) until the arrival of the phonograph in the 1930s (used by linguists like Melville Jacobs and J. P. Harrington), which was then replaced by tape recording technology in the 1950s before video recording technology became widely available in the 1980s (Golla 1995, Johnson 2004, Thieberger & Musgrave 2007). Of course, these analog methods gave way to the rise of digital technology in the latter half of the twentieth century: The digital archiving of language materials finds its origins in the use of computers for social science research in the early 1960s (Austin 2011, Doorn & Tjalsma 2007). The Oxford Text Archive,[5] founded in 1976 by Lou Burnard, represents one of the earliest text archives in use by linguistic communities (Doorn & Tjalsma 2007), and the Linguistic Data Consortium was formed at the University of Pennsylvania in 1992 to address data shortages by serving as a repository and distributor for language resources.[6]

This progression to digital technology brought increasing efficiency and ease for data collection, but not enough attention went toward devising bigger and better ways to archive linguistic material systematically and sustainably. For example, Indiana University began the Archives of the Languages of the World in the mid-1950s to store vast volumes of tape records, but a lack of technical support forced the abandonment of the project (Golla 1995).[7] At least part of the problem stemmed from the fact that traditional archives were not equipped to handle the massive amounts of data being produced, whether in terms of providing long-term storage or managing access by researchers or communities (Johnson 2004). Untold masses of text materials and thousands of hours of recordings, which had been accumulating for decades in the possession of linguists and anthropologists around the world, sat idle—only a fraction of linguistic data managed to make it into dedicated archives (Johnson 2004, Trilsbeek & Wittenburg 2006). This state of affairs did not change much until the

[4]https://www.uaf.edu/anla/about/
[5]http://ota.ox.ac.uk/
[6]https://www.ldc.upenn.edu/about
[7]This collection has been subsumed into the Indiana University Archives of Traditional Music: http://www.indiana.edu/ libarchm/index.php/atm-collections.html.

1990s, which saw the rise of documentary linguistics and a renewed and redefined focus on archiving.

3. Documentary linguistics and a new approach to archiving: 1991–2006 In the early 1990s, a growing number of linguists turned their attention to the problem of mass language endangerment and death (e.g., Hale et al. 1992). These scholars perceived an unprecedented crisis in the field, and the conversation began toward finding solutions: "Obviously we must do some serious rethinking of our priorities, lest linguistics go down in history as the only science that presided obliviously over the disappearance of 90% of the very field to which it is dedicated" (Krauss 1992:10). Soon after, this concern helped fuel Himmelmann's (1998) refinement of *documentary linguistics* (or *language documentation*) as a distinct subfield of linguistics, although some say this was simply a homecoming back to the discipline's roots as a fieldwork-based research enterprise, as mainstream linguistics had become increasingly more theoretical since the generative revolution of the 1950s and 1960s (Conathan 2011, Himmelmann 2006, Thieberger & Musgrave 2007, Woodbury 2003).

But what makes documentary linguistics different from descriptive linguistics? Traditionally, *descriptive* linguistics revolves around the Boasian trilogy of texts, dictionaries, and grammars based on in-depth analyses of primary data from a given language (Himmelmann 1998, Himmelmann 2006, Woodbury 2003, Woodbury 2011). *Documentary* linguistics is much broader and more ambitious in scope. As Himmelmann himself defined it, a language documentation is a "record of the linguistic practices and traditions of a speech community" (1998:166). Woodbury (2003:46–48) usefully elaborated upon this definition by proposing some widely agreed-upon values for proper documentation: A good documentation is diverse, large, ongoing, distributed, and opportunistic with material that is transparent, preservable, ethically created, and portable. Broadly speaking, a documentation provides a sizeable record of a language in use across a range of discourse, furnishing a copious amount of transcribed and annotated audio/video materials accompanied by contextual metadata (Austin 2013, Austin & Grenoble 2007, Johnson 2004). This creates "a lasting, multipurpose record of a language" (Himmelmann 2006:1), which can be employed not only to address language endangerment but also to provide data for linguistics and other disciplines, improve scientific accountability, and maximize the economy of research resources. As such, another element distinguishing modern documentation efforts from those of the past is "concern for long-term storage and preservation of primary data" (Himmelmann 2006:15). We return to this point later.

A handful of major factors enabled the rise of documentary linguistics during this time period (Austin 2012, Austin 2014, Austin & Grenoble 2007, Woodbury 2003). First, of course, was the increased attention to language endangerment. A second factor was the increase in funds for documentary projects, primarily from three major sources: Germany's Volkswagen Foundation, which began the Dokumentation bedrohter Sprachen[8] (DOBES) program in 2000; the Arcadia Trust[9] in

[8]http://dobes.mpi.nl/dobesprogramme/

[9]http://www.arcadiafund.org.uk/about-arcadia/about-arcadia.aspx

the United Kingdom, which started the Endangered Languages Documentation Programme[10] (ELDP) in 2003; and the National Science Foundation and the National Endowment of the Humanities, which together initiated the Documenting Endangered Languages[11,12] (DEL) program in 2005 (Austin 2012, Austin 2014, Woodbury 2003). Other notable funders emerged in this period as well, such as the Community-University Research Alliance[13] and the Aboriginal Research Programme[14] of the Social Sciences and Humanities Research Council of Canada (SSHRC), the Foundation for Endangered Languages[15] (FEL) in the United Kingdom, and the Endangered Language Fund[16] (ELF) in the United States (Woodbury 2011). Finally, modern documentary linguistics was able to emerge due to monumental developments in digital information technology, which enabled more efficient and higher quality recording of audio and video; processing, analysis, and storage of such materials; and the widespread distribution of such information through the internet—all to extents that were previously impossible (Austin 2013, Austin & Grenoble 2007, Bird & Simons 2003, Evans & Dench 2006, Johnson 2004, Woodbury 2003). In 1991 the Australian Institute of Aboriginal and Torres Strait Islander Studies[17] (AIATSIS) created what might be the first digital archive dealing with endangered languages, the Aboriginal Studies Electronic Data Archive[18] (Thieberger 1994).

Along with this new conceptualization of documentary linguistics came a renewed and redefined focus on archiving. From the beginning, archiving occupied one of the four steps laid out in Himmelmann's model of documentation: "presentation for public consumption/publicly accessible storage (archiving)" (1998:171). A host of scholars agreed that archiving is a cornerstone of documentation, (e.g., Austin & Grenoble 2007, Johnson 2004, Rehg 2007, and Woodbury 2003). The reason for this is simple: If we are going to dedicate immense amounts of time, money, and energy to preserve endangered languages, then all of our efforts would be futile without a plan for that information to be put to use safely and sustainably by future generations for a variety of purposes—including facilitating studies in a range of scientific disciplines, enabling verification of data analyses, and producing language teaching materials (Austin 2014, Evans & Dench 2006, Himmelmann 2006, Nathan 2014, Thieberger & Musgrave 2007). This perspective on archiving is considered by some to be another factor distinguishing documentation from description (Himmelmann 2006, Nathan & Austin 2014). With this new outlook on archiving, it was not long before many came to see an inseparable relationship between language documentation and the archive: "All documentation projects should be conceived with an eye toward the ultimate deposit of the recorded data and analysis in an archive" (Austin & Grenoble

[10]http://www.eldp.net/
[11]https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=12816
[12]http://www.neh.gov/grants/preservation/documenting-endangered-languages
[13]http://www.sshrc-crsh.gc.ca/funding-financement/programs-programmes/cura-aruc-eng.aspx
[14]http://www.sshrc-crsh.gc.ca/funding-financement/programs-programmes/priority_areas-domaines_prioritaires/aboriginal_research-recherche_autochtone-eng.aspx
[15]http://www.ogmios.org/index.php
[16]http://www.endangeredlanguagefund.org/
[17]http://aiatsis.gov.au/
[18]http://aseda.aiatsis.gov.au/asedaDisclaimer.php

2007:19). Importantly, it was not just linguists who came to regard archiving as an integral part of language documentation—so did a lot of the people with the money: Organizations like DOBES, EDLP, DEL, and the ELF have come to mandate archiving as part of their documentation project requirements (Austin 2014).

Finally, along with this new view of language documentation, linguists increasingly acknowledged the importance of archiving to Indigenous language revitalization efforts (e.g., Gerdts 2010 and Johnson 2004), which had been gaining steam particularly in the United States since the late 1960s (Gehr 2013). As Hinton (2001) explained, revitalization efforts often begin with a search for existing documentation, which may be housed in large national archives like the Smithsonian or in small, local archives. Moreover, when a strong reliance on native speakers is not possible, the development of pedagogical materials for revitalization efforts, such as dictionaries or language lessons, is often based on archived linguistic documents (2001). The oft-cited case of the Mutsun language represents a famous case for the value of archiving: Records from the nineteenth and early twentieth centuries enabled the production of a grammar in 1977—more than 40 years after the death of the last speaker—as well as subsequent revitalization endeavors (Conathan 2011, Macri & Sarmento 2010). In 1996, one of the most significant American revitalization efforts began when the Advocates for Indigenous California Language Survival[19] held its first Breath of Life Workshop,[20] bringing Indigenous community members to the Berkeley archives to teach them linguistic fundamentals and show them how to use archived materials to facilitate language restoration. Another example of a revitalization program began around 2000 in Canada, when Peter Brand and SENĆOŦEN speaker and teacher John Elliott, Sr. began using the internet to "support Aboriginal people engaged in language archiving, language teaching, and culture revitalization" through the FirstVoices[21] project (Czaykowska-Higgins 2009:31).

By the early 2000s, documentary linguistics had arrived, and it brought a new conceptualization of the power and necessity of archiving. Now linguists faced the question: *How* should we archive?

4. *How* should we archive?: 2000–2010 Documentary linguists recognized the benefits conferred by *digital* archives. For one, digital information is not susceptible to the same problems of physical deterioration that plague wax cylinders, vinyl records, paper documents, magnetic tapes, and other analog materials—whether housed in traditional archives or sitting idle on researchers' shelves (Bird & Simons 2003, Chang 2010, Johnson 2004, Nathan 2011). Some noticed this particular advantage early on, drawing attention to the need to digitally curate such legacy materials: "One of the major tasks of linguistic anthropology in the decades ahead will be to exercise appropriate stewardship over the archival record of American Indian languages" (Golla 1995:152). Other advantages of digital archives include providing much greater capacity for long-term preservation and storage of multimedia data

[19]http://www.aicls.org/
[20]http://www.aicls.org/breath-of-life
[21]http://www.firstvoices.com/

(Nathan 2011), and enabling easier access to and retrieval of information (Trilsbeek & Wittenburg 2006). These capacities shattered conceptions of limitations on both the scope of a given documentary corpus as well as the ability of researchers to fact-check claims directly by going to the data. The following passage embodies this sentiment:

> Digital audio and video recording, portable storage, and the development of software enabling the tagging, management and analysis of collected data raises the stakes for corpus collections. Our traditional published text collection consisted of a few hundred pages of narrative text with interlinear glosses, free translation and explanatory notes, but the modern published corpus may potentially consist of digital audio recordings of data collection sessions, some with accompanying video, and linked to a range of transcriptions representing different kinds and levels of analysis. Where the published text collection once served as the grounding evidence for a linguistic analysis, the digital archive will come increasingly to fill that role. (Evans & Dench 2006:24)

With this recognition of the possibilities granted by digital archives, many documentary linguists seemed mostly unaware that archivists outside of linguistics had already been working for a while to figure out *best practices*[22] for digital archiving (Woodbury 2011). For instance, the Task Force on Archiving of Digital Information was created in 1994 by the Commission on Preservation and Access and the Research Libraries Group, and the task force reported in 1996 the need for trustworthy digital archiving organizations (Chang 2010). Moreover, between 1995 and 2002, the NASA Consultative Committee for Space Data Systems[23] developed the Reference Model for an Open Archival Information System (OAIS), which aimed at requirements for long-term preservation of digital information, including navigating issues with changing user communities and technologies (2010). Years later, "the OAIS Reference Model continues to have wide acceptance in the digital library community, and has become the authoritative model for best practices in digital archiving" (2010:61). Despite an ostensible lack of interdisciplinary communication in this regard, documentary linguists in the early and mid-2000s (e.g., Bird & Simons 2003, Evans & Sasse 2004, and Himmelmann 2006) were becoming increasingly interested in figuring out the best ways to carry out digital archiving of language documentation.

Bird and Simons (2003) took one of the earliest and most important steps toward best digital archiving practices. They called attention to some of the biggest issues facing documentary linguists looking to make data as long-lasting and usable as possible. For example, Bird and Simons noted that "a substantial fraction of the resources being created can only be reused on the same software/hardware platform, within the same scholarly community, for the same purpose, and then only for a period of

[22]According to E-MELD (Electronic Metastructure for Endangered Language Data), best practices for digital archiving of linguistic work are "practices which are intended to make digital language documentation optimally longlasting, accessible, and re-usable by other linguists and speakers" (http://emeld.org/school/what.html).

[23]http://public.ccsds.org/default.aspx

a few years." (2003:579). To fix this problem, they called for a sea change in both technologies and attitudes. In their words: "We need nothing short of an open source revolution, leading to new open source tools based on agreed data models for all of the basic linguistic types, connected to portable data formats, with all data housed in a network of interoperating digital archives" (2003:579).

Another topic in best-practices conversation focused on approaches toward metadata. Metadata, often described as data *about* data, accompanies primary data to provide valuable context and meaning (e.g., speaker identification, date of recording, and genre of text), and is especially useful in determining how data can be located in an archive and how it can and should be used (Austin 2013, Innes 2010, Thieberger & Berez 2011). An important metadata development came in December 2000 with an NSF-funded workshop, Web-Based Language Documentation and Description, held in Philadelphia (Bird & Simons 2003). This workshop gave rise to the founding of the the Open Language Archives Community[24] (OLAC), which is devoted to "(i) developing consensus on best current practice for the digital archiving of language resources, and (ii) developing a network of interoperating repositories and services for housing and accessing such resource." (Bird & Simons 2003:572-573). Among OLAC's contributions are the OLAC Metadata standard[25] and the OLAC Repositories standard,[26] a protocol for harvesting metadata (2003). Another metadata standard arose during this time, too:

The International Standards for Language Engineering Metadata Initiative[27] (IMDI), developed by DOBES, which "is a more comprehensive metadata system that can be used to manage several archival functions, including not only description but also preservation and access" (Conathan 2011:246). Both the OLAC and IMDI schemas have come to be endorsed and adopted by many documentary linguists (Johnson 2004, Himmelmann 2006, Thieberger & Berez 2011).

Other best-practice discussions centered on the collection and management of primary data. For example, Austin (2006) covered ways to manage various forms of data involved in a language documentation, including how to select and use recording equipment, choose data formats (e.g., XML, WAV, or MPEG2), transfer analogue materials to digital form, and process data with software tools like Shoebox. Gippert (2006) discussed the history of and best practices for digitally encoding text (e.g., problems with ASCII and the power of Unicode), including managing structural elements like phrases and clauses. Robinson (2006) talked about the importance of archiving directly from the field to enhance the safety of collected data in conditions that are often inhospitable to electronics. Schroeter and Thieberger (2006) explored the need to have standard data structures, provided to linguists through templates and workflow directives, that can apply across various tools for transcribing and annotating linguistic data. Thieberger (2010) further dealt with data management, location and citation, formation, storage, reuse, and interoperability—while stress-

[24]http://www.language-archives.org/
[25]http://www.language-archives.org/OLAC/metadata.html
[26]http://www.language-archives.org/OLAC/repositories.html
[27]http://www.mpi.nl/imdi/

ing the need for training other linguists in best practices like employing consistent file naming, using OLAC metadata standards, and making data searchable by others. Following Bird & Simons (2003), one of the most important best practices to emerge during this period was the insistence on the use of open-source and uncompressed data formats for collecting and structuring linguistic data (e.g., Good 2011 and Thieberger 2010), which together help stave off obsolescence and make information as rich, long-lasting, and accessible as possible . By 2010, the discussion about *how* to archive even resulted in at least one MA thesis providing a checklist intended to help language documenters choose the proper archive for their deposits (Chang 2010).

Concomitant with these discussions in the literature came the development of organizations and initiatives devoted to implementing and disseminating best practices for archiving language documentation. Established in 2001 after a one-year pilot project, the DOBES program at the Max Planck Institute in the Netherlands mandated that its funded projects adopt "specifications for archival formats, recommendations about recording and analysis formats, and the development of new software tools to assist with audio and video annotation (such as ELAN), and the creation and management of metadata (various IMDI tools)" (Austin 2014:61).[28] From 2001 to 2006, the National Science Foundation funded the Electronic Metastructure for Endangered Language Data[29] (E-MELD) project, which aimed at creating consensus and sharing information on best practices in documentation, including data markup, labels for interlinear glossing, and metadata creation (Austin 2014, Boynton et al. 2006). E-MELD has particular importance because it represented the first time linguists came together to create a significant set of digital standards for documentation. As part of the task of creating stronger networks within the archiving community, the Digital Endangered Languages and Musics Archives Network[30] (DELAMAN) came about in 2003 as an international umbrella body dedicated to creating stronger networks within the archiving community. The push for best practices even resulted in a newsletter that ran from 2004 to 2007, the *Language Archives Newsletter*,[31] which was specifically devoted to issues in archiving (Woodbury 2010).

Furthermore, established archival projects were increasingly going digital (Trilsbeek & König 2014), and new archives emerged with a focus on digital formats and best practices. The ANLC became a founding member of OLAC in 2000, creating an electronic catalog database as well as a digital archive for the Dena'ina Qenaga language (Holton 2014, Holton et al. 2006). The Archive of the Indigenous Languages of Latin America[32] was founded in 2000 at the University of Texas at Austin. Three years later, linguists and musicologists established the Pacific and Regional Archive for Digital Sources in Endangered Cultures[33] (PARADISEC) to digitize and curate field recordings compiled since the 1960s by Australian researchers (Thieberger &

[28]http://tla.mpi.nl/tools/tla-tools/elan/
[29]http://emeld.org/
[30]http://www.delaman.org/
[31]http://www.mpi.nl/LAN/
[32]http://www.ailla.utexas.org/site/welcome.html
[33]http://paradisec.org.au/

Barwick 2012, Thieberger 2013, Thieberger et al. 2015a). After a year of development, the Hans Rausing Endangered Language Project at the School of Oriental and African Studies opened the Endangered Language Archive[34] (ELAR) in 2005 (Nathan 2010, 2014). Modeled on PARADISEC, yet another digital archive opened in 2008: *Kaipuleohone*, the University of Hawai'i Digital Language Archive,[35] which aims to make extant research more discoverable and to preserve language documentation materials (Albarillo & Thieberger 2009, Berez 2013, Berez 2015, Rehg 2007).

From around 2000 to 2010, it appears that documentary linguists had largely succeeded in establishing a general set of (or at least a very rich dialogue around) best technological practices along with initiatives and organizations for digitally archiving language documentation data. However, throughout this period we also see a recognition of various limitations and problems associated with digital archiving. This includes challenges to the idea that a single, comprehensive set of 'best practices' makes sense, given the the wide spectrum of language documentation situations. This critical response has also been observed and discussed at length by Austin (2014:62–65). Austin (2013:4) summarized the situation well:

> Some researchers have emphasised standardization of data/metadata and analysis and "best practices" (e.g., E-MELD, OLAC) while others have argued for a diversity of approaches which recognize the unique and particular social, cultural and linguistic contexts within which individual languages are used.

For example, Bowden & Hajek (2006) pointed out that seemingly 'best' practices are not always relevant or possible to carry out, given varying circumstances in the field: Perhaps there is no electricity; team members may be spread out over wide distances, which inhibits workflow; or local community members may be completely unfamiliar with digital technology. In the face of challenges such as diverging goals and cumbersome workflows, Berez & Holton (2006) noted the difficulties of getting speaker communities—and even other linguists—on board to adopt best practices for long-term data preservation.

On the other hand, arguments also critiqued the limited vision of existing best-practice concepts. For instance, Johnson (2004) and Nathan & Austin (2004) called for richer contextual information to be added to metadata, claiming that existing metadata standards and archival protocols do not go far enough in adding value to data. Nathan (2009) cited the need for an 'epistemology' for audio recording in language documentation, one that goes beyond existing discussions limited to formats and resolution to deal with recording spatial and configuration information as well as controlling signal and noise. Still others pointed out that archived materials will have uses beyond the original purposes for which they were collected and archived: "It is imperative for linguists to understand both the possibilities and the limitations of current archival practices so they can prepare for and advocate for the best possible management of the records they create, and of legacy archival collections" (Conathan

[34]http://elar.soas.ac.uk/
[35]http://kaipuleohone.org

2011:236). Ironically, plenty of time, attention, and resources had been spent developing and promoting best practices regarding documentary linguistic data, but linguists still had not conceived a system to test the effectiveness and longevity of language archives themselves (Chang 2010).

Perhaps the biggest reaction to best-practices conversations has concerned variegated issues of ethics and access (e.g., Dwyer 2006, Green et al. 2011, and Innes & Debenport 2010). Although the digital nature of archives can allow for easier, increased access of archived materials, this is not always a simple matter. For instance, O'Meara and Good (2010) raised issues relating to defining a 'community'; establishing rights to access archived material retroactively; establishing rights and access to "orphan" works that do not have an identifiable copyright holder; and assessing and dealing with sensitivities related to the content of archived materials. Garrett and Conathan (2009) described problems resulting from failures of planning by linguists and archives, which are compounded when parties—whether linguists, speakers, or heritage communities—seek restrictions to access for materials. Although Garrett and Conathan suggested having a consistent, comprehensive, and clear strategy for archiving and developing access restrictions in consultation with heritage communities, this cannot solve every problem. We see this, for example, with informed consent (e.g., Thieberger & Musgrave 2007). Given the fact that linguists and speakers cannot exhaustively anticipate future technological developments and new uses for language documentation data, Thieberger and Musgrave wondered "how the data collector can fully inform the speakers about the nature of the activities to be undertaken" (2007:31). Other ethical dilemmas involve increased public access to sensitive materials, where community members may regard archived data (e.g., narratives, songs, and stories) as sacred, embarrassing, or even dangerous to others (Innes 2010,[36] Macri & Sarmento 2010, Thieberger & Barwick 2012). In such cases, linguists may have an ethical responsibility of "providing as rich a system of ethnographic information as possible," such as ideological statements and behavioral descriptions, in order to ameliorate future problems with the reinstatement or reproduction of archived texts and discourse (Innes 2010:202). Finally, ethical concerns arise from from the fact that "the rules of intellectual property, although set by international standards, often conflict with customs of traditional indigenous groups" (Macri & Sarmento 2010:195).

This has certainly not been an exhaustive account of all the reactions to the "best practices" conversation during this period. However, they do illustrate the broader progression of history: By around 2010, documentary linguists had developed a healthy discourse around both 1) establishing sustainable digital archives that last a long time, permit access to various parties, and provide utility to scientists and speech communities; and 2) grappling with the problems and limitations of trying to squeeze a one-size-fits-all archival approach upon the varied, idiosyncratic contexts of the field. Archiving in language documentation had come a long way in a very short

[36]In the case of her Mvskoke language work, Innes simply chose to stop working with some sensitive materials: "Here, I find that I cannot continue to work on these narratives as this causes my consultants real difficulty and concern" (2010:202).

amount of time, and new discussions soon began around further reconceptualizing the model and role of the archive.

5. Redefining archiving through participatory models: 2010–present Throughout the transition from traditional analog repositories to the power and potential of digital archives, we see the persistence of a "one-way" model of archiving: "providers lodge their materials with the archive and users can (if permissions allow) find and access them" (Nathan 2014:193). This models entails limits on the interaction between depositors and users and between users and archived material (Trilsbeek & Wittenburg 2006). Throughout, the archivist is at the center of the archiving process. In the last few years, however, this situation has changed dramatically with the development of *participatory* archiving models in linguistics. Specifically, one definition of a *participatory archive* is "an organization, site or collection in which people other than the archives professionals contribute knowledge or resources resulting in increased understanding about archival materials, usually in an online environment" (Theimer 2011). The rise of such a model in linguistics seems to have been enabled by four primary factors:

1. The development of community-oriented models of linguistic research
2. The increasing empowerment of Indigenous communities in stewarding their own languages
3. The integration of social media models in archiving
4. The development of participatory models in the archival sciences

5.1 Community-oriented research By late in the first decade of the twentieth century, documentary linguists were increasingly turning to models of research that relied upon collaboration with language communities (e.g., Cameron et al.'s 1992 "empowering" model). Of particular significance is the Community-Based Language Research (CBLR) model outlined by Czaykowska-Higgins in 2009 (author's emphasis):

> Research that is on a language, and that is conducted ***for, with***, and ***by*** the language-speaking community within which the research takes place and which it affects. This kind of research involves a collaborative relationship, a partnership, between researchers and (members of) the community within which the research takes place (24).

The CBLR represents a departure from the traditional model of research in linguistics. For more than a century, research has mostly been carried out by linguists for an audience of linguists, regarding speakers and speaker communities primarily as sources of data—no matter how ethically conscious such engagements might actually be (2009). Although Czaykowska-Higgins was not the first linguist to advocate and practice a collaborative approach to research (e.g., Cameron et al. 1992, Dwyer

2006, and Yamada 2007), she was one of the first to put forth a clear, systematic model for others to follow.

Around this time, there seems to be a shift in the language documentation literature, a stronger acknowledgement of the value of collaborating with communities in linguistic enterprises and producing research that serves the interests of both linguists and speakers (e.g., Good 2011, Dorbin & Holton 2013). This move reconceptualizes the longstanding research paradigm by moving from treating communities as objects of study to "actively including them in the process of documenting their language" (Wilbur 2014:68).

5.2 Empowerment of Indigenous communities Another factor facilitating participatory developments in linguistic archiving has been the fact that Indigenous communities over the last several decades have taken increasing levels of agency and ownership in stewarding their languages through documentation and revitalization (Hinton 2001, Macri & Sarmento 2010). Native communities in the United States, for example, have been stepping up in language scholarship as well as producing materials like phrasebooks, dictionaries, and curricula for revitalization (Hinton 2005). As Indigenous archive activist Allison Boucher Krebs put it (2012:182):

> Whereas historically the flow of information about Indian Country has been away from Indian Country and once outside, about Indian Country by scholars, researchers, and non-Indigenous professionals, today information is flowing back to communities and within communities. The scholars, researchers, and professionals are increasingly likely to be Indigenous.

Of course, this also means that Indigenous communities in the United States, Canada, and Australia have been taking much more active roles in archiving their cultural heritage. In 2005, Hinton noted that the archives at Berkeley were "being used far more by Native Americans than by social scientists for purposes of language and cultural maintenance and revitalization" (24–25), and Holton (2014) observed that ANLA has become an increasingly important resource for revitalization activities in Alaska since the late 1990s.

Indigenous communities have also been taking the reins by creating their own archival institutions (which are often locally based), organizations, and initiatives (Ormond-Parker & Sloggett 2012). In the United States, for instance, The Native American Archives Roundtable[37] was founded in 2005, and a year later the First Archivist Circle[38] issued the Protocols for Native American Archival Materials (Krebs 2012). Moreover, the Administration for Native Americans (ANA) and the the Smithsonian National Museum of the American Indian issued a nearly 300-page reference guide for Indigenous communities interested in establishing archives (ANA 2005). The guide covers an extensive range of subjects, including: 1) why it is important to

[37]http://www2.archivists.org/groups/native-american-archives-roundtable
[38]www.firstarchivistscircle.org/

preserve Native language materials, 2) how to decide what to preserve, 3) what an archive is, 4) how to build an archive infrastructure, 5) how to use existing archives to find language materials, and 6) how to approach archiving costs. As a final example, Alaska's Ahtna community created its own archive, *C'ek'aedi Hwnax*, in 2009 to digitize, curate, and distribute Ahtna language materials—all under OLAC standards and best-practice guidelines undertaken by other archives (Berez et al. 2012, Berez 2013). Such developments exemplify how communities long regarded as objects of study have instead increasingly become leaders in the study and stewardship of their own languages.

5.3 Social networking and archiving A third factor leading to the development of participatory approaches to archiving in linguistics has been a move toward integrating archiving with social networking models (often called "Web 2.0"). Between 2005 and 2010, we saw "the explosive growth of social networking" (Nathan 2011:271), which aims to "link people rather than documents, with a focus on interaction and collaboration instead of passive downloading and viewing of content" (Austin 2014:65). The approach integrating archives and Web 2.0 was pioneered by ELAR in 2010, where "the archive is reconceived as a platform for conducting relationships between information providers (depositors) and information users" (Nathan 2010:111). This integration changes the nature of both access and distribution by allowing parties to negotiate directly with each other—rather than always going through an archivist/archive—which helps address problems such as accessing sensitive materials as well as managing the complexities of growing collections stewarded by small numbers of dedicated staff (Nathan 2010, 2011). This model, of course, shatters traditional boundaries of archiving: The digital archive is not just a place for preserving data; it has been reconceptualized as "a forum for conducting relationships between information providers (usually the depositors) and information users (language speakers, linguists and others)" (Nathan 2011:271). Nathan (2015:53) also discusses the concept of reach, an archive's "multifaceted capacity to successfully provide language resources to those who can gain value from them."

5.4 Development in the archival sciences Finally, as noted by Linn (2014), archival scientists had already been talking about "participatory models" in their own circles since at least the late 2000s. Shilton and Srinivasan (2007), for instance, confronted problematic issues of power entailed by traditional archives. In particular, archives have long directed the selection, collection, and curation of cultural materials from Indigenous communities—who are not involved in the archiving process—to represent those communities: "archives have appropriated the histories of marginalized communities, creat-ing archives *about* rather than *of* the communities (authors' emphasis; 2007:89). To address these problems, Shilton and Srinivasan advocated a Participatory Archiving Model that "encourages community involvement during the appraisal, arrangement, and description phases of creating an archival record" (2007:98). By arising in collaboration with Indigenous communities, a participatory model can help not only to restore power to marginalized people but also to improve the quality

of archives themselves by enhancing their contextual knowledge and value (2007). Huvila (2008:25) built upon this work to formulate the concept of a participatory archive, which has three defining characteristics: 1) *Decentralized curation*, where archivists and participants share curatorial responsibilities; 2) *Radical user orientation*, where the locatability and usability of archived materials takes priority over preservation and the archival process; and 3) *Contextualization of both records and the entire archival process*, which means that archives include knowledge and context provided by others involved in the archiving process, such as a language community. By 2011, participatory models of archiving had become 'sexy' within the archival sciences (Theimer 2011).

5.5 Participatory models of archiving in language documentation Given these four factors, the stage was set for a discourse in documentary linguistics around participatory archiving. By 2011, researchers and archivists were asking themselves how they could expand the usage and impact of archives beyond the limitations of their original conceptions. This entailed a recognition that an archive is not a finished, static repository for data—instead, it is an ever-unfinished research product that involves taking in new information, digitizing old materials, and navigating developments in digital infrastructures, formats, and standards (Albarillo & Thieberger 2009, Holton 2012). Aside from the four factors described above, efforts to expand archives were at least in part also motivated by financial realities:

> In particular, now that some of the major language documentation funding initiatives are coming to an end, the question arises how maximum advantage can be gained from the archiving infrastructures that have been created, for example by encouraging a wider range of people to engage in documenting languages and to deposit their materials into archives, as well as by drawing more users to the various archives (Trilsbeek & König 2014:51–2).

Part of this process involves figuring out who uses archives and for what purposes. Austin (2011), for instance, ascertained that DOBES and ELAR seem to be used primarily by linguists, while ANLA and the California Language Archive are "essentially used by speaker communities or their descendants to access materials for cultural, historical or language-learning purposes." Holton (2012) also found that ANLA users tend to be from Native language communities, who are are often looking for information that is not necessarily, or at least primarily, linguistic. For example, he cited requests for ethnobotanical information, music, and even a eulogy from the nineteenth century—all for non-linguistic purposes. This usage trend in part reflects changing demographics in Alaska, where speaker numbers are declining and language archives often serve as the only records of languages (2012). At the same time, DOBES was exploring how to broaden the impact of its archived data by making it a more accessible resource for scientists and non-scientists interested in language questions (Schwiertz 2012). As part of this effort, DOBES created a new general portal to "attract users

to the archive, facilitate access to the data, and generate new user scenarios and communities" (2012:126).

By 2014, discussions had started exploring the benefits of participatory archiving in documentary linguistics. Green et al. (2011) explained that getting language practitioners involved in both recording their language and making decisions about how to represent it is a good way to encourage not just participation in research but also the long-term availability of data. Furthermore, many linguists (e.g., Gardiner & Thorpe 2014, Garrett 2014, Nathan 2014, Linn 2014, and Woodbury 2014) asserted that participatory archiving models can increase levels of participation in and support for documentary projects among speaker communities, while also maximally engaging audiences and expanding usages for archived material—especially within language communities and other academic disciplines. Simply put, researchers and archivists started to spread the idea that a participatory model might be the best way to get the most out of an archival project.

This has recently led to specific recommendations for participatory models. Woodbury (2014:33) addressed three ways to help archives reach wider audiences "by developing more direct and explicit protocols of communication between documenters and audiences through the medium of language archives." For language documenters, his proposal centers on a "book model," which includes furnishing a guide for exploring a given documentary corpus, explaining the design of the corpus, assigning the corpus to a genre, and providing a narrative about how the data was compiled. For archivists, Woodbury has suggested an "art museum model," based on the fact that such museums curate and provide access to materials. This model includes making the information in archives accessible and discoverable, ensuring that linguists provide adequate descriptions of what they have collected, inviting deposits from people who are not traditional language documenters, holding exhibitions to facilitate public outreach, and getting archives reviewed by both academic and popular outlets to provide public exposure and generate feedback. And for audiences, Woodbury has outlined a 'critic' model that consists of various levels of review for a documentary corpus by a variety of stakeholders (e.g., editors, other language documenters, and archivists).

Linn has recommended a Community-Based Language Archive (CBLA) model, where archives are part of the effort to "bring about community-driven social change through maintaining, revitalizing, or renewing language" (2014:56). Specifically, a CBLA is "an archive or collection that is focused on a language, and that cares for and disseminates documentation that is conducted for, with, and by the language-speaking community within which the documentation takes place and which it affects." (61). Such an archive "actively engages with the relevant community in conducting all levels of documentation, describing and contextualizing, maintenance, and dissemination of information" (61).

In a similar vein, Garrett put forth a model for participant-driven language archiving (PDLA), "an archiving component that assigns role appropriate archiving rights and responsibilities to individuals and communities who participate as 'human subjects' of linguistic research" (2014:68). Although archives have traditionally focused

on building relationships with depositors, a "PDLA's primary objective is to establish direct, web-based, relationships between participants and archives, minimizing the use of depositors as proxies" (69). In the PDLA model, community members become active participants in archiving. They work, for example, to enrich archival resources (e.g., improving or creating metadata) and improve communication between speakers and archives—helping, among other things, to address tricky issues like ongoing informed consent.

Some archives seem already to be moving toward a more participatory model. The Aboriginal and Torres Strait Islander Data Archive, for example, has as an "overarching goal" the "commitment to connect Indigenous Australian communities with research data" (Gardiner & Thorpe 2014:103). In the literature, many of these dedicated discussions of participatory archiving models in documentary linguistics began in 2014, several of which were in the pages of *Language Documentation and Description Volume 12: Special Issue on Language Documentation and Archiving*. This is all quite recent, but it appears that the movement is gaining steam. The next few years will show just where exactly this conversation is going and what its results will be for linguists, other researchers, archivists, and language communities.

6. Conclusion: How are we doing, and where are we going? This overview has divided the history of archiving in language documentation into four general periods:

- Archiving prior to the 1990s, when analog materials were collected and deposited in repositories that were difficult to access by anyone other than a select group of researchers with the requisite dedication, means, and permissions;
- The rise of documentary linguistics in the early 1990s and the subsequent distinction between linguistic *description and documentation*, which engendered both a renewed and redefined focus on archiving and an embrace of digital technology;
- Beginning in the early 2000s, the development of "best practices" for digital archiving and critical reactions addressing the variegated contexts of field situations and ethical issues in language documentation; and
- Since about 2010, developments toward participatory models for linguistic archiving, which break traditional boundaries between depositors, users, and archivists to expand the audiences and uses for archives while involving speaker communities directly in language documentation and archival processes.

Of course, these periods overlap with each other, and the conversations from one period do not—and should not—necessarily end with the beginning of the next.

For example, we are still seeing developments around best practices for digital archiving. Organizations like Innovative Networking in Infrastructure for Endangered Languages (inNET), founded in 2012, are still springing up and seeking better ways to reinforce and extend digital archive networks, facilitate the dissemination of

information to strengthen relationships between archives and the scientific community, promote common archiving standards to help shape archiving policies, and establish relationships between archives and non-scientific communities. Best-practices advocates (e.g., Thieberger 2012) continue to call important attention to the needs for improved methods and tools for language documentation, better metadata and more useful primary data, bigger data storage capacities, and wider promotion of best practices to both linguists and speaker communities.

The critical responses to "best practices" continue as well. Austin (2013:6), for example, says we need to go beyond the normal bounds of best-practice discussions to construct a theory of "meta-documentary linguistics," which he defines as a "documentation of the documentation research itself" that describes "the methods, tools, and theoretical underpinnings for setting up, carrying out and concluding a documentary linguistics research project." Linguists will also keep working on situation-specific solutions to problems in the field that present challenges for a one-size-fits-all approach to archiving (e.g., Bow et al. 2015). Dobrin and Holton (2013:140), for instance, have examined how the priorities and interests of a language community can shift over generations, "reactivating the documentary materials and community-researcher relationships in ways that were not anticipated by anyone involved." Again, Austin (2014:62–65) has more on such critical responses.

The timeline presented here also implies that the development of endangered language archiving since the time of Boas has been an uninterrupted forward trajectory embraced widely by the field. Unfortunately, however, it has not necessarily been the case that linguists—either individually or collectively—have embraced the need for archiving, nor have we agreed upon how to assess the kinds of professional rewards that archiving ought to bring (Thieberger et al. 2015b). Archiving by documentary linguists is still by no means a universal practice, although the number of linguists for whom archiving is a task undertaken at regular intervals—as opposed to waiting until the end of a project or a career—is growing. This has been aided in part by increased awareness of the need to do so, and the falling financial burden of archiving on individuals. Among linguists who do archive regularly, though, most are motivated by personal or professional ideology rather than by discipline-wide expectation or hope of scholarly professional reward.

As an illustration, Gawne et al. (2015) find that very few descriptive linguists are transparent regarding their archiving practice in their publications, including making clear to readers that the primary data is archived, where it is archived, or how to access it. In a survey of more than 100 grammars completed between 2003 and 2012, it was found that only about 10 percent of authors included any reference to the archiving of the primary data upon which the publication was based (2015). This is likely due to the unclear rewards of data management in academia.

In 2010, the Linguistic Society of America passed its *Resolution Recognizing the Scholarly Merit of Language Documentation*,[39] in order to provide academic incentive for archiving by encouraging colleges and universities to consider the products

[39]http://www.linguisticsociety.org/resource/resolution-recognizing-scholarly-merit-language-documentation

of documentation to be valid results of research. The resolution specifically supports the recognition of documentary materials such as the following:

> […] archives of primary data, electronic databases, corpora, critical editions of legacy materials, pedagogical works designed for the use of speech communities, software, websites, or other digital media […] as scholarly contributions to be given weight in the awarding of advanced degrees and in decisions on hiring, tenure, and promotion of faculty. (Linguistic Society of America 2010)

The significance of the resolution is two-fold. First, the resolution acknowledges the value of scholarly work done in the service of increasing linguistic vitality and the inextricability of revitalization efforts from language documentation. Second, it notes that the scholarly products of language documentation go beyond the traditional peer-reviewed journal articles and into the realm of digital products, including archived corpora. Although the resolution is laudable in calling for recognition for archiving practices, it falls short in providing methods to do so. As of yet there is no discipline-wide metric for appraising the quality of preserved linguistic data sets, nor do we know of any departments of linguistics that have made their internal rating system widely available. The number of tenure and promotion cases in which archived collections of annotated data have been given the same weight as journal articles is likely very low. Without the promise of academic attribution, individual linguists have been slow to adopt an archiving workflow or cite primary data in publications.

The value of the historical overview presented here is to point out important trends that have developed within documentary linguistic archiving over the years—especially since the 1990s. At this point, it is also natural to wonder where things may be heading. It seems likely that the next several years will bring further developments in participatory models of archiving. For example, Trilsbeek and König (2014) suggest archives will likely continue to seek expanded audiences (especially in other academic disciplines) and increased community involvement by facilitating the documentation and depositing of archival materials with a range of tools such as smartphones apps. We may also see further development in large-scale, existing e-infrastructure projects (e.g., CLARIN and DARIAH) that will help researchers better share and integrate their work (2014). Moreover, we will also see more critical reactions to participatory models in archiving. What does it mean, for instance, if a community of speakers has no concept of ideas like "digital" and "access" (Robinson 2010, Stenzel 2014)? Importantly, participatory archiving will be part of the process of finding ways to evaluate "the quality, significance and value of language documentation research so that its position alongside such sub-fields as descriptive linguistics and theoretical linguistics can be assured" (Austin 2014:67).

Wherever we end up going, it will surely entail novel and exciting reconceptualizations of archives, expanded audiences, and brand-new uses for language documentation materials.

References

Administration for Native Americans (ANA). 2005. *Native language preservation: A reference guide for establishing archives and repositories*. http://www.aihec.org/our-stories/docs/NativeLanguagePreservationReferenceGuide.pdf

> This is essentially a 'how to' manual for Indigenous communities interested in archiving for the purposes of language documentation and revitalization. As such, it covers a wide range of issues in an informative, practical manner while providing specific, real-world examples. Topics include choosing between (and even building from scratch) a physical or digital archive; concerns of access, copyright, and informed consent; salvaging damaged materials; locating and accessing language materials in existing community, university, government, and private archives; the monetary costs of various aspects of the archival process, including infrastructure maintenance, staffing and labor, and equipment and software; and preserving, copying, and migrating materials.

Albarillo, Emily E. & Nick Thieberger. 2009. Kaipuleohone, the University of Hawai'i's digital ethnographic archive. *Language Documentation & Conservation* 3(1). 1–14. http://hdl.handle.net/10125/4422.

> This article documents the founding and first year of operation of the Kaipuleohone archive in the Department of Linguistics at the University of Hawai'i at Mānoa. The archive is a response to both calls for institutes of higher education to be involved in the creation and preservation of digital collections, as well as the need for preservation of rare endangered language materials. Topics discussed include the purchase of digitization equipment and development of workflow procedures; preservation of materials in ScholarSpace, the University of Hawai'i DSpace repository with an OLAC-compliant metadata catalog; and collaboration with other units on campus like the Music Department, the Anthropology Department, and the Charlene Sato Center for Pidgin, Creole, and Dialect Studies.

Austin, Peter K. 2006. Data and language documentation. In Jost Gippert, Nikolaus P. Himmelmann & Ulrike Mosel (eds.), *Essentials of language documentation* (Trends in Linguistics Studies and Monographs 178), 87–112. Berlin: Mouton de Gruyter.

> A data workflow for language documentation data is presented, alongside some brief overviews of various tools and file formats that the documenter may encounter along the way. The processes of documentation are recording, metadata creation, and capture (or digitization); these are discussed along with backup and file-naming procedures. Processing documentary materials includes linguistic analysis, archiving, and presentation. Although some of the software tools presented are outdated now, the value of this paper lies in recognizing which open formats have remained in use in today's documentary workflow. For example, XML has persisted as a method for storing interlinearized glossed texts.

Austin, Peter K. 2011. Who uses digital language archives? http://www.paradisec.org.au/blog/2011/04/who-uses-digital-language-archives/.

This is a short, informal blog post, but in it Austin explores pivotal questions by asking the leaders of major language archives about their user bases. Austin shares brief replies from ANLA, DOBES, ELAR, and the Survey of California and Other Indian Languages. These responses describe who uses the archives, numbers of visitors (online and in person, if applicable), and their reasons for using the archives. Austin reports important differences: Regional archives are used more by language communities for "cultural, historical or language-learning purposes," but the other archives are used primarily by researchers.

Austin, Peter K. 2013. Language documentation and meta-documentation. In Mari C. Jones & Sarah Ogilvie (eds.), *Keeping languages alive: Documentation, pedagogy and revitalization*, 3–15. Cambridge: Cambridge University Press.

Going beyond traditional ideas of best practices, this piece argues that documentary linguistics also needs a theory of *meta-documentation* that focuses on the theory, methodology, and tools of language documentation—as Austin describes it, "the documentation of the documentation research itself" (4). Austin suggests three different directions for approaching a theory of meta-documentation: 1) deductive, theorizing principles and then applying them to documentation projects; 2) inductive, extracting principles from actual documentation projects; and 3) comparative, examining the role of documentary linguistic metadata in light of what is done in related fields like anthropology and archaeology.

Austin, Peter K. 2014. Language documentation in the 21st century. *JournaLIPP* 3. 57–71.

The author takes a look at the defining characteristics and rise of language documentation, and he discusses changes in the field since 1995. This includes a review of developments in best practices in documentary linguistics, focusing on the efforts of DOBES and the E-MELD project. Importantly, Austin also reflects at length upon critical responses to the emphasis on best practices, which question whether there really is one ideal model for documentary linguistic research. Finally, the author considers developments in archiving, which includes the integration of social networking models and the reconfiguration of relationships between depositors, archives, and users. This article makes a great follow-up companion to Austin and Grenoble's 2007 piece.

Austin, Peter K. & Lenore Grenoble. 2007. Current trends in language documentation. In Peter K. Austin (ed.), *Language Documentation and Description, Volume 4*, 12–25. London: SOAS.

Writing about 15 years after Hale et al.'s seminal 1992 call to action, Austin and Grenoble evaluate the then-current state of language documen-

tation. This includes a review of the theoretical underpinnings and goals of documentary linguistics, discussion of the kinds of projects language documentation can facilitate—especially linguistic research and language revitalization—as well as comments on issues of best practices and access rights. The authors also discuss the factors behind the emergence of documentary linguistics in the late twentieth century (e.g., technological advancements and the development of digital archives). The piece concludes with reflection upon important theoretical issues, including delineating the boundary between documentary and descriptive linguistics as well as defining a "comprehensive" documentation of a language.

Berez, Andrea L. 2013. The digital archiving of endangered language oral traditions: Kaipuleohone at the University of Hawai'i and C'ek'aedi Hwnax in Alaska. *Oral Tradition* 28(2). 261–270.

This article compares two small-scale digital language archives—Kaipuleohone at the University of Hawai'i, and C'ek'aedi Hwnax, which serves the Ahtna Alaska Native community of south central Alaska—in terms of their relevance to oral history research. The former was developed primarily to fulfil the language data preservation needs of an academic department that is known for its linguistic fieldwork in the Asia-Pacific region, while the latter was developed in response to community concerns for the preservation of and access to records of their own linguistic heritage. Both were built according to best practices for digital endangered language preservation and both are members of OLAC, although the audiences they serve are quite different.

Berez, Andrea L., Taña Finnesand & Karen Linnell. 2012. C'ek'aedi Hwnax, the Ahtna Regional Linguistic and Ethnographic Archive. *Language Documentation & Conservation* 6. 237–252. http://hdl.handle.net/10125/4538.

This article details the development of C'ek'aedi Hwnax, the Ahtna Regional Linguistic and Ethnographic Archive in Copper Center, Alaska. C'ek'aedi Hwnax, founded in 2010, was the first OLAC-compliant, Indigenously administered digital language archive in North America. Discussed here are the history of Native Language archiving in the state of Alaska; the identification of the need within the Ahtna community to collect, preserve, and disseminate records of Ahtna language; and the establishment of the archive under the Ahtna Heritage Foundation, including funding, staffing, purchasing equipment, training, digitization, and policy development.

Berez, Andrea L. 2015. Reproducible research in descriptive linguistics: Integrating archiving and citation into the postgraduate curriculum at the University of Hawai'i at Manoa. In Amanda Harris, Nick Thieberger & Linda Barwick (eds.), *Research, records and responsibility: Ten years of PARADISEC*, 39–51. Sydney: Sydney University Press.

> The notion of reproducible research, in which researchers provide the dataset upon which scientific claims are based, is explored in the context of linguistics. As in other fieldwork-based sciences, true replicability is often not possible for linguistics, but reproducibility is often possible. The author discusses an initiative in the linguistics department at the University of Hawai'i to increase reproducibility by requiring PhD students to the archive primary data sets upon which dissertations are based, and then to cite back to that data in the text of the dissertation.

Berez, Andrea & Gary Holton. 2006. Finding the locus of best practice: Technology training in an Alaskan language community. In Linda Barwick & Nicholas Thieberger (eds.), *Sustainable data from digital fieldwork*, 69–86. Sydney: University of Sydney Press.

> The training component of the NSF-sponsored *Dena'ina Archiving, Training and Access* project included two types of training: 1) A three-week class during the summer of 2005 in basic language technology at the Dena'ina Language Institute in Soldotna, Alaska, which was designed for young members of the Dena'ina community; and 2) Four semesters of training in advanced multimedia technology applications to linguistics graduate students. While it had been expected that both learner groups would adapt easily to best practices for language data sustainability, it later became apparent that this expectation ignored community member expectations and interests for the role of technology in language revitalization.

Bird, Steven & Gary Simons. 2003. Seven dimensions of portability for language documentation and description. *Language* 79(3). 57–582.

> This landmark paper discusses seven problem areas, or *dimensions*, that potentially affect the portability of digital data in language documentation and description. These are content, format, discovery, access, citation, preservation, and rights. The authors propose value statements for the field of linguistics with regard to each of these dimensions in order to encourage discussion among linguists toward the development of best practices.

Bow, Catherine, Michael Christie & Brian Devlin. 2015. Shoehorning complex metadata in the Living Archive of Aboriginal Languages. In Amanda Harris, Nick Thieberger & Linda Barwick (eds.), *Research, records and responsibility: Ten years of PARADISEC*, 115–131. Sydney: Sydney University Press.

> The authors present an interesting case study that highlights complications with implementing best-practice approaches in archiving. Specifically, Bow et. al examine challenges involved when attempting to "shoehorn" complex and varied types of data into the standardized approach of an accessible digital archive. For example, the authors discuss conflicts between scientific nomenclature standards and the terms actually used

in language communities; problems trying to fit data into strict categorization protocols, such as when controlled vocabularies oversimplify the complexities of particular Aboriginal language materials; and difficulties determining which materials to include or exclude.

Bowden, John & John Hajek. 2006. When best practice isn't necessarily the best thing to do: Dealing with capacity limits in a developing country. In Linda Barwick & Nicholas Thieberger (eds.), *Sustainable data from digital fieldwork*, 45–56. Sydney: University of Sydney Press.

This one of many papers from the mid-to-late 2000s that questions the relevance of 'best practices' when working with endangered languages in developing countries. The authors examine the success of digital documentation workflows in the Waima'a speaking community of East Timor. The project trained and employed a local assistant in the full digital workflow, to great success, but the authors determined that in the end the archival resources are ultimately of little value to the Waima'a community, which favors instead traditional paper publications.

Boynton, Jessica, Steven Moran, Anthony Aristar & Helen Aristar-Dry. 2006. E-MELD and the School of Best Practices: An ongoing community effort. In Linda Barwick & Nicholas Thieberger (eds.), *Sustainable data from digital fieldwork*, 87–98. Sydney: University of Sydney Press.

This article outlines the development of the Electronic Metastructure for Endangered Languages (E-MELD) project in general, and the School of Best Practice website developed under E-MELD in particular. 'The School' was one component of the five year E-MELD project which was designed to instruct field linguists and anyone in possession of analog endangered language materials in the digitization and care of those items. The article discusses the various stages of development of The School, including identifying the need for such a resource; reaching the appropriate audience; and designing various instructional components like a showroom of case studies and a 'classroom' area with short articles on various topics.

Cameron, Deborah, Elizabeth Frazer, Penelope Harvey, M. B. H. Rampton, & Kay Richardson (eds.). 1992. *Researching language: Issues of power and method*. London: Routledge.

This book presents some of foundational work underlying participatory approaches to archiving. Cameron et al. define and delineate a model of "empowering research," which they describe as research undertaken *on, for, and with* language communities. This model contrasts with 'ethical' and 'advocate' research, both of which fail to incorporate fully interactive methods, the agendas of the people being researched, and a commitment to sharing the knowledge generated through research. In light of the conceptualization of an empowerment model, the editors present four case

studies from their own work to furnish comparative material for reflection upon power and methodology in linguistic research.

Chang, Debbie. 2010. *TAPS: Checklist for responsible archiving of digital language resources*. Dallas: Graduate Institute of Applied Linguistics MA thesis.

The TAPS (target, access, preservation and sustainability) checklist is developed as a metric to assist depositors in assessing the quality of archival practices when selecting a repository for digital endangered language materials. The checklist is then tested at nine digital archives. TAPS was developed for use by nonspecialists by selecting and comparing relevant components from other tools already in existence for assessing digital repositories. These tools are also discussed, although they are not necessarily geared to language repositories, and the author also reflects on the need to develop more formal tools for assessing language archives.

Conathan, Lisa. 2011. Archiving and language documentation. In Peter K. Austin & Julia Sallabank (eds.), *The Cambridge handbook of endangered languages*, 235–254. Cambridge: Cambridge University Press.

Most linguists who regularly deposit their materials in an archive are only familiar with some aspects of the archiving workflow. This article presents the entire archiving process from the point of view of archival science, but with special attention to the needs of endangered language records. The stages in the workflow are appraisal and accession (assessing whether a collection is of enough value to warrant archiving, and the legal process by which an archive acquires materials for deposit), arrangement and description (the hierarchical grouping of materials and the use of metadata to provide information about the records for later finding), preservation (the long-term commitment to care for the physical form and intellectual content of the materials), and access and use (the mobilization of materials for educational and other purposes).

Czaykowska-Higgins, Ewa. 2009. Research models, community engagement, and linguistic fieldwork: Reflections on working within Canadian Indigenous communities. *Language Documentation & Conservation* 3(1). 15–50. http://hdl.handle.net/10125/4423.

This paper proposes a model for ethical linguistic fieldwork based on the author's experiences working in Canadian First Nations communities. The model, termed *community-based language research*, or CBLR, calls for research projects to be designed *for, with*, and *by* members of an endangered language community. In this model, linguists are full collaborative partners in the research, but they are not the primary agents of the research. The paper discusses other models of linguist-focused research and reflects on why one might choose to adopt the CBLR approach when working in Indigenous communities. The author also considers challenges that may arise in collaborative research programs.

Dobrin, Lise M. & Gary Holton. 2013. The documentation lives a life of its own: The temporal transformation of two endangered language archive projects. *Museum Anthropology Review* 7. 140–154.

Dobrin and Holton addresses a critical issue related to archiving, ethics, and access: The viewpoints and interests of a language community can change throughout the life of a project. Case studies explore Dobrin's Arapesh research in Papua New Guinea and Holton's work with Dena'ina in Alaska. In both cases, Indigenous communities became increasingly interested in documenting their own languages and interacting with extant collections of linguistic material held in digital archives. As such, the authors advise that documentary linguistics and archiving be approached as works in progress that are attuned to the wishes of language communities.

Doorn, Peter & Heiko Tjalsma. 2007. Introduction: Archiving research data. *Archival Science* 7(1). 1–20.

Coming from the discipline of archival science, this article introduces the concept of archiving research data (as opposed to archiving public records). Doorn and Tjalsma provide very useful information concerning the historical development of archives for research data as well as the advent and challenges of preserving digital information. In the latter half of the article, the authors survey the main issues and contemporary trends regarding demands on data archiving. This includes discussion of organizational infrastructures for data facilities, data strategies at national and international levels, issues of open access and data availability, and more.

Dwyer, Arienne M. 2006. Ethics and practicalities of cooperative fieldwork and analysis. In Jost Gippert, Nikolaus P. Himmelmann & Ulrike Mosel (eds.), *Essentials of language documentation* (Trends in Linguistics Studies and Monographs 178), 31–66. Berlin: Mouton de Gruyter.

The first half of this chapter introduce basic ethical concepts related to language documentation (e.g., rights and responsibilities of fieldworkers and informed consent), and also legal aspects of data ownership and copyright. The second half is much more practical in nature, and offers a framework for ethical language documentation under the aegis of 'the five Cs': criteria, contacts, cold calls, community, and compensation. The value of this chapter is its clarity of presentation for those new to fieldwork and language documentation.

Evans, Nicholas & Hans-Jurgen Sasse. 2004. Searching for meaning in the Library of Babel: Field semantics and problems of digital archiving. In Linda Barwick, Allan Marett, Jane Simpson & Amanda Harris (eds.), *Researchers, communities, institutions and sound recordings*, 1–31. Sydney: University of Sydney.

The authors contribute to best-practice discussions by exploring challenges involving the archiving of semantic documentation. Evans and

Sasse assert that technological advancements have greatly expanded our abilities to collect and store sound recordings, but this has not necessarily been accompanied by parallel developments in capturing and conveying the *meaning* of these recording (e.g., explaining gestures, cultural context, or language-specific semantic relationships). The authors present case studies to illustrate the problem, and they advocate developing appropriate archiving technology—such as multi-layered annotations created over time and involving contributions from a variety of relevant parties—to facilitate the documentation of meaning.

Evans, Nicholas & Alan Dench. 2006. Introduction: Catching language. In Felix K. Ameka, Alan Dench & Nicholas Evans (eds.), *Catching language: The standing challenge of grammar writing* (Trends in Linguistics Studies and Monographs 167), 1–39. Berlin: Mouton de Gruyter.

This is, first and foremost, the introduction to a volume about writing descriptive grammars, but Evans and Dench nonetheless engage ideas very relevant to archiving in documentary linguistics. For example, they discuss the progression of technology that has changed not only the kinds of linguistic data we collect but also how we interact with, store, and preserve this information. This includes the expectation that digital archives will be used increasingly for purposes such as testing linguistic analyses, but this entails significant implications for questions of access and data-stewardship best practices.

Gardiner, Gabrielle & Kirsten Thorpe. 2014. The Aboriginal and Torres Strait Islander Data Archive: Connecting communities and research data. In David Nathan & Peter K. Austin (eds.), *Language Documentation and Description, Volume 12: Special Issue on Language Documentation and Archiving,* 103–119. London: SOAS.

Gardiner and Thorpe overview ATSIDA, a part of the Australian Data Archive that places an emphasis on collaboration and relationship building with researchers and language communities. The authors discuss the development, structure, and stakeholders of ATSIDA. They describe the archive's operations and furnish a look into the particulars of data curation and preservation as well as protocols designed to connect language communities with linguistic, cultural, and historical research data. Gardiner and Thorpe also explore the challenges and opportunities that have arisen during the establishment of ATSIDA, which should be valuable for anyone interested in participatory archiving.

Garrett, Edward. 2014. Participant-driven language archiving. In David Nathan & Peter K. Austin (eds.), *Language Documentation and Description, Volume 12: Special Issue on Language Documentation and Archiving*, 68–84. London: SOAS.

In this article pertaining to participatory models of archiving, Garrett outlines the motivations and preliminary requirements for implementing what he calls *participant-driven language archiving* (PDLA). He claims

that existing archives have focused too much on building relationships solely with depositors, ignoring opportunities to involve the people who are the 'human subjects' of documentary linguistic research. In particular, Garrett explains that participants can enrich archived resources and address challenges of informed consent. The author explores some of the potentials and challenges of the PDLA model, including negotiating access, repatriating resources, and facilitating payment for language consultants.

Garrett, Andrew & Lisa Conathan. 2009. Archives, communities, and linguists: Negotiating access to language documentation. *Linguistic Society of America Annual Meeting*. http://www.ailla.utexas.org/site/lsa_olac09/conathan-garrett_lsa_olac09.pdf

Garrett & Conathan present several case studies from their own experiences to illustrate conflicts involving access to archived materials related to languages of California and the western United States. Such problems have hindered collaboration between archives, linguists, and heritage communities. Examples include failures to create access protocols, attempts by linguists or language communities to restrict access, and "turf disputes" between parties with stakes in archived materials. Garrett & Conathan review some archival protocols designed to help facilitate collaboration with communities while advocating for their rights, and they discuss lessons learned from these case studies.

Gehr, Susan. 2013. *Breath of Life: Revitalizing California's native languages through archives*. San Jose: San Jose State University MA thesis.

This thesis is an oral history of the *Breath of Life* workshops held biennially since 1996 by the Advocates for Indigenous California Language Survival at the University of California, Berkeley. Gehr begins by surveying the history of Native American language revitalization efforts since the mid-twentieth century, with special focus on the role of archives and archived/archival material. She interviews participants, linguists, and archivists involved in the workshop and presents thoughts about future revitalization efforts.

Gerdts, Donna. 2010. Beyond expertise: The role of the linguist in language revitalization programs. In Lenore A. Grenoble & N. Louanna Furbee (eds.), *Language Documentation: Practice and Values*, 173–192. Amsterdam, Philadelphia: John Benjamins Publishing Company.

Based on her own experiences with the Halkomelem language, the author addresses the tension that can sometimes arise between members of an endangered language community and linguists in the context of language revitalization. She discusses the kinds of skills that linguists can bring to a revitalization project, and potential misunderstandings about linguists' roles and abilities. She also presents her experiences of what Na-

tive language communities tend to want an academic linguist to provide, and what the needs of revitalization programs are.

Gippert, Jost. 2006. Linguistic documentation and the encoding of textual materials. In Jost Gippert, Nikolaus P. Himmelmann & Ulrike Mosel (eds.), *Essentials of language documentation* (Trends in Linguistics Studies and Monographs 178), 337–361. Berlin: Mouton de Gruyter.

The first half of this chapter discusses issues of character encoding, especially as it applies to presenting non-English (rather, non-ASCII) characters in textual materials. 8-bit to 32-bit encoding and Unicode are presented, along with some recommendations for avoiding character encoding problems (much of the discussion will be useful today, if one is in possession of older digital materials). The second half of the chapter discusses content-driven markup of textual structure, and proposes HTML as a potential way to get the benefits of true markup—XML—without too much trouble. XML is also discussed briefly.

Golla, Victor. 1995. The records of American Indian linguistics. In Sydel Silverman & Nancy J. Parezo (eds.), *Preserving the anthropological record*, 143–157. New York: Wenner-Gren Foundation for Anthropological Research.

Golla's chapter summarizes vital information about the history of linguistic anthropology in North America, primarily since the late nineteenth century. He discusses the various types of records that have been created and collected by scholars, which includes lexical compilations, texts, file slips, sound and video recordings, and digital files. Golla also describes the history and collections of some of the most important archives preserving Native American linguistic material. The chapter concludes with a look at the challenges of preserving these records while properly training future generation of scholars to steward and study them.

Good, Jeff. 2011. Data and language documentation. In Peter K. Austin & Julia Sallabank (eds.), *The Cambridge handbook of endangered languages*, 212–234. Cambridge: Cambridge University Press.

Good discusses conceptual issues surrounding the nature of data in language documentation, which includes primary data as comprised of direct recordings of speech events and the transcriptions, or written representations, of those events. Primary data are contrasted with descriptive resources like texts, dictionaries, and grammars. The author also discusses the differences between data structure on the one hand, and implementation or presentation on the other. Also presented are the notions of proprietary versus open formats; markup; archival, working, and presentation formats; and metadata.

Green, Jennifer, Gail Woods & Ben Foley. 2011. Looking at language: Appropriate design for sign language resources in remote Australian Indigenous communities. In

Nick Thieberger, Linda Barwick, Rosey Billington & Jill Vaughan (eds.), *Sustainable data from digital research: Humanities perspectives on digital scholarship*, 66–89. Melbourne: Custom Book Centre.

> Sign languages are common in Arandic communities in Central Australia. These endangered languages are generally used by people who also use spoken language, and are culturally valued for use in certain rituals, and in situations like hunting and at times when audibility is disadvantageous. The authors describe a project to document, preserve, and promote Arandic sign through digital resource development. The project was designed to maintain respect for the dignity and desires of the communities by recording video in natural bush settings, by eliciting in local languages, and through careful editing. The authors also describe their data storage, annotation, and web publication procedures.

Hale, Ken, Michael Krauss, Lucille J. Watahomigie, Akira Y. Yamamoto, Colette Craig, LaVerne Masayesva Jeanne & Nora C. England. 1992. Endangered languages. *Language* 68(1). 1–42.

> This collection of six essays appeared as a collection in the journal *Language* following a symposium at the 1991 Linguistic Society of America annual meeting. Hale's first essay introduces the collection and touches on language endangerment as the potential loss of cultural and intellectual diversity. Krauss's celebrated essay, described more fully below, is a call to arms for linguists to organize against language endangerment. Watahomigie and Yamamoto discuss reactions to language loss in Native America with particular emphasis on Hualapai in reference to both the American Indian Languages Development Institute and the Native American Languages Act. Craig discusses legislation from the 1980s in Nicaragua known as the Autonomy project under which several language planning projects were implemented for the Indigenous languages there; Craig focuses on the Rama Language Project and its successes. Jeanne proposes a Native American Language Center, which would be dedicated to a range of support and research activities for Native American languages, and staffed by and serving the concerns of speakers of Native American languages. England reflects on the role of Mayan language scholarship in Guatemala. Hale's second essay considers more deeply the value of linguistic diversity to humanity.

Himmelmann, Nikolaus. 1998. Documentary and descriptive linguistics. *Linguistics* 36. 161–95.

> In this, the definitive article now commonly cited as launching the subfield of language documentation as distinct from descriptive linguistics, the author describes the activities of language documentation as the creation of "a record of the linguistic practices and traditions of a speech community" (166). Practical and theoretical considerations are presented for the

four steps of language documentation: 1) decisions about which data to collect; 2) recording the data; 3) annotation, the transcription and translation of the data with commentary; and 4) preservation and presentation. Also discussed are ethical and privacy considerations, as well as guidelines for collecting a documentation that is varied in genre and spontaneity.

Himmelmann, Nikolaus. 2006. Language documentation: What is it and what is it good for? In Jost Gippert, Nikolaus P. Himmelmann & Ulrike Mosel (eds.), *Essentials of language documentation* (Trends in Linguistics Studies and Monographs 178), 1–30. Berlin: Mouton de Gruyter.

This is the introductory chapter to the first edited volume on language documentation proper. Eight years after the publication of Himmelmann 1998, the author further refines this field of linguistic inquiry, and defines a language documentation as "a lasting, multipurpose record of a language" (1). He also discusses the value of language documentation to other disciplines both inside and outside of linguistics, and presents a format for a documentation. This format includes records of observable linguistic behavior; indications of metalinguistic knowledge including paradigms, usage scenarios, and other generalizations; lexical databases; and the apparatus. The *apparatus* is defined as the set of information that is used to interpret and understand the rest of the documentation, including metadata, transcriptions, translations, ethnographic sketches, glossing conventions, and the like.

Hinton, Leanne. 2001. Language revitalization: An overview. In Leanne Hinton & Kenneth Hale (eds.), *The green book of language revitalization in practice*, 3–18. San Diego: Academic Press.

In this first chapter of a guide to language revitalization, Hinton surveys language shift and endangerment as well as various approaches to revitalization. This includes discussion of the role of archives in revitalization. For instance, archives play a vital part at the starting point of revitalization efforts, when communities seek out existing material on their languages. Archived materials also serve as critical resources for the creation of language-teaching materials, such as reference grammars and language lessons. Accordingly, Hinton discusses programs like Breath of Life, which aim to increase access to archives for Indigenous communities.

Hinton, Leanne. 2005. What to preserve: A viewpoint from linguistics. In Administration for Native Americans (ed.), *Native language preservation: A reference guide for establishing archives and repositories*, 24–26. Washington, D.C.

This is a very brief selection from a guidebook for Indigenous communities about archival matters related to their languages (see ANA 2005 above). Nonetheless, Hinton touches upon several important themes and issues: Indigenous communities are increasingly enlisting archives in the

service of language maintenance and revitalization, particularly in the creation of dictionaries, curricula, and the like; archived language materials often lack crucial metadata, such as detailed annotations and transcriptions; and speakers and collectors must determine together the access conditions for their archived data.

Holton, Gary. 2012. Language archives: They're not just for linguists any more. In Frank Seifart, Geoffrey Haig, Nikolaus P. Himmelmann, Dagmar Jung, Anna Margetts & Paul Trilsbeek (eds.), *Language Documentation & Conservation Special Publication No. 3, Potentials of Language Documentation: Methods, Analyses, and Utilization*, 111–117. Honolulu: University of Hawai'i Press. https://scholarspace.manoa.hawaii.edu/handle/10125/4523.

In this short chapter, Holton provides an insightful look at how language archives are actually used. He draws upon his experience at ANLA to present examples demonstrating that the audiences and uses of an archive can go far beyond the founding aims of linguists simply preserving language data. Holton describes, for example, an ethnoastronomy project relying upon ANLA's archived sources. He also discusses community efforts to revitalize Eyak, where ANLA is the only surviving source of information about the language. Thus, Holton advises archives to facilitate non-linguistic uses for their materials and to position linguistic data to create derived products in the service of language revitalization.

Holton, Gary. 2014. Mediating language documentation. In David Nathan & Peter K. Austin (eds.), *Language Documentation and Description, Volume 12: Special Issue on Language Documentation and Archiving*, 37–52. London: SOAS.

A recurring thread in best-practice discussions concerns negotiating and facilitating access to archived materials, but Holton calls attention to a critical point: Providing access alone is not enough to ensure that such materials are actually used. This problem is particularly significant when language maintenance and revitalization efforts are involved. As such, this article proposes that archives must *mediate* between collections and users. Using his experiences at ANLA as a case study, Holton suggests how archives can make their materials more accessible and more relevant to language communities, which requires that archives work closely with the people they aim to serve.

Holton, Gary, Andrea L. Berez, & Sadie Williams. 2006. Building the Dena'ina language archive. In Laurel Evelyn Dyson, Max Hendricks, & Stephen Grant (eds.), *Information technology and indigenous people*, 205–209. Hershey: Idea Group.

This paper discusses the development of the Dena'ina Language Archive, a digital archiving project created under the aegis of the NSF-sponsored Dena'ina Archiving, Training, and Access project. Dena'ina is an Athabascan language spoken in south central Alaska, and under this project the

Dena'ina language materials in ANLA were digitized and made available online. Metadata were made discoverable through OLAC and were embedded in a value-added online portal known as qenaga.org (*qenaga* means 'language' in Dena'ina). The project represented an early digital collaboration between linguists, language technologists, and community members in an Alaska Native language.

Huvila, Isto. 2008. Participatory archive: Towards decentralised curation, radical user orientation, and broader contextualisation of records management. *Archival Science* 8(1). 15–36.

Building upon the groundwork laid by Shilton and Srinivasan (2007), Huvila explicitly formulates the concept of a "participatory archive." He describes the development of this idea through a case study of two projects building digital historical archives in Finland. The three defining characteristics of a participatory archive are: 1) *decentralized curation*, 2) *radical user orientation,* and 3) *contextualization of both records and the entire archival process*. This model radically reconfigures the responsibilities of and interactions between archivists, depositors, and users throughout the archival process.

Innes, Pamela. 2010. Ethical problems in archival research: Beyond accessibility. *Language & Communication* 30(3). 198–203.

Innes offers a brief-but-significant exploration of ethical considerations in archiving. This article relates her experiences working to prepare for publication Mary Haas' archived notes on Mvskoke. Innes encounters a major problem: Some members of the language community felt that particular narratives were inappropriate for certain audiences, and that other texts were even dangerous. This case study raises critical issues of obtaining and documenting informed consent, managing access to archived materials, and navigating tensions between the language ideologies of a community and those of scholars who expect data to be open and available.

Innes, Pamela & Erin Debenport. 2010. Editors' introduction. *Language & Communication* 30(3). 159–161.

Although this is but a short introduction to an entire journal issue devoted to ethics and language documentation, it is worth reading to hear from the editors themselves about what motivated the production of such a volume: Documentary linguistics had spent plenty of time and resources developing "best practices" for many of the technological and archival aspects of documentation, but the same dedication had not been committed to exploring the ethical implications of these aspects.

Johnson, Heidi. 2004. Language documentation and archiving, or how to build a better corpus. In Peter K. Austin (ed.), *Language Documentation and Description Volume 2*, 140–153. London: SOAS.

Johnson's article is a must-read primer for understanding the relationship between archiving and language documentation. She offers an informative review of the role of archiving in early and modern documentary linguistics, along with a description of the progress of technology used in such endeavors. For anyone looking for a quick guide on where archiving fits into documentary linguistics, Johnson provides a breakdown explaining "who should archive, and where, why, when, and how one should archive" (3). The bulk of this article covers the ethos and best-practice methodology of archiving language documentation, spanning topics such as data formats, access permissions, item labelling, and metadata.

Krauss, Michael E. 1974. Alaska Native language legislation. *International Journal of American Linguistics* 40(2). 150–152.

This brief describes the 1972 passing of four bills in the Alaska State Legislature concerning Alaska Native Languages. Senate Bill 421 authorized mandatory bilingual education in state schools where students speak a Native language; Senate Bill 422 authorized the establishment of the Alaska Native Language Center at the University of Alaska; Senate Bills 424 and 423 appropriated funds to the other two bills respectively. The text of all four bills are presented.

Krauss, Michael. 1992. The world's languages in crisis. *Language* 68. 4–10.

The most-cited of the essays edited by Hale and appearing together in *Language* (1992), this piece starts by citing some sobering figures about language vitality in North America and beyond. Krauss proposes a cline of statuses for vitality including "endangered," "moribund," and "safe." Endangered languages are compared to endangered species, and the author draws parallels about the expected reaction of the scientific community in face of endangerment.The essays ends with the admonishment that linguistics not "go down in history as the only science that presided obliviously over the disappearance of 90% of the very field to which it is dedicated" (10).

Krebs, Allison Boucher. 2012. Native America's twenty-first-century right to know. *Archival Science* 12. 173–190.

This article provides valuable historical and cultural context related to the increasing self-empowerment of Indigenous people in the United States over the course of the last several decades. Krebs evaluates two initiatives supporting the development of libraries, archives, and information centers for Indigenous communities: 1) the Institute of Museum and Library Services' Grants to Indian Tribes, and 2) the Fourth Museum of the National Museum of the American Indian. Of particular value here is the overview of activist Vine Deloria Jr.'s advocacy for an Indigenous 'right to know,' along with Krebs' timeline, which breaks down relevant developments re-

garding the relevant interplay between federal, citizen, and professional organizations.

Linn, Mary S. 2014. Living archives: A community-based language archive model. In David Nathan & Peter K. Austin (eds.), *Language Documentation and Description, Volume 12: Special Issue on Language Documentation and Archiving*, 53–67. London: SOAS.

Linn outlines a proposal for a Community-Based Language Archive (CBLA), a radical departure from traditional models of archiving. In a CBLA, the archive engages with a language community throughout every component of the archiving process. Along with explaining the concept, Linn provides a case study of her experiences integrating the CBLA model while transforming collections and building new ones at the Sam Noble Oklahoma Museum of Natural History. This article also includes a useful overview of literature exploring participatory and community-based approaches to archiving and language research.

Linguistic Society of America. 2010. *Resolution recognizing the scholarly merit of language documentation*. http://www.linguisticsociety.org/resource/resolution-recognizing-scholarly-merit-language-documentation

This resolution, passed in 2010 by 'a sense of majority' within the Linguistic Society of America, declares the outputs of language documentation for scholarly and community use—including dictionaries, grammars, text collections, digital data sets, web products, and more—to be considered academic output for the purposes of hiring, tenure, and promotion.

Macri, Martha & James Sarmento. 2010. Respecting privacy: Ethical and pragmatic considerations. *Language & Communication* 30(3). 192–197.

Macri & Sarmento provide a helpful, brief case study that illustrates ethical problems involved in archiving sensitive materials. This article details issues encountered by researchers transcribing and coding notes in the J. P. Harrington Database Project, which aims to create resources for use by a variety of academic and non-academic audiences. In particular, notes have involved gossip and hearsay, sensitive customs, sacred sites, and even potentially physically dangerous knowledge. Macri and Sarmento raise important questions about conflicts between international standards and Indigenous communities, and deciding who—if anyone—can speak for a community.

Nathan, David. 2009. The soundness of documentation: Towards an epistemology for audio in documentary linguistics. *Journal of the International Association of Sound Archives* 33. 50–63.

A critique of so-called 'best practices' in language documentation that encourage the use of ever-advancing technologies without truly understanding the goals and impacts of audio recording, this article encourages *critical listening* when making recordings. One aspect of this includes giving

serious consideration to signal-to-noise ratio: Determining what counts as signal and what counts as noise should be guided by the aims of the documentation project. Another aspect is the consideration of psychoacoustic effects of capturing spatial information through advanced stereo techniques like ORTF. It is argued that critical listening will produce better documentation than carelessly adopting the latest advancements in media like video.

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.

Nathan describes how ELAR has attempted to implement the properties of Web 2.0 (e.g., social networking and interaction online) in order to restructure and enhance the experiences of its depositors and users. This moves the archive beyond a traditional role as a data repository. Instead, ELAR now aims to facilitate relationships between parties involved in archiving. Nathan argues that this approach is better equipped for managing issues of access (especially sensitivities and restrictions) as well as the diversity of resources held by ELAR.

Nathan, David. 2011. Digital archiving. In Peter K. Austin & Julia Sallabank (eds.), *The Cambridge handbook of endangered languages*, 255–273. Cambridge: Cambridge University Press.

In some sense this handbook chapter is a companion to Conathan 2011, in that it addresses specifically the digital aspects of archiving within the larger framework of archive curation. The author discusses the nature of digital data and digital encoding; several sections are dedicated to describing extant digital archives, their services, and their policies; and the author ends by touching on data migration, the archiving of video, and archive assessment.

Nathan, David. 2014. Access and accessibility at ELAR, an archive for endangered languages documentation. In David Nathan & Peter K. Austin (eds.), *Language Documentation and Description, Volume 12: Special Issue on Language Documentation and Archiving*, 187–208. London: SOAS.

This article illustrates a shift *in practice* toward a participatory model for ELAR, one of the most important archives involved in documentary linguistics. Nathan describes how ELAR has integrated a social networking approach to reconfigure the way the archive interacts with—and facilitates interactions between—its depositors and users. This, of course, is a departure from the traditional 'one-way street' model of archiving. He walks the reader through the ELAR protocol for navigating resources as well as searching and browsing, and he explains how this approach enhances access for various types of users.

Nathan, David. 2015. On the reach of digital language archives. In Amanda Harris, Nick Thieberger & Linda Barwick (eds.), *Research, records and responsibility: Ten years of PARADISEC*, 53–79. Sydney: Sydney University Press.

> The author discusses the concept of *reach* as a measurement of the capacity of an archive to provide materials to the appropriate audience. Ten facets of reach are defined: acquisition, audiences, discovery, delivery, access management, information accessibility, promotion, communication ecology, feedback channels, and temporal reach.

Nathan, David & Peter K. Austin. 2004. Reconceiving metadata: Language documentation through thick and thin. In Peter K. Austin (ed.), *Language Documentation and Description, Volume 2*, 179–187. London: SOAS.

> In this critical addition to best-practices discussions, Nathan and Austin put forth a distinction between 'thin' and 'thick' metadata. They argue that most attention in documentary linguistics goes toward the former, which does not provide enough value for linguists and speech communities interested in working with language materials. *Thin* metadata is primarily for cataloguing, mostly aimed at facilitating resource discovery. On the other hand, *thick* metadata involves more context—such as transcriptions, commentary, and time-aligned annotations—and is intended to enhance the access and use of archived materials.

Nathan, David & Peter K. Austin. 2014. Editors' introduction. In David Nathan & Peter K. Austin (eds.), *Language Documentation and Description, Volume 12: Special Issue on Language Documentation and Archiving*, 4–16. London: SOAS.

> This is the introduction to "the first journal publication symmetrically targeted at both language documentation and archiving" (6). As such, it presents a helpful overview of the papers inside the publication. However, this chapter also offers value in its own right. In particular, Nathan and Austin furnish a useful glance at the relationship between archiving and language documentation. They also point out issues that recur throughout their volume: community curation, the promotion of archived language resources, the contextualization of archived materials, the 'form' of documented material (e.g., structure and granularity), and the conceptualization of archiving as a publishing.

Nordhoff, Sebastian & Harald Hammarström. 2014. Archiving grammatical descriptions. In David Nathan & Peter K. Austin (eds.), *Language Documentation and Description, Volume 12: Special Issue on Language Documentation and Archiving*, 164–186. London: SOAS.

> Much of the best-practices talk in archiving has revolved around primary data, and so Nordhoff and Hammarström call attention to the need for a methodology of archiving *grammatical descriptions*. Grammatical descriptions are based on primary data but entail different information types

and structures, and their users have specific needs for retrieving information at certain levels of granularity. Given these differences, the authors recommend a semantic-markup architecture based upon the Text Encoding Initiative (TEI). They present a systematic appraisal of existing TEI schema as well as special TEI elements, which could facilitate the archiving and access of grammatical descriptions.

O'Meara, Carolyn & Jeff Good. 2010. Ethical issues in legacy language resources. *Language & Communication* 30(3). 162–170.

This article offers a critical contribution to best-practice recommendations in archiving. O'Meara and Good examine the pilot phase of the Northeastern North American Indigenous Languages Archive to probe vital ethical issues surrounding the establishment of rights and access to archived language resources. In particular, the authors raise questions related to four areas: 1) the notion of 'community,' 2) establishing rights and access retroactively, 3) establishing rights and access to resources without an identifiable copyright holder, and 4) navigating concerns associated with sensitive materials.

Ormond-Parker, Lyndon & Robyn Sloggett. 2012. Local archives and community collecting in the digital age. *Archival Science* 12. 191–212.

Ormond-Parker and Sloggett focus on Aboriginal communities in Australia to take an important look at the increasing self-empowerment of Indigenous people in archiving. This, of course, has been fueled in part by the proliferation of digital tools and technology. The authors identify the benefits of such developments for these communities, which include economic development, community empowerment, and the creation of opportunities for young people. At the same time, however, Ormond-Parker and Sloggett argue that community-driven efforts are often not equipped to handle the various threats inherent to digital archiving. As a solution, the authors recommend a national framework to support community-controlled archives.

Rehg, Kenneth L. 2007. The Language Documentation and Conservation Initiative at the University of Hawai'i at Mānoa. In D. Victoria Rau & Margaret Florey (eds.), *Language Documentation and Conservation, Special Publication No. 1, Documenting and Revitalizing Austronesian Languages*, 13–24. Honolulu: University of Hawaii Press. http://hdl.handle.net/10125/135.

Although it focuses on one initiative at a single university, Rehg's piece is a useful treatment about putting into practice some of the most crucial themes from the history of archiving in linguistics. This includes best-practices training for linguists in the theory, methods, and ethics of language documentation. Rehg also describes efforts to create collaborative research models that benefit linguists and non-linguists alike. As such, he outlines then-developing plans to create a digital archive at the University

of Hawai'i, one that safely stores data in accordance with the desires of speech communities. This archive, named Kaipuleohone, opened in 2008.

Robinson, Laura. 2006. Archiving directly from the field. In Linda Barwick & Nicholas Thieberger (eds.), *Sustainable data from digital fieldwork*, 23–32. Sydney: University of Sydney Press.

Depositing materials into an archive on a regular basis has not always been part of the linguist's workflow, so this author discusses her own procedures for developing a regular archiving practice while on a year-long fieldwork trip to the Philippines. She describes her solar power configuration, her digitization workflow, and her metadata documentation workflow. She sent her data regularly to PARADISEC via the postal service during this period. Although archiving from the field has become *de rigeur* since this article was written, it is important to remember that this was not always common practice.

Robinson, Laura. 2010. Informed consent among analog people in a digital world. *Language & Communication* 30. 186–191.

The ethical bind that comes with obtaining informed consent about digital dissemination of language data from people with no knowledge of the internet is discussed in the context of the author's fieldwork with a remote community of Agta speakers in the Philippines. Institutional review boards will often allow oral, as opposed to written, consent in cases of non-literate consultants, but the author argues that because researchers have a moral obligation for informed consent, consultants with no knowledge of the internet could be considered a vulnerable class when the researcher wants to disseminate data online. The two solutions available—nondissemination of that data versus assuming speakers would want their data to be disseminated online "if they only understood"—are presented as equally paternalistic.

Schroeter, Ronald & Nick Thieberger. 2006. EOPAS, The EthnoER online representation of interlinear text. In Linda Barwick & Nicholas Thieberger (eds.), *Sustainable data from digital fieldwork*, 99–124. Sydney: University of Sydney Press.

The authors describe the initial development phase of EOPAS, a tool designed to convert the normal outputs of a digital language documentation workflow into presentation formats suitable for online viewing. The tool primarily works with time aligned transcripts (e.g., those from ELAN and Transcriber) and interlinear text (e.g., Toolbox). EOPAS transforms the validated XML output of those other tools into EOPAS XML via stylesheets. The resultant file is then stored alongside the original media file for display; at the time, a tool known as Annodex was being explored as a streaming delivery option, and other HTML displays were also developed.

Schwiertz, Gabriele. 2012. Online presentation and accessibility of endangered languages data: The general portal to the DOBES Archive. In Frank Seifart, Geoffrey Haig, Nikolaus P. Himmelmann, Dagmar Jung, Anna Margetts & Paul Trilsbeek (eds.), *Language Documentation & Conservation Special Publication No. 3, Potentials of Language Documentation: Methods, Analyses, and Utilization*, 126–128. Honolulu: University of Hawaii Press. http://hdl.handle.net/10125/4526.

This very brief chapter belongs to conversations about expanding the audiences and uses of archives. As one of the primary funders of endangered language documentation work, DOBES maintains a large archival collection of data from its projects. In order to expand the archive's user base and increase access to materials, DOBES launched a general web portal in March 2013. With a bare-bones approach, Schwiertz walks through the structure and features of the portal, describing how it aims to serve researchers, depositors, language communities, and the general public.

Shilton, Katie & Ramesh Srinivasan. 2007. Participatory appraisal and arrangement for multicultural archival collections. *Archivaria* 63. 87–101.

Shilton & Srinivasan offer perhaps the first contribution to the discussion around participatory models in archival sciences. As institutions creating collective memory, archives often fail to include different ethnic and cultural communities in the foundational archival practices of appraisal, arrangement, and description. This contributes to imbalances in power and representation for historically marginalized people. As such, Shilton and Srinivasan recommend 'rearticulating' appraisal and arrangement as community-driven, participatory processes. In doing so, a participatory model can improve the quality of archives, preserve more local knowledge and context, and help empower people traditionally left out of the archiving process.

Stenzel, Kristine. 2014. The pleasures and pitfalls of a "participatory" documentation project: An experience in northwestern Amazonia. *Language Documentation & Conservation* 8. 287–306. http://hdl.handle.net/10125/24608.

Stenzel presents her experiences documenting languages in the Amazon, providing a critical response in the ongoing discourse around collaborative and participatory research models in documentary linguistics. The piece is primarily a narrative history of Stenzel's four-year project, with perhaps the most valuable contribution coming from her discussion of the various 'pitfalls' she encountered. This includes a host of "logistical, technical, cultural, and philosophical" challenges, which all have a bearing on important issues like project sustainability, accountability, and the complex human relationships that provide the underpinnings for collaborative projects.

Theimer, Kate. 2011. Exploring the participatory archives: What, who, where, and why. *Annual Meetings of the Society of American Archivists*. http://www.slideshare.net/ktheimer/theimer-participatory-archives-saa-2011.

> Although this is a brief conference presentation, Theimer's contribution is another good example of conversations in archival sciences about participatory models of archiving, which had been taking place for several years before penetrating the field of linguistics. Theimer helpfully introduces her concept and definition of 'participatory archiving,' which entails contributing knowledge and resources in a (typically) online environment. Moreover, she outlines a distinction between engagement and participation. This is a slideshow rather than an article, so this piece is best considered together with a paper like Shilton and Srinivasan 2007 or Huvila 2008.

Thieberger, Nicholas. 1994. Report on the AIATSIS visiting research fellowship, Aboriginal Studies Electronic Data Archive: A report to AIATSIS Council on the conclusion of the Visiting Research Fellowship. http://trove.nla.gov.au/work/33785959?q&versionId=41559386.

> This report contains a summary of the structure and operations of the Aboriginal Studies Electronic Data Archive, which was established in 1991 and is now integrated with AIATSIS. This piece also describes various projects undertaken by the archive, including the AIATSIS Aboriginal Dictionaries Project, a workshop on copyright, and more. The value of this report primarily lies in its historical information and thorough accounting of the activities of what might be the first digital archive dedicated to endangered languages.

Thieberger, Nicholas. 2010. Anxious respect for linguistic data: The Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC) and the Resource Network for Linguistic Diversity (RNLD). In Margaret Florey (ed.), *Endangered Languages of Austronesia*, 141–158. Oxford: Oxford University Press.

> This chapter is a prime example of best-practice discussions in linguistic archiving: Thieberger presents a thorough walkthrough of recommended methods for creating and storing language documentation data. He draws upon his own experience documenting the Oceanic language South Efate and working with PARADISEC to provide specific advice for proper data management and workflows, making data locatable and citable, choosing file formats and software tools, and more. Additionally, this chapter discusses the operations of PARADISEC and stresses the importance of training academics and speaker communities to employ best-practice methods in the documentation of endangered languages.

Thieberger, Nicholas. 2012. Using language documentation data in a broader context. In Frank Seifart, Geoffrey Haig, Nikolaus P. Himmelmann, Dagmar Jung, Anna

Margetts, & Paul Trilsbeek (eds.), *Language Documentation & Conservation Special Publication No. 3, Potentials of Language Documentation: Methods, Analyses, and Utilization*, 129–134. Honolulu: University of Hawaii Press. http://hdl.handle.net/10125/4527.

> In this short chapter, Thieberger provides critical commentary related to making language documentation data as long-lasting, accessible, and useful as possible. Topics include creating data that can be reused and migrated to different formats and media to survive for generations; providing proper methods training in documentation and data management for academic and speech communities; encouraging repositories to conform to accepted data management and curation standards; meeting the evolving needs of users in an increasingly social media-oriented environment; and, of course, creating incentives for parties involved to follow best practices.

Thieberger, Nicholas. 2013. Curation of oral tradition from legacy recordings: An Australian example. *Oral Tradition* 28(2). 253–260.

> This piece is a brief introduction to PARADISEC, aimed at an interdisciplinary audience interested in the world's oral traditions. Thieberger summarizes the mission, history, and operations of PARADISEC. Discussion includes the technical features of the archive, annotations and transcriptions, and trainings offered by PARADISEC. Thieberger also describes how interested researchers can use the archive to access online recordings and their accompanying analyses.

Thieberger, Nicholas & Linda Barwick. 2012. Keeping records of language diversity in Melanesia: The Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC). In Nicholas Evans & Marian Klamer (eds.), *Language Documentation & Conservation Special Publication No. 5, Melanesian Languages on the Edge of Asia: Challenges for the 21st Century*, 239–253. Honolulu: University of Hawaii Press. http://hdl.handle.net/10125/4567.

> Thieberger & Barwick present an overview of the context behind the creation of PARADISEC and a summary of how the archive operates. PARADISEC is a cutting-edge digital repository for recordings primarily from the region around Australia (but open to materials from around the world), and aims to make such materials available to researchers and communities. Founded in 2003, the archive has long been a best-practices leader, being designed specifically to interoperate with researcher workflows, accommodate the domains and standards of different disciplines, and consider ongoing ethical and technological developments.

Thieberger, Nicholas & Andrea L. Berez. 2011. Linguistic data management. In Nicholas Thieberger (ed.), *The Oxford handbook of linguistic fieldwork*, 90–118. Oxford: Oxford University Press.

This article is a guide to managing digital workflows for language documentation both in and out of the fieldwork setting. Good data management in a documentation project is likened to building a house: When the foundation is solid, the house is long-lasting and extensible. The article discusses a wide range of topics of interest to the documentary linguist who is preparing to develop procedures for managing digital data, including the difference between data and metadata; the distinction between form and content (e.g., form-driven markup versus content-driven markup); and a workflow for well-formed linguistic data from field to archive to presentation. The authors offer suggestions for planning for data management well in advance of fieldwork, including planning for archiving and developing procedures for consistent file naming and data backup. Finally, the paper discusses the principles behind a relational metadata database, the value of regular expressions in data manipulation, and creating well-structured time-aligned interlinear glossed texts.

Thieberger, Nicholas, & Simon Musgrave. 2007. Documentary linguistics and ethical issues. In Peter K. Austin (ed.), *Language Documentation and Description, Volume 4*, 26–37. London: SOAS.

This article discusses vital ethical concerns that have arisen in linguistics due to developments in technology and modern language documentation. Thieberger and Musgrave focus primarily on informed consent and data ownership and rights. For example, researchers must grapple with the fact that language documentation is more intrusive than traditional descriptive data collection, and documentary linguists cannot predict all future uses for their data. Moreover, archives have become central to language documentation, which introduces a third party that must be taken into account when constructing consent. The authors also address issues regarding the ownership of language data and the products derived from them.

Thieberger, Nick, 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 PARADISEC*, 1–15. Sydney: Sydney University Press.

The introductory chapter in a volume to commemorate the tenth anniversary of the founding of PARADISEC, this piece describes the founding of the archive in 2002 and reflects on its evolution over the following decade. At the time of writing, the archive houses some 94,500 files on 860 distinct languages worldwide. Technical specifications are described, including the development of Nabu, the archive's catalog software. The authors also provide examples of academic and community uses of PARADISEC collections over the years. PARADISEC now rates five stars on the Open Language Archive Community metric and holds the European Data Seal of Approval.

Thieberger, Nick, Anna Margetts, Stephen Morey, & Simon Musgrave. 2015b. Assessing annotated corpora as research output. *Australian Journal of Linguistics* 36. 1–21.

> This paper represents an important step in the valuation of documentary linguistics corpora as scholarly output. The authors explore options for valuing corpora in the Australian research context, although they note that these discussions can and should take place in other countries as well. Options considered include publishing corpus reviews, which would be similar to book reviews; and a publication or 'journal' model, in which corpora are 'published' in a serial publication. The authors propose a peer review process for corpora that is similar to the peer review process of traditional publications, under the auspices of the Australian Linguistics Society, and they include discussion of parameters for assessing the accessibility and quality of corpora.

Trilsbeek, Paul & Alexander König. 2014. Increasing the future usage of endangered language archives. In David Nathan & Peter K. Austin (eds.), *Language Documentation and Description, Volume 12: Special Issue on Language Documentation and Archiving*, 151–163. London: SOAS.

> Trilsbeek & König approach crucial issues of using existing infrastructures to expand the usage and audiences of digital archives that preserve endangered language materials. This includes discussion of acquiring additional materials by facilitating and increasing contributions from language communities; integrating with existing large-scale e-infrastructures to furnish users with access to more data and research tools; and making endangered language data more available to researchers in disciplines other than linguistics by finding means to enrich metadata and provide useful annotations, transcriptions, and translations.

Trilsbeek, Paul & Peter Wittenburg. 2006. Archiving challenges. In Jost Gippert, Nikolaus P. Himmelmann & Ulrike Mosel (eds.), *Essentials of language documentation* (Trends in Linguistics Studies and Monographs 178), 311–336. Berlin: Mouton de Gruyter.

> This article surveys the challenges of digital archiving by assessing the 'three key players' involved: depositors, users, and archivists. Each places different demands upon the archive, and a given key player has motivations, goals, and preferences that differ from those of the others. Trilsbeek and Wittenburg review these demands and the conflicts they create, and they discuss interactions between an archive's key players. The article also examines conflicts generated by an archive's need to preserve data for the long term while meeting the short-terms needs of various user groups. Finally, it offers a valuable look at legal and ethical issues of access and managing access to archived materials.

Wilbur, Joshua. 2014. Archiving for the community: Engaging local archives in language documentation projects. In David Nathan & Peter K. Austin (eds.), *Language Documentation and Description, Volume 12: Special Issue on Language Documentation and Archiving*, 85–101. London: SOAS.

> Wilbur describes his experiences with the Pite Saami Documentation Project working with local archival institutions to improve access to language materials for speech communities. Modern archiving of language documentation materials is primarily digital, online, and aimed at a global audience. However, Wilbur notes that this can create barriers for many communities interested in accessing information about their own language and culture. Such barriers include a lack of requisite technological infrastructure or computer and language skills. Wilbur presents a case study to illustrate the benefits and challenges of working with national, regional, and municipal institutions to overcome these barriers.

Woodbury, Tony. 2003. Defining documentary linguistics. In Peter Austin (ed.), *Language Documentation and Description Volume 1*, 35–51. London: SOAS.

> In this edited version of a plenary address from the 2003 annual meeting of the Linguistic Society of America, Woodbury provides an overview of the relatively new field of language documentation. The motivations for documentation include changes in technology, an increased interest in linguistic and social diversity, and, of course, the language endangerment crisis. The author notes that one of the defining characteristics of the field as distinct from other areas of inquiry is the discourse-centered approach of documentation, wherein attention to naturally occurring speech takes a place of importance alongside more traditional endeavors like language description. The author also addresses the need for a theorization of language documentation, and he discusses specific projects in Alaska and Peru.

Woodbury, Anthony. 2011. Language documentation. In Peter K. Austin & Julia Sallabank (eds.), *The Cambridge handbook of endangered languages*, 159–211. Cambridge: Cambridge University Press.

> Woodbury's chapter is dedicated to defining language documentation in a handbook on endangered languages more generally. He traces the development of the field as having its roots in the Americanist tradition, especially the ethnographically rich fieldwork of Franz Boas. Boas' practices and values then transferred via his student Sapir to structural era scholars including Emeneau and Haas, then to Krauss, and even to Gumperz in the 'ethnography of speaking.' The author also discusses the relationship between documentation and community-based language work and values, making the point that good documentation can be widely useful in practical and emblematic ways in language revitalization programs.

Woodbury, Anthony C. 2014. Archives and audiences: Toward making endangered language documentations people can read, use, understand, and admire. In David Nathan & Peter K. Austin (eds.), *Language Documentation and Description, Volume 12: Special Issue on Language Documentation and Archiving*, 19–36. London: SOAS.

The author provides advice to language documenters, archivists, and audiences for improving the frequency and purpose of usage of archival collections. Documentary linguists can make their collections more valuable by creating corpus guides, including good descriptions of the documentation project activities, and sharing fieldwork journals. Archivists can increase usage by making collections easily discoverable and accessible; asking depositors to create collection guides (or creating one when the depositor is no longer available); and following practices undertaken by art museums, including guest curators and 'exhibits.' Audiences (e.g., journal editors) can increase the value of collections by encouraging reviews of archival collections.

Yamada, Racquel-María. 2007. Collaborative linguistic fieldwork: Practical application of the empowerment model. *Language Documentation & Conservation* 1(2). 257–282. http://hdl.handle.net/10125/1717.

Yamada presents a case study of linguistic fieldwork designed to meet the needs of both academic and speech communities. Linguists working to document endangered languages can struggle to achieve their own professional and academic goals while balancing the needs and desires of the communities with which they work. Yamada provides examples from her own work with speakers of the Cariban language Kari'nja to illustrate a model of collaborative, community-based linguistic research. She describes several projects, including the creation of pedagogical materials, collaborative linguistic analysis, and the repatriation of previous language recordings.

Ryan E. Henke
rhenke@hawaii.edu

Andrea L. Berez-Kroeker
andrea.berez@hawaii.edu

elpublishing

Language Documentation and Description

ISSN 1740-6234

This article appears in: *Language Documentation and Description, vol 2*. Editor: Peter K. Austin

Language documentation and archiving, or how to build a better corpus

HEIDI JOHNSON

Cite this article: Heidi Johnson (2004). Language documentation and archiving, or how to build a better corpus. In Peter K. Austin (ed.) *Language Documentation and Description, vol 2*. London: SOAS. pp. 140-153

Link to this article: http://www.elpublishing.org/PID/026

This electronic version first published: July 2014

Language documentation and archiving, or how to build a better corpus

Heidi Johnson

1. Introduction

Archives have historically played a central role in the description of endangered languages. This is not surprising, since there is little sense in collecting data on languages that are disappearing if there is no plan for preserving that data. Archiving materials for the already nearly extinct languages of North America was an essential goal of the pioneers of Americanist linguistics: Franz Boas, Edward Sapir, and their intellectual descendants. They diligently deposited all their fieldnotes (and later, audio recordings) in archives and museums such as the Smithsonian Institution.

These archived materials have since formed the basis for decades of linguistic research. Archives facilitate collaboration across generations of researchers and have enabled the production of some of the greatest contributions to the field. One excellent example is the Onondaga-English/English-Onondaga Dictionary (Woodbury 2003) which was based on both the author's own fieldwork and on archived texts and earlier dictionaries.

Archives also support the maintenance and revitalization of endangered languages, by making materials from earlier periods, when the language was more widely spoken and a greater range of forms and genres were still alive, available to the speakers and their descendants. An example of this is the J.P. Harrington Database Project at the University of California, Davis, which is digitizing and publishing on the web the descriptive data he collected in the early 1900's (Macri, Golla, and Woodward 2004). These newly-available recordings and texts are being used in "the monthly 'language lessons' that are being held by members of the Juaneño Band of Mission Indians at San Juan Capistrano. Rather than the usual vocabulary drills and tutoring in a practical orthography, the Juaneños gather to listen to tape recordings of the last fluent speaker of their language, Anastacia Majel, dubbed from aluminium discs that Harrington and his nephew, Arthur, made in the mid-1930s" (Golla 1996).

Similar stories can be told around the world. In the nineteenth and early twentieth centuries, language materials consisted entirely of written text data. Transcriptions taken as direct dictation, notes of elicitation sessions, translations, field notes, and analyses, were all produced on paper. Linguists and anthropologists were careful to preserve these painstakingly produced materials by depositing them in archives, which were well-equipped to preserve such collections. These texts were accessible to researchers who were able to travel to the archive and work with the original, often

Heidi Johnson (2004) Language documentation and archiving, or how to build a better corpus. In Peter K. Austin (ed.) *Language Documentation and Description, Vol 2*, 140-153. London: SOAS.

hand-written, materials. As technological developments progressed, recordings of speech were made, and these were also duly deposited in archives (see De Graaf and Shiraishi, this volume, for a brief history of similar developments in Russia). Unfortunately, recordings on wax cylinders, vinyl disks, and open-reel magnetic tapes are not so easily accessed after a few decades. They are also difficult to copy, placing them at risk of destruction by natural forces such as mould and oxidation.

As recording technologies improved, making recordings in the field became easier and easier, but traditional archives are not well equipped to manage collections of recordings. They have neither the means to preserve them for the long term nor to make them accessible to researchers and speakers. There are exceptions, such as the Indiana University Archives of Traditional Music[1], whose mission is precisely the long-term preservation of recorded materials, on their original media. They make copies on cassette tapes on request.

There are few such repositories for analogue recording media in the world, however, and somehow during the middle decades of the twentieth century linguists stopped trying to deposit their language documentation in archives and museums (except in Australia where the Australian Institute of Aboriginal and Torres Strait Islander Studies has had an active tape archiving policy since 1964). There must be thousands, if not tens of thousands, of recordings of speech in endangered languages on open-reel and cassette tapes squirreled away in the attics and offices of linguists and anthropologists around the world[2]. These primary data have not been publishable, and so perhaps have been less valued by the field as a whole. The only "documentation" that has been available to the world at large, which includes speakers of endangered languages, has been the highly refined distillations of languages that are published as grammars, dictionaries, and scholarly articles.

At the end of the twentieth century, this gloomy picture was transformed by the development of digital media for audio and video recordings, and by the Internet, which facilitates the global dissemination of digital text and media. Now, digital archives make it possible to preserve language documentation permanently and disseminate it widely. The emergence of documentary linguistics (see Himmelmann 1998, Woodbury 2003), accompanied by publicity about endangered languages (such as Webster 2003), and the efforts of international projects such as the Hans Rausing Endangered Languages Project (HRELP) and the DoBeS project of the Volkswagen Foundation. Documentary linguistics is characterised by integration with information and communications technology, which enables researchers to capture, store, and utilize enormous amounts of information (Woodbury 2003, Bird and Simons 2003, Nathan, this volume, Thieberger, this volume).

[1] http://www.indiana.edu/~libarchm/

[2] Dietrich Schüler, Austrian Sound Archive, estimates that 80% of recordings are in private hands (pers. comm.) — Editor.

Fortunately, this explosion of interest and capabilities has been accompanied by developments in the creation of digital archives. There are already several digital archives for endangered language materials ready to receive the documentation being produced today, and to digitize and archive legacy materials from previous decades. The Digital Endangered Languages and Musics Archive Network (DELAMAN[3]) has been formed to co-ordinate efforts and thus improve service to the field. The workshop at which this paper was originally presented was one result of DELAMAN's collaboration.

A list of current DELAMAN members is maintained on the DELAMAN website. Researchers are encouraged to contact any of these archives for information and assistance preferably at an early stage of their language documentation project.

2. Archiving whys and wherefores

This section attempts to answer the basic questions about archiving: who should archive, and where, why, when, and how one should archive.

2.1 Who should archive?

Any researcher who accepts funding from public sources, such as universities and private foundations like HRELP that have public application procedures, has an obligation to produce a public good. Archived materials are public goods, even if access and use is restricted to protect the rights or wishes of the speakers whose words are recorded therein. The resources are still preserved for future generations. Any researcher who works on an endangered language, and thus with an endangered language community, has an obligation to produce materials that can be used by that community well into the foreseeable future and beyond. In other words, all documentary linguists should archive at least a substantial portion of the documentation materials that they produce.

2.2 Where should you archive?

An archive is a trusted repository created and maintained by an institution with a demonstrated commitment to permanence and the long-term preservation of archived resources. A ***collection***, or ***corpus***, is the body of documentary materials created by linguists and native speakers in the course of their research. Note that digitization alone does not constitute archiving. Digital media are actually more vulnerable to loss and obsolescence than are analogue media. The open-reel tapes stacked in museum basements are far easier to retrieve and convert to digital form than a digital recording

[3] http://www.delaman.org

stored on a DVD-RAM double-sided storage disk, the drives for which were made for only one year.

Documentary linguists should seek help from the archive that serves their funding agency, (e.g., HRELP or DoBeS[4]), or the region or language area in which they work (e.g., ANLC[5], AILLA[6], or PARADISEC[7]). Consult the list of DELAMAN archives and feel free to write to any member for advice if you do not find an appropriate archive there.

2.3 Why should you archive?

Documentary linguists should archive their language documentation in order to:

- preserve recordings of threatened languages for future generations;
- facilitate re-use of primary materials (e.g. recordings and fieldnotes) for:
    - language maintenance and revitalization programs;
    - typological, historical, comparative studies;
    - any kind of linguistic, anthropological, or other study that you won't do;
- foster development of both oral and written literatures for endangered languages;
- make known what documentation there is for which languages.

You should also archive your language documentation to further your own career. Archiving can be considered a form of publishing: even if the materials themselves are archived with highly restricted access conditions, the metadata (see section 3.4) is published in the archive's catalogue. You should list all materials that you have archived on your curriculum vitae, so that future employers will know how much work you have done.

Archived materials should also be cited in scholarly and other publications, just as we cite any other published work. This enables those who read a work to locate the primary materials on which that work is based. It also ensures that the speakers whose knowledge and artistry are preserved in the documentation materials are given proper credit for their contributions.

DELAMAN and other organizations such as the Open Language Archives Community (OLAC[8]) are working to devise a standard format for citing archived materials. It will probably look something like the following:

[4] Dokumentation Bedrohter Sprachen, http://www.mpi.nl/DOBES

[5] Alaska Native Language Center, http://www.uaf.edu/anlc/

[6] Archive of the Indigenous Languages of Latin America, http://www.ailla.utexas.org

[7] Pacific and Regional Archive for Digital Sources in Endangered Cultures, http://www.paradisec.org.au

[8] http://www.language-archives.org

Sánchez Morales, Germán. (1994). "Satornino y los soldados." Heidi Johnson, (Researcher.) [online.] Archive of the Indigenous Languages of Latin America. http://www.ailla.utexas.org. ZOH001R010. Access=public.

Note that both the narrator (creator) of the text and the researcher who collected it are mentioned. This models the relationship between the author of a chapter of a book and the editors who put the book together and see it through to publication.

2.4 What should you archive?

Language documentation ideally consists of examples of the full spectrum of language forms and uses that the language community employs (see also Himmelmann, 1998). In general, documentary linguists should try to record, in audio and/or video, as much information as their means and their consultants allow. We should also make an effort to ensure that at least some part of what we record is amenable to open publication, so that some sort of introduction to the language and the culture of its speakers can be made visible to the rest of the world. For many of these peoples, obscurity is a grave historical wrong, pushing them to the margins of world events and facilitating their ultimate destruction.

Of course, we must always discuss the pros and cons of publication for each event or discourse that we record with the speakers, carefully documenting their wishes with respect to future uses of the recordings. In short: it is important to get permission before you start recording (recording includes taking photographs). There is a little more about permissions and intellectual property rights in section 3.2.

With that caveat, what kinds of things are good candidates for archival preservation? Here are some likely genres:

- public events: ceremonies, oratory, dances, chants;
- narratives: historical, traditional, myths, personal, children's stories;
- instructions: how to build a house, how to weave a mat, how to catch a fish;
- literature: oral or written, poetry, any creative work that people may offer;
- conversations: anything that's not gossip or too personal, e.g. conversations about a recent school event or holiday;
- transcriptions, translations, and annotations of recordings, in which anonymity is preserved if necessary;
- field notes, elicitation lists, orthographies - anything other people might find useful;
- datasets, databases, spreadsheets and other secondary (unpublishable) materials;
- sketches of all kinds: grammar, ethnography;

- photographs of speakers and public events.

There are also recordings that should not be archived, such as anything that would cause injury, arrest, or embarrassment to the speakers. One example is the collection of interviews conducted by Pamela Munro and her students with Zapotecs living in Los Angeles about their illegal border crossings (Pamela Munro, pers. comm.). Sacred texts that must not be heard or seen by outsiders are another example.

2.5 When should you archive?

In addition to making regular backups, ideally, you should archive everything you produce as soon as you return from the field, to make sure that nothing is lost. In practice, it usually takes a little time to prepare a field corpus for archiving (see section 3). Note that concerns about losing primary access to the research potential of your documentation should not prevent you from archiving as soon as possible. Students especially are encouraged to archive their corpora with password protection or some other restriction that allows them sole access, to give them time to finish their theses. It is expected that restrictions on access created to protect researchers' concerns will expire after an appropriate length of time (five years, in most cases).

2.6 How should you archive?

You should prepare your corpus according to the guidelines established by the archive where you will deposit it. Review the guidelines published on its website or write to the management for information. If there is no archive for your language or region, the general rules for corpus management given in section 3 will help you ensure that your language documentation is ready for archiving as soon as a suitable institution is established.

3. Building a better corpus

Documentary linguists typically produce a plethora of materials in a wide range of formats, including audio and video recordings, digital and manuscript texts, spreadsheets, and databases. All of these things can be archived.

Steven Bird and Gary Simons identify seven factors affecting the portability of language documentation materials, where portability refers to the continuing usefulness of such materials across time, disciplines, and functions (Bird and Simons 2003). The seven 'pillars' are: content, format, discovery, access, citation, preservation, and rights. Content was discussed briefly in section 2.4. Access and preservation are considered primarily the responsibility of the archive in this guide. Discovery refers to the ability of other interested persons, in our case generally speakers and academic researchers, to locate and access the resource. Metadata, or catalogue information, is what makes

discovery possible[9]. It is also what makes proper citation of the resource possible, so that pillar will not be given its own section here. And since I am an archivist who has had to deal with all manner of legacy materials, I add an eighth fundamental pillar to the foundations for a preservable corpus: labelling.

The following crucial elements in building a better corpus will be discussed here:

- format;
- permissions (rights);
- labelling;
- metadata (discovery and citation).

3.1 Formats

Some data formats are more amenable to long-term preservation than others. These formats may not generally be the most convenient to work with or to use in presentations, publications, or computer-based displays. We distinguish three classes of contexts in language documentation: archival, presentation, and working (the last is where researchers manipulate, edit, annotate their data, etc.) The archival media materials should be uncompressed (this especially applies to digitization of an analogue original) and text data should be in eXtensible Markup Language (XML) structured files. Presentation and working formats can be derived from the archival format.

This is rather abstract, but given the proliferation of digital formats in recent years, it is worth expanding a bit to make clear. The following table of examples of each class of contexts should make these definitions concrete:

	a grammar	a recording	a film
archival context	XML	wav (at least 44.1Khz/16 bits)	MPEG2
presentation context	pdf, html	mp3	Quicktime
working context	MS Word	ATRAC (on minidisk)	proprietary digital camera formats

The general requirements for archival-quality (master copy) formats are that they be:

- non-proprietary; that is, their encoding is in the public domain;
- amenable to forward migration to new formats over time;
- portable, re-useable, repurposeable;
- the best possible reproduction of the original (if not the originals themselves.)

[9] For a different conception of the roles of 'metadata' see Nathan and Austin, this volume.

Legacy materials should be digitized. New materials should be recorded in archival formats. For example: new audio recordings should be created in PCM wav format at a sample rate of at least 44.1Khz with a bit depth of at least 16. DAT recorders, CD recorders, flash ram recorders, and high-density minidisk (Hi-MD) recorders all meet this requirement. Archive the original, and use your copy to produce mp3 files and cassettes for your consultants, make sound snippets for interactive multimedia dictionaries (see Nathan, this volume), and whatever other creative purposes you can devise.

3.2 Permissions

Define a policy concerning Intellectual Property Rights (IPR) and develop a consistent practice for obtaining consent, e.g. forms and/or recorded statements. Learn how to talk to your consultants about IPR. The best source for information on developing your policy will be other researchers who have worked in your region or language community, who are familiar with the customs and mores of the area, and your native-speaker consultants. Note the IPR status of each resource in its metadata (section 3.4).

As mentioned above, you should always get permission before recording anything. Getting permission means discussing the potential uses and abuses to which the recordings and other documentary materials that you and your consultants produce may be subject over time. Generally, we seek permission to publish language documentation for use only for academic, educational, and other non-commercial purposes.

If your consultants are familiar with forms, you could ask them to sign a licence agreement, such as the one on AILLA's website. You could expand this form to include every potential use that you can imagine, such as the following:

- archiving with the following access conditions:
  - open public access for non-commercial purposes;
  - access restricted by password;
  - access restricted for a certain length of time;
  - permission must be granted by a specific agency or individual;
  - special conditions (to be specified) apply.
- other publications, such as books or CD-ROM;
- excerpts published and/or used in classrooms.

If your consultants are unwilling or unable to sign a form, you could record on audio or video a statement of their agreement to specified uses of their works. This recording would then become a part of the archive's documentation for the work.

Although the legal and ethical issues are complicated, particularly when viewed from a global perspective, it is not really that hard to talk to the consultants we work with about potential uses of their work. It is incumbent upon us all to learn how to talk to speakers about intellectual property rights and publication, to take the time in the field for full discussion of all related issues, and to document in permanent form the resulting agreement between speakers and researchers, so that the archive can handle the documentation materials appropriately in the future[10]. That said, worries about property rights should never be used as an excuse for not archiving documentation for present and future generations.

3.3 Labelling

Nothing could possibly be more important than labelling every single item you produce — each track, tape, disc, notebook, digital file, photograph — with **RUTHLESS CONSISTENCY**.

Give this some serious thought. Your system must be infinitely extensible, ensure that related parts can be put back together, and facilitate sorting and general corpus management. You should be using this system from the very start, so figure it out before you begin your project. Think of it as your 'hit-by-a-bus' insurance: if something happens to you, another person will be able to make sense of your corpus, so that the speakers and others who are depending on you to do a good job are not disappointed (though of course, they will be grieved).

The first step is to decide what constitutes an archival object in your corpus. This is not necessarily the same thing as a digital file, and not necessarily the same thing as a unit of media, such as a CD. Consider the difference between a digital video casette, an MPEG2 file, and a documentary film: the file encodes the film which resides on the casette (along with, perhaps, other films). The file will go in the archive and be converted to whatever new format comes along in a decade (or less); the film will be described in the metadata, cited in articles, and 'repurposed' into alternative formats, such as CD-ROMs and BBC special broadcasts. The casette will probably end up in a landfill somewhere. But you still have to label it, so that the archivist, the BBC producer, and you can locate the file to view the film.

The useful 'object' over the long term is the content — the film, in our example above. This should generally be the basic object in your labelling scheme, if possible. In handling legacy materials, archivists often resort to considering the carrier (tape, casette, disk) as the basic object, simply because we can't understand the intellectual content it contains well enough to distinguish one story from another. But for new materials, this should not be a problem. Each individual story (song, interview, etc.)

[10] Note that IPR restrictions can be subsequently changed to be more or less restrictive and are not set in stone. It is important that consultants understand this flexibility.

that you record onto your high-density minidisc constitutes a separate archival object and should be labelled accordingly.

One other factor that must be considered is ensuring that related things are kept together by your labelling scheme. Language documentation materials often come in sets, or bundles, of related items. The prototypical example is an audio recording of a narrative with an annotation text that includes the transcription and translation of the recording. These two things may exist on different media in different physical locations — like a DAT tape in a storage box and an Shoebox file on your hard drive — for the duration of your project (although increasingly documenters use time-aligned annotations that link digital media and text — see Thieberger, this volume, for an example). Your labelling scheme sholud ensure that they can be properly paired by someone else and that they will be archived together. Long recordings may span several carriers, resulting in parts 1 and 2 (or more): these must be labelled so that people can listen to the whole recording in the proper order. Some people make both audio and video recordings of the same discourse event: be sure that your labels allow this relationship to be recovered. You may want to consider some member of the set as primary and the others as secondary. For example, an audio recording is primary, while transcriptions, translations, and other annotations are obviously derivative products, and thus secondary. A dataset that you construct during analysis may be regarded as a single object in itself and receive its own label.

I strongly recommend using a numeric labelling system (that may appear to be an opaque, and user unfriendly) and keeping track of all the details in an auxiliary database, spreadsheet, index cards, or some other sortable form. Numeric labels, which should be unique, are infinitely extensible and compact; this means you will be able to fit them on tiny media labels and use them as keys in your database. Do not use titles of stories: you have no idea how many versions of “El Tigre” you will ultimately end up recording. Always write labels on everything in good indelible black ink using clear, legible print. In the following subsections, I give you three examples of extensible labelling schemes.

3.3.1 AILLA labels

At AILLA, we label every resource, which in our archive refers to a bundle of related files, and every file inside that bundle. The resource label is used to sort the collection and appears in citations of archive resources. If you used something like this, you would make the resource label the key in your supporting database, and write it on the CD, minidisc, notebook, diskette, or any other thing that includes a part of this resource’s bundle of related files.

Our labels work like this:

ZOH001R010	the 10th resource in the first deposit for language ZOH (Zoque of Oaxaca)

ZOH001R010I001.wav	the audio recording in wav format
ZOH001R010I001.txt	the Shoebox interlinearization in text format

We use the language code[11] as the first element so that all the materials in the archive for a given language will sort together. This is extremely helpful if you are working with more than one language. The deposit number helps us manage the archiving workflow. The zeros make sure that all the files in the archive will sort properly; they aren't necessary if you have fewer than 100 or so objects to manage.

3.3.2 Participant initials plus a media type code

A participant is a person who plays an important role in the creation of a resource. The central participants are the speaker who narrates a story, sings a song, or contributes to elicitation sessions, and the researcher who elicits all this verbal behaviour. You could use a labelling scheme based on the initials of your consultants; this would let you sort entries in your database so that all the materials created by or with a given consultant would fall together. If more than one consultant has the same initials, you'll have to add some letters to distinguish them. Examples from my work with Germán Sánchez Morales are:

gsm1_au1	audio recording part 1
gsm1_au2	audio recording part 2
gsm1_sb	Shoebox interlinearization of the audio
gsm1_tx1	text, notes
gsm1_ph1	photo of Germán

The next resource that you create with this consultant will be labelled gsm2_xx etc. Resources that you create by yourself, such as morphological paradigms, will be labelled with your own initials, numbered in the same fashion, and included in your corpus management database.

3.3.3 Label by media unit

This is a very straightforward way to manage recordings made on removable media such as CDs, minidisks or DAT tapes:

md1t1	minidisc 1, track 1

[11] The language codes used by all members of the Open Language Archives Community come from the Ethnologue language codes developed by the Summer Institute of Linguistics. This set encodes several thousand languages and is thus the most complete set of such codes available. For more information or to search for a code, visit the Ethnologue web site at http://www.ethnologue.com/.

md1t1_sb1 Shoebox database for that minidisk track

This method is not likely to be much help for materials produced on one large hard disk or flash memory card. However, as long as your labels are consistent and your materials described fully in your corpus management database, it really doesn't matter which scheme you employ.

3.4 Metadata

One of the reasons that labelling is so important is that it makes it possible to associate all sorts of useful information with each object in your corpus by means of a metadata record. This information is essential for portability, in the fullest sense of the word (Bird and Simons 2003). Metadata catalogue information is especially vital for digital materials, because they are not amenable to direct inspection, as is a book or other printed matter. Metadata facilitates discovery of archived resources, since it provides an assortment of terms for which researchers and speakers can search using interfaces such as the OLAC Search Engine (Hughes, Kamat, and Bird 2004). The metadata record for a resource also provides a place to maintain information about the intellectual property rights inherent in that resource, such as the full names of its creators (and copyright holders) and any special terms and conditions of use.

At an absolute minimum, the metadata for any resource must include:

- creators' full names: this is required for proper citation[12];
- name of the language: be specific! Zoque of San Miguel Chimalapa, Oaxaca, Mexico, not just Zoque;
- date of creation: use the primary (recording) date for all related items if you want, but be sure to note the date of each recording;
- place of creation: again, be specific;
- access restrictions: note any special conditions or restrictions on the use of the resource. Include a password, if necessary;
- genre keyword: this will be dependent on your choice of schema (see below). Keywords, such as narrative, dataset, word_list, make it easier for people to find the resources they are looking for.

There are two metadata schemas (sets of elements) that have been defined for use by the linguistic community. The OLAC schema is based on the metadata elements used by libraries and other disciplines[13]. The IMDI (International Standards for Language Engineering Metadata Initiative[14]) schema was developed by the Max Planck Institute

[12] In some cases, eg. where consultants request anonymity, you may wish to use abbreviations and store the full names in a password protected file.

[13] Dublin Core Metadata Initiative, http://www.dublincore.org

[14] http://www.mpi.nl/IMDI

for Psycholinguistics on behalf of the DoBeS project. It is specifically designed for cataloguing language documentation materials and bundling related items together properly.

You can choose either the IMDI or the OLAC schema for your corpus. If you already know which archive you will be depositing your corpus with, use the one they require. Detailed documentation of each schema can be found on the respective websites.

Label every metadata entry with the same label that you use for the resource. List every related item in the metadata. Add as many notes about the circumstances of the participants and the creation of the resource as you can while they are still fresh in your mind. The provenance, or history, of a documentation resource is often of great interest to future generations of community members. Always be thinking about your consultants' great-grandchildren when you work with an endangered language.

If you are using the IMDI schema, you can use the IMDI Corpus Browser, downloadable from the IMDI website, to manage your corpus. AILLA, which also uses the IMDI schema, will have a Shoebox 5.0 metadata template available from its website by the time this volume is published. We also offer paper forms that you can download in a variety of formats. You can create your own metadata editor easily enough, using any database or spreadsheet program that you happen to have handy.

4. Conclusion

Doug Whalen has written: "we are poised to see a revolution [in the field of linguistics] caused by an unprecedented level of access to the raw materials of our discipline, using tools that have only recently become available ... The vanguard of the revolution will be those who study endangered languages" (Whalen 2003:30). I hope that this brief guide to corpus management will help ensure that these unprecedented quantities of materials documenting endangered languages are indeed accessible for speakers and researchers for generations to come.

5. References

Bird, Steven and Gary Simons (2003). Seven dimensions of portability for language documentation and description, *Language* 79/3: 557-582.

Golla, Victor (1996). *Newsletter of the J.P. Harrington Conference*. Number 10: May 1996. [online] http://www.rock-art.com/jph/nl10.htm. Accessed 2004-06-04.

Himmelmann, Nikolaus P. (1998). Documentary and descriptive linguistics, *Linguistics* 36: 161-195.

Hughes, Baden, Amol Kamat, and Steven Bird (2004). The OLAC Search Engine, presented at *Workshop on Linguistic Databases and Best Practice*; EMELD Language Digitization Project, Detroit, Michigan, July 16-18, 2004. [online.] http://www.emeld.org/workshop/2004/hughes3-demo.html. Accessed 2004-07-09.

Macri, Martha J., Victor Golla, and Lisa Woodward (2004). *J.P. Harrington Database Project.* [online.] http://cougar.ucdavis.edu/nas/NALC/JPH.html. Accessed 2004-07-09.

Webster, Andy (2003). Digital race to save languages, *BBC News World Edition*, Thursday, 20 March, 2003, 09:02 GMT. [online.] http://news.bbc.co.uk/2/hi/technology/2857041.stm. Accessed 2004-07-09.

Whalen, Douglas H. (2003). How the study of endangered languages will revolutionize linguistics, XVII International Congress of Linguists, Prague, Czech Republic, July 24-29, 2003. To appear in *Linguistics Today*, Piet van Sterkenburg (ed.), Amsterdam: John Benjamins.

Woodbury, Hanni (2003). *Onondaga-English/English-Onondaga Dictionary*. Toronto: University of Toronto Press.

Woodbury, Anthony C. (2003). Defining documentary linguistics, in Peter K. Austin (ed.) *Language Documentation and Description, Vol 1*: 35-51. SOAS.

elpublishing

Language Documentation and Description

ISSN 1740-6234

This article appears in: *Language Documentation and Description, vol 12: Special Issue on Language Documentation and Archiving*. Editors: David Nathan & Peter K. Austin

Archives and audiences: Toward making endangered language documentations people can read, use, understand, and admire

ANTHONY C. WOODBURY

Cite this article: Anthony C. Woodbury (2014). Archives and audiences: Toward making endangered language documentations people can read, use, understand, and admire. In David Nathan & Peter K. Austin (eds) *Language Documentation and Description, vol 12: Special Issue on Language Documentation and Archiving*. London: SOAS. pp. 19-36

Link to this article: http://www.elpublishing.org/PID/135

This electronic version first published: July 2014

Archives and audiences: Toward making endangered language documentations people can read, use, understand, and admire

Anthony C. Woodbury

University of Texas at Austin

1. Introduction[1]

Language documentation leads to the accumulation of linguistic records in vast quantity. In just over a decade, archives established by two major funders of endangered language documentation, the DoBeS archive at the Max Planck Institute in Nijmegen, the Netherlands, and the Endangered Language Archive (ELAR) of the Hans Rausing Endangered Language Project at SOAS in London, have archived 10.5 terabytes and 8 terabytes, respectively, of text, sound, video and photographs. And at the University of Texas at Austin, our digital Archive of Indigenous Languages of Latin America (AILLA) has archived 1.9 terabytes, with another 2 terabytes waiting to be processed. Likewise, the analogue archive of the Alaska Native Language Center, in the space of two decades (1960-1980), collected all then-extant documents and recordings in and on Alaska's 20 indigenous languages, amounting to about 5,000 distinct items (Krauss and McGary 1980), which has now grown to 15,000 items according to Holton (2012: 105).[2]

Digital archives of language documentation have, of course, much in common with traditional ones. In archives of both kinds, as Conathan (2011) describes, records are assembled into corpora. These corpora can be collections of records taken from various sources that are related to a given theme; or they can be *archival fonds*, i.e., records emanating from a single project or group or individual. Potential records are appraised, and if selected, they are accessioned, arranged and described by means of metadata, guides and finding aids of various kinds, which make them accessible. This is best

[1] I gratefully acknowledge discussions, training and influence on these issues over the years from many people, especially Peter Austin, Heidi Johnson, Christian Kelleher, Susan Kung, David Nathan and Joel Sherzer.

[2] See www.mpi.nl/resources/data/dobes for DoBeS; www.elar-archive.org for ELAR; www.ailla.utexas.org for AILLA; and www.uaf.edu/anla/ for the Alaska Native Languages Archive.

Anthony C. Woodbury 2013. Archives and audiences: Toward making endangered language documentations people can read, use, understand, and admire. In David Nathan and Peter K. Austin (eds.) *Language Documentation and Description*, *Vol 12,* 19-36. London: SOAS.

done with the widest possible range of user interests in mind, for the present and future.

Still, digital archives are different. Materials that are 'born digital' can be accessioned relatively easily. Digital archives have huge capacities, so that for text and audio (but not yet video), space is barely a problem. Digital archives can be searched quickly, in many more ways than traditional archives can. They are available instantly, any time, anywhere, and do not require a (perhaps long) trip to a particular place for a visit. And they can be reproduced easily. These differences all have impacts on how archives are conceived, and on what we ask of them.

One such impact, perhaps an indirect one, is that documenters, funders, and archivists have increasingly viewed digital language archives as a means for primary dissemination, that is, they see archiving as a kind of enhanced, permanent publication. Documenters now often carry out projects whose primary *goal* is the creation of an accessible language documentary corpus, which they then ask archivists to preserve and (in most cases)[3] make widely available. They may do this whether or not the corpus reaches print or publication in a more traditional form. In effect archives are becoming a means for communicating results to a wide range of audiences. And this in turn may affect how archives work with more traditional corpora whose creators (or assemblers) assign to themselves a less explicitly authorial role.

In this paper I want to make some suggestions for how language documenters can properly pursue this view of their work. I want to explore how documenters might produce documentations that people can read, use, understand and admire: documentations that genuinely address their audiences (Section 3). I also want to explore how archives can accommodate such efforts (Section 4). And I want to explore what audiences themselves can contribute, so that the efforts do not grow in a vacuum (Section 5). First, however, let us attempt an initial characterization of documentary audiences.

[3] Document creators or donors may also archive materials that they wish to preserve but keep under controlled access; the need for effective solutions is made more urgent by the potential accessibility of digital archives.

2. Who are (or could be) the audiences?

I have argued (Woodbury 2011) that at their best, language documenters want their material, however conceived and assembled, to engage diverse audiences:

- community members interested in family, neighbors, community identity, verbal art, education, reclamation, or nostalgia
- scientists interested in philology, ethnohistory, human ecology, language typology, or linguistic theory
- humanists interested in linguistic expression and its products
- general publics with any of these interests, and more.

Language documentation can be so multi-purpose because it has discourse, that is, records of naturally occurring speech of *any* kind, at its core, even if it also includes such linguist-specific products as paradigms, word lists, and recorded elicitation sessions. A particularly helpful stance is that of Holton (2012), recognizing that even in the short term, archives made by linguists 'aren't just for linguists'. This implies that in selecting material and building archives, we should try to imagine the widest range of possible audiences. As Conathan (2011: 238) puts it, '[o]ver time, the importance of records may change and records may be put to unanticipated uses.'

The way we conceive of audiences widens further when we take a broad view of how a corpus (or the documentation project from which it comes) might be *theorized* (see Woodbury 2011: 161); that is, how the corpus might be said to cohere or 'add up'. There are various possibilities:

- a so-called Noah's Archive, a one-time sampling of the uses of a language for a grammar, dictionary, or thumbnail linguistic ethnography
- a more specific collection, such as the addition of something new, such as conversational data, to an existing corpus for a language lacking such conversational materials (cf. the Aleut conversational corpus recorded by Alice Taff – see elar.soas.ac.uk/deposit/taff2006aleut)
- a database of insect names, with pictures and scientific identifications
- a collection of songs, with text and musical transcription, such as is described for Iwaija *Jurtbirrk* songs in Barwick et al. (2007)
- a set of videotaped and transcribed experiments designed to answer a specific set of questions, as described in San Roque et al. 2012.

These and untold other corpus theorizations engage still further audiences.

Finally, the ever-changing form of digital records can create new audiences, as Nathan (2010) and Thieberger (2012) have discussed. For example, the *Jurtbirrk* songs just mentioned also appeared as a CD with the appealing title *Jurtbirrk Love Songs of Northwestern Arnhem Land* (Barwick et al. 2005).

It is still not clear, however, that we are engaging our audiences as fully as we could be. Austin (2011) surveys audience use of several archives mainly to gauge who the audiences are, and concludes that regional archives have especially engaged communities while academic project-linked archives like DoBeS have more academic followings. I cannot help noting that counts of visits or visitors, where available, are often not as high as we might wish.[4]

It may be that our audiences need more help 'getting in'. Even when metadata for individual resources and the collection as a whole are relatively detailed, there can be a feeling that one is lost in a thicket. To me this is unfortunate, because as both a documenter and an archivist, I feel our material is compelling and that it could interest and intrigue many more people.

3. Proposals for language documenters

A simple *book model* can provide an initial analogy for language documenters interested in shaping how their documentary corpus goes forth to potential audiences via digital archives. Books of transcribed, translated (and sometimes analyzed) texts, which are part of the so-called Boasian trilogy, are the most prominent example. More heterodox is a genre of publication we might call a volume of *language materials* (a term which sometimes appears in the title, e.g., McDonald and Wurm 1979). Knut Bergsland's (1959) published compendium of material on Atkan and Attuan Aleut can serve as an outstanding example. It contains an introduction providing context for the work, as well as commentaries, annotations, analyses, translations, exegeses, and footnotes for individual texts from his own field work as well as other sources. Beyond this, Bergsland creates an organization and flow through his material by inching along the Aleutian Chain from east to west via elaborate presentations of proper tribal, geographical, and personal names (55 dense

[4] 200 distinct visits per year for the Alaska Native Language Archive; 1,200 unique visitors in 2010 for the DoBeS archive; 460 registered users and 1,000 annual downloads at ELAR; and 3,937 registered users at AILLA. [*Editors' note*: these figures are from 2010 and thus now somewhat out of date.]

double-column pages supplemented with maps and photos) before turning to texts and translations, again moving from east to west and presenting materials from a two century period ending with his own field work (73 more pages). What might have been a random collection of materials is woven into a coherent form with a strong trajectory.

Some key ingredients of Bergsland-type framing have also been mentioned in recent writing on language documentation. Let me review them.

Himmelmann (2006: 21) usefully distinguishes between an apparatus for the documentary corpus as a whole, and an apparatus for individual sessions. Whole corpus apparatus would include general metadata as well as what he calls 'general access resources' such as an introduction, statements of overall corpus conventions and, optionally, general descriptive analyses such as grammars or ethnographies. Session apparatus would include individual session metadata as well as annotation and other ancillary materials.

Nathan (2010) and Austin (2013) add to this a notion of *meta-documentation*, 'the documentation of your data itself, and the conditions (linguistic, social, physical, technical, historical, biographical) under which it was produced' (Nathan 2010). Likewise, Conathan (2011) emphasizes the importance of creating (what we might call meta-documentary) context within archives themselves by using original order and provenance as key principles for assembling and organizing archives.

A final distinction, which is also implicit in the book analogy, is raised by Nathan and Austin (2004) under the rubric of *thin* and *thick metadata*. The metaphor is borrowed from Clifford Geertz's vision of ethnography, where the former is metadata meant to help find items, while the latter is rich, context laden, and (potentially) shades into annotation (see also Evans and Sasse 2007; Woodbury 2007 for further exploration). Surely we would expect narrative elaboration in a book, not merely a basket of tags.

Let us consider now some specific proposals that draw and expand upon the book model.

3.1. Documenters themselves should make a guide to their documentary corpus

Traditional archives describe their materials with a variety of tools, including 'catalogue records, finding aids, inventories, [and] subject guides' (Conathan

2011: 245).[5] Holton (this volume) refers to this as *mediating* the corpus. My proposal, then, is for documenters to take an active, authorial role in this process, just as they would as authors of a book; that is, to take on the mediation of their own corpus.[6] Typically, documenters are asked by archives to organize resources hierarchically into 'bundles' or 'sessions' pertaining to recorded speech events, elicitation sessions, or research protocols, and at times to group these into larger, superordinate categories. They are also expected to create metadata based on this hierarchical organization. Sometimes, there is a short prose description of the collection. What is missing, and what I think of as having special importance, is a longer prose statement that introduces the collection, gives background, and indicates how it might be used. At a minimum it would introduce and link to items (or sets of items) of content in the collection. It would also contain, or link to, the normal elements of a book introduction, such as:

- a description of conventions
- ethnographic, geographic, and sociolinguistic setting
- thumbnail guides to grammar and lexicon
- a survey of research, of bibliography, and of extant documentation and
- meta-documentation, i.e., information on the circumstances of the archived material's creation (about which there is more below).

Beyond the minimum, it would narrate a path (or multiple paths, fitting the interests of different types of audiences) through the material, much as Bergsland did for his Aleut material. It could also receive a boost by referring

[5] For example, the National Anthropological Archives of the U.S. National Museum of Natural History lists on its website (www.nmnh.si.edu/naa/guides.htm) guides for sets of collections as well as finding aids for individual collections. The latter contain both prose introductions to a collection, including a biography when the collection is the papers of one individual, and hierarchically organized inventories and catalogues of materials, sometimes with descriptions.

[6] In traditional archives, this work is normally undertaken by archivists. Certainly at our AILLA archive, and I suspect in other digital language archives, there are not enough resources for archives to do this work. It is therefore all the more important for documenters to jump into the breach. In addition, documenters will have a much more intimate knowledge of the structure and content of their materials than an archivist can have.

to, and including within the archive, other finished products such as a written ethnography or grammar, and, if appropriate, materials from other sources (again, just as Bergsland did).

3.2. Include meta-documentation: describe the design of activities or projects from which the corpus arose; offer a theorization of the corpus (or several, from different perspectives); and describe the appraisal process in assembling the corpus

For a corpus that arises from a language documentation project it is important to describe the design of the project (the participants and stakeholders and their goals and roles in the project) and to discuss how the corpus may be theorized, i.e., said to cohere (see Woodbury 2011: 161). Here are some suggestions:

- if the corpus arose from a single research project, thoughtfully describe the research, its purpose, design, methods, and expected contributions (and include any before-the-fact research proposals, such as grant applications, or after-the-fact project reports)
- if it arose over the course of many projects, describe the evolution or trajectory of the work
- if it arose in community contexts, or as a joint activity among participants with different or only partially intersecting goals, try to document alternative views of the project's theorization and intentions: for one participant a corpus could be 'intonation data' while for another, that same corpus could be 'stories by old people'.

For any corpus, it is a valuable part of the meta-documentation to describe the appraisal process, i.e., the criteria according to which materials were selected for inclusion. For example, in archiving a recorded narrative along with a transcription and translation, one might describe why one chose to include (or exclude) rough drafts of the translation, or an audio or video recording that documents the process of transcription and translation (cf. Woodbury 2007). Describing the appraisal process is especially important if the corpus includes a collection of materials arising in widely separate contexts (e.g., an assembly of discovered manuscripts or audio recordings), since it also provides a basic theorization for the collection.

These elements of meta-documentation could be part of the guide (Section 3.1), or the guide could link to other discussions along these lines among the 'general access resources'.

3.3. Think of your documentary corpus as belonging to a genre

Books are normally classified by genre, overtly or tacitly. The same can hold of documentations or documentary collections. For linguist-documenters, the long-practiced *texts* genre in the Boasian grammar-dictionary-texts trilogy may seem the default, barely needing discussion beyond indicating that the texts inform grammatical and lexical discovery. Nevertheless it is worth laying out even such basic assumptions, as part of a corpus theorization.

New and different genres continue to emerge as language documentation evolves and as new audiences for it are considered. This is clear from a quick inspection of collections in AILLA, which includes, for example:

- Amith and Castillo García (Ongoing) – a collection of about 100 mostly folktale narratives by several speakers of Mixtec, and as such easily fitting the genre of storybook or folklore collection
- Bohnemeyer (Ongoing) – a set of 15 video-recorded spatial reference experiments performed by Yucatec Maya speakers that is to be part of a much larger set of experiments performed across 13 Mesoamerican languages (Bohnemeyer 2007-), fitting at minimum an experimental corpus genre, but perhaps more generally, comprises a data-accountable version of a scientific monograph
- Sherzer (Ongoing) – a collection that spans decades of audio recordings, texts, books, articles and images encompassing 'wordlists, narratives, poetry, sketches, books, chants, songs, oratory, permissions, photographs, ethnographies, descriptions, articles, conversations, commentaries, greetings/leave-takings, educational materials, meetings, instructions, ceremonies, and elicitation.' As such it amounts to an extended, eclectic ethnography of speaking.

It is particularly important for documenters (or compilers) not only to leverage the 'branding value' of familiar genres when their material fits, but also to create, define, and give substance to new genres that may be special to language documentary collections, always with a view toward communicating with potential audiences (cf. Csató and Nathan 2003:74).

3.4. Write narratives, logs, and journals

Writing narratives about the production of a corpus, during its production or after the fact, puts the material into a clear real-world context, and can aid greatly in its interpretation. This can take the form of an overall narration of the project by one person, whether involved in the project or not, indicating

participants and establishing the sequence and course of the work as well as its goals and its real or perceived setbacks. For example, our Chatino Language Documentation Project generated a series of frank, group-autobiographical annual reports, as well as a narration of our sometimes winding path into complex tonal systems (Cruz and Woodbury, Submitted).

Likewise, narrations can be associated with particular resources. In the Chatino project, we encouraged each other to write 'journals' into our daily metadata summaries which now occur as (somewhat unwieldy, but unquestionably 'thick') entries in an AILLA metadata field titled 'description'. This one by Hilaria Cruz, for example, runs to three paragraphs:

> The information in this resource documents the ceremony of the changing of the local traditional authorities in San Juan Quiahije. This ceremony began at night on December 31, 2009 and culminated at noon on January 1, 2010.
>
> These events take place at the city hall in San Juan which is located in the main square of town. Some of these recording[s] document prayers conducted by the authorities and their families at the church and several other spiritual points in San Juan such as the local cemetery. There are recording[s] of community guards, *ne74 skan4*, and other community members, *nten14 kchin1* at the porch of the city hall in San Juan. There is a conversation inside the city hall with higher ranking authorities, a conversation with higher ranking authorities about the ritual places where they place the candles when they go pray, and the last prayer by two head elders conducted in the main altar of city hall. They are Wenceslao Cruz Cortés and Evencio Cruz Apolonio. Evencio was the head judge and Wenceslao the head elder. This is the last post that Wenceslao will ever serve at city hall. Wenceslao will now transition to be part of the Consejo of Elders in the community. They both served their post for a year. Wenceslao and the other authorities pray at the Catholic church. There are prayers inside the church, a recording outside the city hall, a recording of us preparing to record the ceremony of the changing of the authorities, ambient noise, and recordings outside the city hall.
>
> The actual ceremony of the changing of the authorities, including a speech by Wenceslao, is found along with a conversation between head elders and former authorities, and a recording of the inauguration of new authorities. Finally there is a conversation between elders and lower ranking authorities. The elder Eligio Vásquez asked them to speak. (H. Cruz et al. 2009).

For collaborative projects, such narratives give insight into the assumptions and perspectives of individual contributors, and degrees of interpretive context not attainable by most typical 'short-answer' metadata. Resource (or resource-bundle) specific narrations can serve, in miniature, the same role the 'guide' does for the whole collection.

4. Proposals for Archivists

If the book model suggests how language documenters might communicate with their audiences, perhaps an *art museum model* can offer suggestions to archivists.[7] Art museums are like archives in that they are charged with curating their holdings. Particular audiences may enjoy increased access to certain holdings, but general audiences encounter them in exhibitions, arranged around familiar kinds of themes (e.g. 'paintings by Cassatt', 'New York abstract expressionism', 'Munch's relationship to photography', or even just 'our nicest stuff') and introduced in general terms section by section or room by room, and in particular terms next to each item. Artists' letters or notes, studio photographs, critical discussions and so on may also appear in the exhibition or as part of an analytic exhibition catalogue. Exhibitions may also include materials from other archives, obtaining material 'on loan' in order to contribute to their theme.

With this in mind, here are some further proposals for archivists.

4.1. Make collections accessible and resources discoverable

Most digital language archives are already browsable by depositor, language, and collection (see Trilsbeek and König, this volume, and Nathan, this volume). This is a first step that facilitates proposals made above for documenters. But for many such archives (our AILLA archive included), basic finding aids are not always available. Ideally, alongside basic metadata, each collection will have a guide (as described in Section 3.1) which provides a structured overview.

[7] I owe this analogy to Heidi Johnson, who has long discussed the possibility of creating special collection 'exhibitions' in AILLA; it applies very well to the National Anthropological Archives website with its online exhibitions: www.nmnh.si.edu/naa/exhibits.htm

4.2. Ensure that collections are well described, including meta-documentation that indicates the theorization for the collection

Traditionally, this is simply what archivists do. But in the present context, where resources are spread thin and documenters are invited to describe their own collections, archivists can act as overseers of the work, seeing to it that documenters provide adequate description and helping them where possible.

In cases where documenter input is impossible (e.g., where a documenter is no longer alive), the archive can still compose a general guide and even a theorization of the collection from the archivist's standpoint. Often, it is possible to use other existing materials such as notes, diaries, or publications to reconstruct a basic narrative about the creation of the materials.

4.3. Consider the role of 'guest language archivist'.

Just as museums have guest curators, it is possible to invite deposits from people who are not language documenters in the usual sense, but who have access to records worthy of collecting and archiving and who wish to serve as guest archivists. An analogy from the world of books would be a philological edition of documents, such as Goddard and Bragdon's (1988) compendious 2-volume, *Native Writing in Massachusett*. It contains a preface which includes:

- meta-documentation about their project
- an introduction giving the context of the documents and a discussion of their import from different disciplinary perspectives
- an edition of the documents, including photographs of them, transliterations, translations, and philological discussion,
- a grammar and word index based on the material.

4.4. Consider holding exhibitions

A distinction is rightly drawn between preservation and presentation (Good 2011, Holton 2011). Nevertheless, archives can find ways to preserve archival integrity while creating standalone, outreach exhibitions of fixed duration. Such exhibitions could provide many benefits:

- giving prominence to the work of particular projects, documenters or archivists
- manifesting the archive's commitment to outreach

- providing get-it-in-final-shape-or-else! deadlines to counteract the tendency for archives always to be 'under construction'
- functioning as experiments in identifying and attracting audiences, welcoming them, and addressing their interests.

Furthermore, a suitable framework might enable cross-archival curation on the art museum model, e.g. an archive could hold an exhibition on a particular language or theme that draws not only from its own holdings, but also from other archives, or from ephemeral sources such as websites or social media outlets.

At the same time, this proposal raises challenges for the archivist.[8] These include that of deciding what collections to feature, as well as being aware of cultural sensitivities associated with diverse materials when they are assembled this way in a public context.

4.5. See that collections and exhibitions get reviewed

Finally, like museums and book publishers, archives can take an active interest in getting themselves, their individual collections, or their exhibitions reviewed by interested audiences. This can mean traditional academic review (see below) but it can also include review in popular media, community media, newsletters, and blogs. This would not only provide publicity and outreach, but also feedback on how to address a broader range of audiences.

5. Proposals for audiences

If there is a fruitful model for the contribution that audiences can make, surely it is that of the *critic*. Because our audiences are diverse (see Section 2), we should think of critics of many kinds, not only academic. Nevertheless, it is expedient first to sketch specific proposals for academic and other documentation producers and archivists in their own roles as critical audiences, in the hope of encouraging critics from other audiences as well.

[8] I thank one of the anonymous reviewers of this paper for pointing these out.

5.1. Journal editors can commission reviews of language documentations and may offer criteria or guidelines for the reviews

Journals do not normally solicit reviews of archived language collections, but there is no reason they could not. They might also be encouraged to give guidelines for such reviews. For example, they might ask for:

- a basic description of the collection
- a statement of its scope, purpose, and theorization (how it 'hangs together')
- evaluation of the collection's theorization (how clearly it 'hangs together')
- an evaluation of how clearly the collection is contextualized
- an assessment of its technical attributes (design, systematicity, clarity of data management, and adequacy of transcription, translation, and annotation)
- the relationship and importance of the documentation *vis-à-vis* related documentation and scholarship
- an assessment of likely audiences, and how well their needs are addressed, including those of audiences far in the future
- mention of how the reviewer is situated with respect to the collection and audience, e.g. stating the reviewer's own interests in the material and whether s/he is a native speaker of the language(s) represented.

5.2. Documenters or archivists (or anyone) at-large can volunteer to write reviews, letting review criteria emerge from the task at hand.

Theorizations and intended audiences of documentary collections can be quite different, and that might call for different review approaches. The best review practices may emerge over time, once the review genre(s) become more widespread. Moreover, reviews might be submitted not only to linguistic journals but also to journals in other disciplines, depending on the focus of the corpus.

Whether academic review criteria are stipulated or emergent, they play two very important roles. First, they lead to the establishment of authentic quality standards, evaluative criteria, and expectations, providing conventions for both the documentation discipline and the genre(s) of its reviews. Second, the set of standards, criteria and conventions, and of course the reviews

themselves, become a means for recognizing and rewarding good work. Once such standards are established, it will become easier to explain to university tenure committees that alongside publications, one of the key achievements we look for in seeking tenure for a documentary linguist is a well-reviewed documentary corpus in a highly-regarded archive (this would then operationalize the recognition by the Linguistic Society of America of corpus creation as research, for example).

5.3. Other documentation users can likewise establish criteria and perspectives for evaluating language documentary corpora

Ideally, reviews and other discussion of documentary corpora might emerge in popular contexts such as blogs, community language pages, or social media, and in contexts focused, locally or globally, on education, language activism, the arts, humanities, or the sciences. This too could help to change and improve how language documentations address audiences.

These three proposals for reviewing leave some questions wide open. When is a collection ready for review? Must the archiver or documenter first say it is ready, or is it fair game as soon as it becomes partly visible? Can reviews address sets of materials from various documenters or archives, alongside individual collections? Should academic reviewers have privileged access to restricted materials in order to arrive at an informed assessment?

An important question is whether 'the researcher will be subjected to judgement from someone who will never adequately understand the research experience and may make judgements that could be uninformed or inadequately informed ones ... [t]he review could end up superseding the worth of the collection'.[9] Indeed, any review can be ill-informed and destructive, and archival records might be especially vulnerable precisely because of their uniqueness. Nevertheless, records that are flawed or quirky in some respects may still be of unique scientific or historical importance (see, for example, Hinton's (1994) discussion of both the quirks and lasting value of the archival legacy of the linguist J. P. Harrington). But the possibility of negative reviews might make potential donors of materials have second thoughts, or present a career impediment, rather than a career reward, to young documenters, and thus endanger, rather than promote, the archiving and dissemination of language documentation.

[9] I am grateful to the anonymous reviewer of this paper who raised this question.

But I would still defend the reviewing proposals, on several grounds. Above all, if we really wish to communicate we have to be willing to receive reactions, and if we wish to communicate to wide audiences, those reactions may not always reflect our own perspectives, experience, and expertise. I think that well-intentioned people will take the uniqueness factor into account when evaluating contributions, especially in an academic context. Also, readers of reviews are likely to understand, or take with a grain of salt, the range of reactions they may encounter. Finally, when academic reviews of collections become commonplace, their established norms and standards will better situate any particular review.

6. Conclusion

At present, documentary linguists put enormous effort into collecting and archiving their work in newly created language archives. At the center of this work are records of natural speech, which, because they are at the center of human social and intellectual life, are of very wide interest. Yet despite the labor of documenters and the interest inherent in the material, the work does not appear to be reaching wide audiences. I have made specific proposals for documenters, archives, and audiences to help resolve this problem by developing more direct and explicit protocols of communication between documenters and audiences through the medium of language archives.

I urge documenters to take authorial control of their work, as they would if each archived collection were a book of language materials:

- make a guide to your own documentary corpus
- include meta-documentation: describe the design of activities or projects from which the corpus arose, offer a theorization of the corpus (or several, from different perspectives), and describe the appraisal process used to select and assemble the corpus
- write narratives, logs, and journals
- think of your corpus as belonging to a genre.

To some extent, all this means documenters taking on some of the work traditionally done by archivists. In turn, I urge archivists to assist (and if necessary prod) documenters to meet standards rooted in traditional archival practice, and to act as active promoters of communication with audiences, on the model of a museum:

- make collections accessible and resources discoverable
- ensure that collections are well described, including meta-documentation that indicates the theorization for the collection

- consider a role for a 'guest language archivist'
- consider holding exhibitions
- see that collections and exhibitions get reviewed.

Finally, I urge audiences to be active participants in the process, as critics, on the assumption that that is the only real way to complete the circle of communication:

- journal editors can commission reviews of language documentations and offer guidelines and criteria for the reviews
- documenters or archivists (or anyone) at-large can write reviews, letting review criteria emerge over time
- other documentation users can likewise establish criteria and perspectives for evaluating language documentary corpora.

These proposals create significantly new and challenging roles for documenters, archivists, and audiences. In particular, archivists are asked to pass to documenters (and community members, see Linn, this volume) such key elements of their traditional roles as assessment and archival description, all while retaining ultimate responsibility as overseers. And audiences are implored not to ignore archives but instead to figure out how to use them and to take an active interest.

I think that with these proposals in mind, or at least the goals they represent, it might be possible to begin to address a lopsided situation in documentary linguistics in which we as documenters continue to produce materials, and as archivists continue to preserve them, without making connections to the rest of the world that come anywhere close to our rhetoric of value and loss.

References

Amith, Jonathan and Rey Castillo García. Ongoing. *Mixteco language documentation project.* The Archive of the Indigenous Languages of Latin America: www.ailla.utexas.org. Media: audio, text. Access: 89% restricted. [accessed 2011-10-16]

Austin, Peter K. 2011. Who uses digital language archives? Endangered languages and cultures. Blog archive. www.paradisec.org.au/blog/2011/04/who-uses-digital-language-archives/ [accessed 2011-10-16]

Austin, Peter K. 2013. Language documentation and meta-documentation. In Mari Jones and Sarah Ogilvie (eds.) *Keeping languages alive: Documentation, pedagogy and revitalization*, 3-15. Cambridge: Cambridge University Press.

Barwick, Linda, Bruce Birch and Joy Williams. 2005. *Jurtbirrk Love Songs of Northwestern Arnhem Land,* Batchelor Press, Batchelor NT. Book and CD.

Barwick, Linda, Bruce Birch and Nicholas Evans. 2007. Iwaidja *Jurtbirrk* songs: Bringing language and music together. *Australian Aboriginal Studies*, 2007(2), 1-34.

Bergsland, Knut. 1959. Aleut dialects of Atka and Attu. Transactions of the American Philosophical Society, n.s. 49(3), 3-128. Philadelphia.

Bohnemeyer, Juergen. 2007-. *Spatial language and cognition in Mesoamerica.* www.acsu.buffalo.edu/~jb77/Mesospace.htm [accessed 2011-10-16]

Bohnemeyer, Juergen. Ongoing. *Mesospace collection*. The Archive of the Indigenous Languages of Latin America: www.ailla.utexas.org. Media: video. Access: 100% restricted. [accessed 2011-10-16]

Conathan, Lisa. 2001. Archiving and language documentation. In Peter K. Austin and Julia Sallabank (eds.) *The Cambridge Handbook of Endangered Languages*, 235-254. Cambridge: Cambridge University Press.

Cruz, Emiliana and Anthony C. Woodbury. Submitted. Finding a way into a family of tone languages: The story and methods of the Chatino Language Documentation Project. In Steven Bird and Larry Hyman (eds) *Language Documentation and Conservation, Special Issue: How to study a tone language.*

Cruz, Hilaria (researcher), Wenceslao Cruz Cortés (speaker) and Evencio Cruz Apolonio (speaker). 2009. The Change of Authorities. *Chatino language documentation project collection*. The Archive of the Indigenous Languages of Latin America: www.ailla.utexas.org. Media: audio, video, image. Access: public. Resource: CTP006R016. [accessed 2011-10-16]

Csató Éva and David Nathan 2003. Multimedia and the documentation of endangered languages. In Peter K. Austin (ed.) *Language Documentation and Description,* Vol. 1, 73-84. London: SOAS.

Evans, Nicholas and Hans-Jürgen Sasse. 2007. Searching for meaning in the Library of Babel: Field semantics and problems of digital archiving. In Peter K. Austin (ed.) *Language Documentation and Description,* Vol. 4, 58-99. London: SOAS.

Gippert, Jost, Nikolaus P. Himmelmann and Ulrike Mosel. (eds.) 2006. *Essentials of language documentation.* Berlin, New York: Walter de Gruyter.

Goddard, Ives and Kathleen Bragdon. 1988. *Native writings in Massachusett. Parts 1 and 2.* Philadelphia: The American Philosophical Society.

Good, Jeff. 2011. Data and language documentation. In Peter K. Austin and Julia Sallabank (eds.) *The Cambridge Handbook of Endangered Languages*, 212-234. Cambridge: Cambridge University Press.

Himmelmann, Nikolaus P. 2006. Language documentation: What is it and what is it good for? In Gippert, Jost, Nikolaus P. Himmelmann and Ulrike Mosel (eds.) *Essentials of language documentation*, 1-30. Berlin, New York: Walter de Gruyter.

Hinton, Leanne. 1994. *Flutes of Fire*: *Essay on California Indian Languages*. Berkeley: Heyday Books.

Holton, Gary. 2011. The role of information technology in supporting minority and endangered languages. In Peter K. Austin and Julia Sallabank (eds.) *The Cambridge Handbook of Endangered Languages*, 371-399. Cambridge: Cambridge University Press.

Holton, Gary. 2012. Language archives: They're not just for linguists any more. In Frank Seifart, Geoffrey Haig, Nikolaus P. Himmelmann, Dagmar Jung, Anna Margetts and Paul Trilsbeek (eds.) *Potentials of Language Documentation: Methods, Analyses, and Utilization*, 105-110. *Language Documentation and Conservation Special Publication No 3.*

Krauss, Michael E. and Mary Jane McGary. 1980. Alaska Native Languages: A bibliographical catalogue, Part One: Indian languages. *Alaska Native Language Center Research Papers 3.*

McDonald, Maryalyce and Stephen A. Wurm. 1979. Basic materials in Wankumara (Galali): grammar, sentences, and vocabulary. *Pacific Linguistics Publications*, B65. Canberra: Pacific Linguistics.

Nathan, David and Peter K. Austin. 2004. Reconceiving metadata: Language documentation through thick and thin. In Peter K. Austin (ed.) *Language Documentation and Description*, Vol. 2, 179-187. London: SOAS.

Nathan, David. 2010. Archives 2.0 for endangered languages: from disk space to MySpace. *International Journal of Humanities and Arts Computing,* Volume 4(1-2), 111-124.

San Roque, Lila, Lauren Gawne, Darja Hoenigman, Julia Colleen Miller, Alan Rumsey, Stef Spronck, Alice Carroll and Nicholas Evans. 2012. Getting the story straight: Language fieldwork using a narrative problem-solving task. *Language Documentation and Conservation*, 6, 135-174.

Sherzer, Joel. Ongoing. *Kuna collection.* The Archive of the Indigenous Languages of Latin America: www.ailla.utexas.org. Media: audio, text, image. Access: 0% restricted. [accessed 2011-10-16]

Thieberger, Nicholas. 2012. Using language documentation data in a broader context. In Frank Seifart, Geoffrey Haig, Nikolaus P. Himmelmann, Dagmar Jung, Anna Margetts and Paul Trilsbeek (eds.) *Potentials of Language Documentation: Methods, Analyses, and Utilization*, 129-134. Language *Documentation and Conservation Special Publication No 3.*

Woodbury, Anthony C. 2007. On thick translation in language documentation. In Peter K. Austin (ed.) *Language Documentation and Description*, Vol. 4, 120-135. London: SOAS.

Woodbury, Anthony C. 2011. Language documentation. In Peter K. Austin and Julia Sallabank (eds.) *The Cambridge Handbook of Endangered Languages*, 159-186. Cambridge: Cambridge University Press.

LANGUAGE DOCUMENTATION & CONSERVATION

Vol. 13 (2019), pp. 545–563
http://nflrc.hawaii.edu/ldc
http://hdl.handle.net/10125/24901
Revised Version Received: 5 Aug 2019

Public access to research data in language documentation: Challenges and possible strategies

Mandana Seyfeddinipur
SOAS University of London, UK

Felix Ameka
Leiden University, The Netherlands

Lissant Bolton
British Museum, UK

Jonathan Blumtritt
University of Cologne, Germany

Brian Carpenter
American Philosophical Society, USA

Hilaria Cruz
University of Kentucky, USA

Sebastian Drude
Clarin, The Netherlands

Patience L. Epps
University of Texas Austin, USA

Vera Ferreira
SOAS University of London, UK

Ana Vilacy Galucio
Museu Paraense Emilio Goeldi, Brazil

Brigit Hellwig
University of Cologne, Germany

Oliver Hinte
University of Cologne, Germany

Gary Holton
University of Hawaii, USA

Dagmar Jung
University of Cologne, Germany

Irmgarda Kasinskaite Buddeberg
UNESCO, France

Manfred Krifka
Leibniz Zentrum Allgemeine Sprachwissenschaft, Germany

Susan Kung
University of Texas Austin, USA

Miyuki Monroig
World Intellectual Property Organization, Geneva

Ayu'nwi Ngwabe Neba
University of Buea, Cameroon

Sebastian Nordhoff
Free University Berlin, Germany

Brigitte Pakendorf
Université de Lyon, France

Kilu von Prince
Leibniz Zentrum Allgemeine Sprachwissenschaft, Germany

Felix Rau
University of Cologne, Germany

Keren Rice
University of Toronto, Canada

Michael Riessler
University of Freiburg, Germany

Vera Szoelloesi Brenig
Volkswagen Stiftung, Germany

Nick Thieberger
Paradisec, University of Melbourne, Australia

Paul Trilsbeek
Max Planck Institute for Psycholinguistics, The Netherlands

Hein van der Voort
Museu Paraense Emilio Goeldi, Brazil

Tony Woodbury
University of Texas Austin, USA

E-ISSN 1934-5275

The Open Access Movement promotes free and unfettered access to research publications and, increasingly, to the primary data which underly those publications. As the field of documentary linguistics seeks to record and preserve culturally and linguistically relevant materials, the question of how openly accessible these materials should be becomes increasingly important. This paper aims to guide researchers and other stakeholders in finding an appropriate balance between accessibility and confidentiality of data, addressing community questions and legal, institutional, and intellectual issues that pose challenges to accessible data.

1. Introduction Over the past two decades Open Access to research publications has become increasingly valued by researchers, funding organizations, and the general public.[1] There is an increasing expectation that the products of publicly funded scientific research should be open to all. More recently this expectation is being extended not only to the products of research but also to the primary data from which those results derive. Providing access to primary data facilitates reproducible research, ensuring scientific accountability for research results while also increasing transparency, efficiency, and collaboration (cf. Berez-Kroeker et al. 2018). Another type of challenge arises from statements such as the Berlin Declaration on Open Access,[2] which affects Open Access publications. The Berlin Declaration requires that "[t]he author(s) and right holder(s) of [Open Access] contributions grant(s) to all users a free, irrevocable, worldwide, right of access to, and a license to copy derivative works, in any digital medium for any responsible purpose, subject to proper attribution of authorship". While not legally binding, such declarations can conflict with community interests where limitations on access might be important, or where communities are concerned that their materials might be misappropriated and used for commercial purposes.

The issues raised by the Open Access Movement are impacting all areas of linguistics, but they are particularly significant within documentary linguistics, given the focus of this subfield on primary data. This paper discusses issues surrounding public access to data produced by language documentation projects, i.e., projects which create collections of annotated recordings of people speaking about their lives, cultures, and histories. The tensions arising from the nature of the projects are manifold and relate to privacy and copyright issues, among others (cf. Janke 1998; Brown 2003; Thieberger & Musgrave 2007).

Since the emergence of documentary linguistics as a sub-discipline in the late 1990s, recording and preserving culturally relevant materials, natural dialogues, and oral literature have been important for research, documenting and preserving cultural heritage, and providing community members with access to language data. Accessi-

[1]A chronological overview of the Open Access Movement can be found at https://legacy.earlham.edu/ peters/fos/timeline.htm (Accessed 21 May 2019) – a timeline created by Peter Suber (one of the Open Access pioneers), which covers the period up to 2008. Beyond 2008, this timeline was continued in wiki form at the Open Access Directory and can be consulted at http://oad.simmons.edu/oadwiki/Timeline (Accessed 21 May 2019). A visualised timeline is also available at https://symplectic.co.uk/open-access-timeline/ (Accessed 21 May 2019). For a critical reflection on the definition(s) of Open Access and its implications for indigenous knowledge sharing see Christen (2012), and Singer (2014).

[2]https://openaccess.mpg.de/Berlin-Declaration (Accessed 10 April 2018).

bility is fundamental to the field of documentary linguistics; as summarized by Himmelmann (1998:165), "it is simply a feature of a scientific enterprise to make one's primary data accessible to further scrutiny". However, while Open Access might be seen as an ideal from the open research perspective (OECD 2015), *fully* open data are not always possible or desirable from a cultural, ethical, and privacy perspective (cf. Dwyer 2006; Rice 2006; 2011; Austin 2010; van Driem 2016, among others, for detailed discussions on ethical issues in language documentation).[3] This is because language documentation projects typically produce audio and video recordings which may contain personal or politically sensitive content, or material that is culturally inappropriate to share (cf. Brown 2003:229ff; Christen 2012:2875). This content consists of a variety of genres of natural speech, including traditional stories, histories, cultural activities, procedural accounts, conversational interactions between people, and traditional knowledge, as well as gossip, personal stories, and political discussions. We need to be aware of the colonial nature of academic research, as "imperialism and colonialism brought complete disorder to colonized peoples, disconnecting them from their histories, their landscapes, their languages, their social relations and their own ways of thinking, feeling and interacting with the world" (Smith 1999:28). The role of archives in making material available can be seen as both a continuation of neocolonialist methods, and as a postcolonial repatriation, because restricting access to primary records, which academics are often criticized for, is also seen as bad practice.

Following Christen (2012:2883), "knowledge can (and does) die if it is not used. But it also needs to be used and circulated within an articulated ethical system". Because of the nature of the content of the recordings, access to them may be restricted for several reasons. From a community perspective, recordings may be considered sensitive and not appropriate for Open Access because of their personal or political nature or because knowledge is not seen as shareable with non-community members (cf. Christen 2015). Moreover, researchers might fear that data made publicly available before they fully analyze it may be mined by others who will scoop the original researcher.[4]

Responding to these concerns, many digital archives working with endangered language materials and communities have implemented graded access restrictions.[5] In some instances, depositors are able to specify who should have access to their recordings. In addition, most archives require users to agree to an ethical code of conduct prior to accessing materials, or they may restrict use to educational or academic non-commercial purposes. Strictly speaking, these types of restrictions do not constitute Open Access, as they place an additional barrier between the user and the data and may restrict the way the materials are used and repurposed. For the pur-

[3]It should be mentioned at this point that the Open Access Movement is not trying to make everything open regardless of sensitivities and nuances. Even the strongest supporters of Open Access recognize that open access is not appropriate for every situation.

[4]This fear is reflected in the tendency for PhD students to put embargos on data deposited with language archives. This shows, furthermore, that scooping in itself is more a problem of the academic career and less a problem of the reusage of data.

[5]Examples of such archives are those that are members of the Digital Endangered Languages and Musics Archives Network (DELAMAN). http://www.delaman.org/. (Accessed 10 April 2018).

poses of this paper we will refer to this type of access as Public Access. Some archives may place further restrictions on access to some items, such as requiring users to request access to recordings directly from the depositor. This type of access would not be considered public access.

This paper aims to guide researchers and other stakeholders in finding an appropriate balance between accessibility and confidentiality of data, addressing community questions and legal, institutional, and intellectual issues that pose challenges to accessible data. The paper is organized as follows. We first address issues around communities in §2, then turn to legal issues and ownership of data in §3. Following this, we examine institutions and public access, including a discussion of costing models and archives, in §4. We then turn to data types and the access challenges connected to them in §5, and end with a discussion of credit and control in §6. In all cases, we first set out some of the challenges posed by the goal of public access, and then identify strategies as recommendations that might be used to address those challenges.

2. Communities and public access

This section introduces the types of community issues that may arise from public access to language documentation data and examine some strategies that can be used to address these issues.

2.1 Challenges

Communities and researchers are often concerned about certain types of material being made publicly available. This could be because the material is sacred, spiritual, or even secret in content, is intimately connected to communities' traditional knowledge and genetic resources, or because the material is politically sensitive or identifies individuals in ways that are potentially harmful to them. Communities may be suspicious about how publicly accessible material might be used, and how outsiders might profit from the material. A further challenge arises from the question of how to ensure, in regions with little or no internet access, that the concept of worldwide digital sharing can be explained, with all its consequences.

Community perspectives and concerns about notions of authorship and intellectual property rights, and who has the right to determine to whom material can be made available, may also vary (cf. Whimp & Busse 2000). Determining who has the cultural and legal authority to provide consent may be complex where individuals, families, or communities hold rights to specific stories, songs, dances, or other cultural expressions. Community membership and rights to speak for the community may be contested. Legal and cultural rights and authorities may be affected by clan membership, gender, and individual issues, and there may be groups or institutions within the community who compete for authority. There is also regional variation in attitudes to ownership and control of knowledge and language. In some places language is owned and knowledge must be bought, or used only by the knowledge-holders (cf. Wilkins 1992). Researchers or depositors must be aware of these specificities and provide information about such attitudes in their collection metadata, so that archive staff and users are aware of them.

2.2 Strategies The concerns raised above can be addressed through discussions within the language community, and by working together to implement an ethical framework for ownership, intellectual property and access. Informed consent – ensuring that speakers are aware of the potential harm caused by their participation in a language documentation project – can provide a vehicle for addressing some community concerns. It entails that speakers determine ownership and who will have access to materials resulting from the documentation. Whether informed consent is mandatory because of conditions set by a university, a funder, or a community, discussing issues around consent is essential in understanding intellectual property rights and access. See Fluehr-Lobban 1994, Grinevald 2006, and Robinson 2010, among others, for detailed discussions on informed consent.

Ownership of material and questions around access options need to be discussed early, both with individuals and the wider community, with discussion continuing on a regular basis, and these discussions should be situated within an appropriate ethical framework. Questions such as the following can be considered in the process of understanding and dealing with the specificities of ownership of the data collected: Is this story one that anyone in the community has the right to tell? Does this version belong to a particular person, while in some sense it also belongs to a family?

What level of access the speaker or the community wishes to give to materials resulting from documentation is another topic that needs attention. Here are some important questions to be considered when discussing this issue: Who can listen to, view, or read particular materials, and what does it mean if anyone in the world could do this? Can only a family or a family member listen to, view, or read a story? Could people from a neighboring village listen to, view, or read this material? What about someone from a more distant urban area? How about a government official? Just what these categories are will differ from place to place, making some degree of ethnographic understanding necessary. Additionally, to deal with access issues from within the community, one should also ask beforehand how data made available on the internet might be used.

Workshops can be held to discuss these topics. Notions of authorship, ownership, and accessibility, addressing questions such as those given above, can be discussed. Training can be provided for individual speakers, who can then explain the issues to others. Examples from existing archives can be used to highlight what an archive is, how authorship is indicated, and conditions on access. Likewise, researchers can be educated as to community concerns about access.

More formally, consent should be documented in an appropriate form for the individual and the community: Where written consent is not suitable, speakers' agreement can be recorded orally, as can relevant discussions with the wider community. Community sensitivity to material may vary depending on its format – video, audio, or written – and linguists and archives should be aware that community restrictions might in fact apply only to particular components of a given data set.

Any consent obtained should take into account both authorship and access conditions. Individual and community views of consent can change over time, and these issues should be discussed around any recordings that might be viewed as sensitive,

either with respect to authorship or use. Some material may be deemed inappropriate for archiving and may thus be retained by communities or individuals, or else destroyed. Recordings that were not deemed sensitive at one time might come to be viewed as such at a later point in time and vice versa, so these issues must be revisited regularly in order to ensure that community and individual interests are respected and that appropriate access levels are implemented. Therefore, informed consent should include discussion of the level of access (open, or restricted in some way), and this discussion should be included as part of the collection's metadata.

In some (or perhaps many) cases, truly "informed consent" around access may be unachievable, as the concept of worldwide digital sharing, its scope, and the potential for materials to be misused or misinterpreted is not easily explained. The aim is for informed consent to be as informed as possible. It may be appropriate to err on the side of caution and restrict access, at least in the early stages of research.

Further considerations relate to potential uses of the material that may violate community interests and access agreements. For example, there is a risk that ethnobotanical or artistic material drawn from Open Access deposits could be used in ways that fail to recognize community intellectual property rights, and even for commercial gain – in spite of explicit licenses which prohibit such uses. These risks can be at least partially mitigated by archive-based requirements for registering users, tracking downloads to allow better oversight of the use of the content, and providing clear ethical guidelines on legitimate uses of the material. These risks also need to be weighed against the colonial legacy of withholding materials from the people who have a direct interest in them.

3. Legal issues, ownership, and public access Just as communities can challenge Open Access to materials, legal and ownership issues also present challenges. This section introduces some of these challenges.

3.1 Challenges In some jurisdictions research permits are required in order to conduct a language documentation project, and the permits may place explicit restrictions on access to research data. Where permit processes require researchers to guarantee that research outcomes will not be used for non-research related purposes, particularly commercial gain, violations (actual or perceived) may lead to the loss of a permit and to further implications for a researcher's career. Many universities also require that an ethics protocol be approved before research can begin. The research cannot take place without the permission of the appropriate people or institutions (cf. Bowern 2010; O'Meara & Good 2010; Næss & Hovdhaugen 2011; Good 2018).

Legislation regarding research data varies according to jurisdiction. In some countries there is a requirement that research data (particularly data seen as including personal information) be destroyed once the research is complete. If language documentation data is not exempt from this, a justification can be made for its preservation in an archive, which must happen before a researcher collects the data and requires informed consent to do so, as mentioned in §2. In some publicly-funded archives all archived material is required by freedom of information laws to be made openly

available upon request, as is the case, for instance, with recorded information held by public authorities in England, Wales and Northern Ireland, and by UK-wide public authorities based in Scotland.

Different ethical standards and regulations governing access and copyright may have repercussions for collaboration and working across international boundaries. Researchers must observe the local legal frameworks that apply in all countries where they work, conforming to data protection and privacy laws, obeying national copyright regulations and intellectual property rules, and respecting freedom of information laws. Intellectual property rights may apply differently to original recordings and written texts, as opposed to transcriptions, translations, and other annotations.

3.2 Strategies Researchers should be aware of legal issues and requirements in their institutions, resident countries, the countries and communities in which they conduct research, and the countries in which work will be archived. It is also important to keep in mind that where research permits are required these might include restrictions on data use and access. Researchers should also be informed about these requirements well in advance of beginning the research.

Moreover, researchers must also understand the intellectual property implications of documenting traditional knowledge. Traditional knowledge refers to the "knowledge, know-how, skills and practices that are developed, sustained and passed on from generation to generation within a community" (cf. WIPO 2016a). Due to its low level of legal recognition in many countries, traditional knowledge is not easily protected by the current intellectual property system, which "typically grants protection for a limited period to new inventions and original works by individuals or companies" (cf. WIPO 2016a). Intellectual Property law typically vests copyright in language documentation materials with the individuals who made the recordings – i.e., linguists, anthropologists, etc. – rather than the speakers. This means traditional knowledge holders do not have legal ownership over the materials and cannot determine their legal use (see Macmillan 2013 for a discussion about legal protection of tangible and intangible cultural heritage; see also Khan 2018). In this sense, prior informed consent is essential to clearly assign copyright to speakers, negotiate appropriate licensing, and ensure that communities and individuals can exercise rights over the material provided and that these are acknowledged accordingly.

One strategy for avoiding the strongest implications of the Berlin Declaration for Open Access publications and similar documents is to use the Creative Commons Non-Commercial license.[6] This prevents material from being used in textbooks available for sale, in language schools which charge fees, and on websites which run advertisements to finance costs. However, this strategy may also limit reuse where language initiatives rely on such income streams to finance their operations. Perhaps a more effective strategy for avoiding the commercial use of materials is the Creative Commons Share-Alike license, under which all derivative content must again be made

[6](Accessed 10 April 2018).

freely available.[7] This means, for instance, that a movie made using content from the collection must be available under the same open license.

Different archives have different license or "deed of gift" standards. Some require that copyright be assigned or licensed to the archive, while others stipulate that data creators or authors retain copyright. Other archives require the depositor to apply a Creative Commons license to their research publications.

4. Institutions This section examines institutions broadly, including archives. Issues relating to access, data types, and users of archives are addressed below in §5.

4.1 Challenges Public access to research data requires long-term archiving of language data. This in turn requires a long-term commitment by institutions to maintaining and developing technology to sustain archives and avoid data graveyards. This involves costs, and institutions require models to meet those costs over a sustained time period.

Currently, systematic standardized policies concerning data management and accessibility for funders, researchers, and archives are lacking. Such policies would entail creating interfaces and developing the usability of archives, while meeting high standards for deposits, with reports on usage and impact. There is little training available yet in this kind of data management (see Gawne et al. 2017).

Archiving and maintaining archives comes at a cost, and there is a cost to providing high quality presentations and interfaces, but there is also a cost to not doing so (see Thieberger 2014). Digital archives must be maintained and offer new functions, services, and modes of display that make the data as accessible as possible.

4.2 Strategies One strategy for resolving this challenge is funding. If funding were available to support the work institutions need to do, the skills and talent could be found to do it. Institutions involved in archiving (including museums, galleries, archives, libraries, and research centers) need to collaborate to identify common solutions, both in technology and costing, to ensure continuing support. Systematic, standardized policies concerning data management will be of value to funders, researchers, and archives. The following suggestions should be incorporated into the workflows of the institutions dealing with archiving:

- Restricted access must be justified. See §5.2 and §6.
- Data management, curation, archiving, and publishing should be properly budgeted for beyond a project's lifespan.
- Embargo periods for primary researchers should have time limits and should expire unless a longer time period is explicitly sought. See §6.2.2.
- Implementation of policies should be monitored and researchers' compliance verified through annual performance reports of both researchers and archives.

[7](Accessed 10 April 2018).

Data management does not happen automatically; researchers must be trained in data management techniques. This can be addressed by introducing training through university-level courses in data management and archiving. Field methods courses might include an introduction to workflow management, metadata, access levels, ethical considerations, licensing, and informed consent. Archives could also develop online resources, including video tutorials, in order to ensure thorough coverage of the ethical and practical issues involved. Training for archivists should cover legal and ethical issues. Textbooks and other materials should be developed to allow this, with funding allocated for their creation (see §5).

With respect to the fundamental issue of funding for archiving and making research data accessible, collaboration between archives on a technical level and the sharing of solutions between institutions can minimize costs. Archives need to assess the true costs of curation and archiving, taking into account ingestion, curation, loading, storage, managing access regulations, agreeing on access with speaker communities, and so on, and must seek appropriate sources of funding. Like individual researchers, institutions must be aware of legal requirements regarding making materials available. Researchers need to understand the costs of curation and archiving, and must work with funders to find ways of continuing to fund these beyond the timespan of a grant.

5. Archives Archives as institutions are discussed above in §4. This section examines archives with respect to access, focusing on data types and users. Archives play a critical role in public access to research material, as it is through archives that materials are made discoverable and accessible. While depositors may be better prepared to curate their materials, in practice this task ultimately falls to the archive, which has responsibility for the curation and long-term storage of materials.

5.1 Data types, access conditions, and public access As discussed in §2, providing access to certain types of data may be problematic. A variety of data types are listed in Table 1, together with issues that they may face and possible strategies for dealing with the challenges.

As indicated in Table 1, most material can be made Open Access or accessible through log-in, while access conditions may be appropriate for sensitive material, according to the direction of the speaker or their community. In some cases anonymization may provide a solution, with the researcher undertaking the anonymization with the assistance of archival staff. Metadata can indicate that participants should not be identified: they can be referenced as "anonymous", or people and locations can be given pseudonyms.

5.2 Archives and their depositors This section addresses technical aspects of depositing in archives; see §6 on more personal aspects.

Table 1. Issues and solutions for different data types

Data type	Issues	Strategies
Descriptive metadata	Unproblematic in most cases.	Participants in recording sessions and their personal details as well as locations can be anonymized if necessary. Metadata sets can be hidden while collections are in construction.
Child language data	Minors are protected by national and international laws. It may be necessary to restrict access to voices and images. Consent given by legal guardians may require renegotiation once children come of age; provision must be made for obtaining children's consent later on.	Metadata and anonymized transcripts may be made available. Materials can be archived with restricted access for research use. Video and audio can be stored offline (mandatory in some countries for data pertaining to children).
Original texts, transcripts, and annotations	Less personally identifying than audio/video/images. Intellectual property rights must be respected. Some content may be problematic (see §2.2 on avoidance of harm).	Texts, transcriptions, translations, and some tabular data may be made available where other media are restricted. Can be anonymized. Certain content may need to be restricted. Redacted texts could be made publicly accessible. A limited embargo period may be permitted for students or for first use by researchers.
Multimedia (audio, video)	Contains personally identifying information. Various potential consequences for speakers and communities.	May need to be restricted. Can be made available, if personal rights are cleared and intellectual property rights respected.
Experimental data	Generally unproblematic. Already anonymized.	Existing guidelines from APA, university ethics committees, etc. must be respected.
Location data	Geographical coordinates of certain objects, events or natural resources may be commercially interesting (loggers, poachers, mineral prospectors, bio-pirates, etc.) and may put the community and their area at risk.	Restrict any information that is likely to be problematic. Provide mediated access, if there is a possibility of inappropriate use of the information. Consider withholding from archival collection, if accidental release of data would prove irreversibly problematic.
Sensitive material	Potential monetary value (e.g., ethnobotanical material)	Can be made available to registered users, with clear guidelines for usage and a clear trail of use.
Legacy materials	Not easy to determine access restrictions as there is often no indication of informed consent or sensitivity.	The default is for such data to be publicly available, unless there are legal restrictions or concerns around sensitivity. It should be acknowledged that the material has unclear copyright conditions and can be taken down, if anyone is aggrieved by it (the 'takedown principle', cf. e.g., Urban, Karaganis, & Schofield 2017). Crowdsourcing may be used to enrich metadata and identify possible access issues.

5.2.1 The challenges Archives rely on depositors as intermediaries between themselves and communities, for obtaining informed consent and providing metadata, licenses, and access restrictions. This reliance on the depositor can create problems regarding the handling of personal rights, traditional knowledge, and copyright and licensing rights, especially with older collections where a depositor is no longer available or has not nominated a legal successor to make decisions for the collection, does not have a long-term relationship with the speakers, or where informed consent has not been obtained.

5.2.2 Possible strategies Clear statements of rights and licenses and unambiguous access conditions are crucial for archives to be able to implement the intentions of individuals and communities. From the outset of a project, researchers should work with archives to address issues of licensing and access, to develop a succession plan stating who will be responsible for materials in the future, and to make plans for future treatment of restricted materials. While restricted materials are generally not favored by archivists, community wishes regarding access restrictions must be respected. At the same time, it is too easy for researchers or archives to use 'community sensitivity' as an excuse for not making their records available, resulting in the age-old colonial extraction of materials that do not then find their way back to the source community. In a reflective review of the relationship between Indigenous Knowledge and Open Access, Christen (2012:2889) concludes:

> Incorporating a wider range of ethical and cultural concerns into our digital tools subverts the narrow notions of information freedom and the cultural commons that presently characterize our discussion of the commons. Memes like 'information wants to be free' and general calls for 'open access' undo the social bearings of information circulation and deny human agency. Shifting the focus away from information as bits and bytes or commodified content, indigenous cultural protocols and structures for information circulation remind us that information neither wants to be free nor wants to be open; human beings must decide how we want to imagine the world of knowledge-sharing and information management in ways that are at once ethical and cognizant of the deep histories of engagement and exclusion that animate this terrain.

Archivists can provide guidance and training on obtaining informed consent for archiving as part of creating a data management plan. Clear rules should set out what is expected in terms of access regulations; these might include embargo periods for which any access restrictions must be properly justified. Otherwise, materials for which no justification for an embargo is provided should be made accessible. Funders and archives might share information about depositors' track records on access and archiving.

5.3 Archives and their users

5.3.1 The challenge Language archives must be designed to meet the needs of a variety of users with different expectations and requirements, and these expectations and requirements may change with time (cf. Wasson et al. 2016). Users may include the following:

- Scientific researchers, both in linguistics and other fields, e.g., ethnography, history, cognitive science. They require: good access to data, including detailed search options; streaming and download options; easy ways to reference specific data; ability to upload new annotations without compromising existing ones.
- Speakers of the language and community members. They require: an interface in an appropriate local and/or national languages; metadata and transcriptions in a national language; search capabilities for individuals, places, types of recordings, etc.; an interface suitable for use in schools and other community contexts. Parts of the collection may be accessible only to the community or only to individuals in the community.
- General public, museums etc. Materials and resources that are particularly accessible and interesting, often for extraneous reasons, can be highlighted as "showpiece of the month", etc.; interfaces and transcriptions can be in global languages other than English; holdings described in the language of the general public; links to and from Wikipedia articles and other collaborative platforms.

5.3.2 Strategies to address needs of different user types Different users may have different access rights. For instance, access might be by log-in via a client certificate-based authentication and/or Shibboleth for scientific researchers, and there might be parts of the collection that are restricted in use and available only to community members, or perhaps only to selected community members. Other parts of the collection might be open to all.

Public access includes access to materials by the speakers and their community. Community access deserves somewhat more attention than it currently receives, and can be affected by a variety of factors in different regions. Archived records may not be findable by speakers for a variety of reasons, including:

(a) language barriers;

(b) lack of bandwidth/internet access;

(c) speakers/community not being aware that recordings exist or are available;

(d) inaccurate metadata;

(e) lack of technical skills and computer literacy; and

(f) an interface and data structure that is difficult to use.

Such issues can be addressed by publicizing archive metadata through local cultural agencies and other institutions (e.g., schools, museums, local government), and working to improve access to archive sites. The interface, minimally the metadata catalogue, can be provided in a local language and appropriate training offered. If people do not have access to the internet or computers, tablets or notebooks can be set up in a school or other institution as a local archive. Funders could cover reasonable costs for capacity building and providing local access, with these being implemented by the researcher, the archive, or both, depending on the situation. This must include ongoing training, and to be effective, researchers should work with communities to understand and implement their perspectives on what is needed. Periodic reviews of ownership and access conditions by all relevant parties will likely be helpful. It is important to keep in mind that there is no one-size-fits-all solution – there is both regional variation and variation over time (including changes in technology and in community access to and ability to use technology).

The work of documentation has the potential to be expropriative – collecting and disseminating recordings of indigenous people speaking in their languages is problematic. As Smith (1999:99) notes, "[i]ndigenous knowledges, cultures and languages, and the remnants of indigenous territories, remain as sites of struggle".

However, archive work is typically driven by non-indigenous university-based researchers who have taken on the responsibility of making the research of the university available outside academia. This action counteracts an earlier expropriation, that of the academic researcher who kept recordings safe but did not know how to return them to the source communities, or did return them periodically on analog media that had a short life span.

5.4 Embedding in institutions Some archives are embedded in larger institutions (as opposed to community-based archives, for example) and must follow internal policies, including internet security protocols, choice of specific models and systems for archiving. While institutional policies may conflict with various archival practices, we suggest a commitment to provide public access should form a general archiving principle. Note that prior agreements with depositors may be legally binding; for instance, access levels and other similar requirements need to be preserved.

6. Credit, control, and public access Concerns within communities about making data public were addressed in §2 and §5.1. This section addresses concerns by researchers about making data public.

6.1 Credit, control, and the researcher This section addresses two concerns of researchers: (1) identifying who should be attributed credit may be difficult, or contested; and (2) researchers or research teams may be ambivalent about making a collection available as they are concerned that their contribution to gathering, transcribing, glossing, and translating the material will go unrecognized. We focus on credit

with regard to researchers and communities. Funders are generally acknowledged in a footnote rather than through authorship (we recommend footnote acknowledgement of funding for all archived collections as well as for publications).

Documentation teams should discuss who will be credited in references to the data collection, and how. Major language consultants (transcribers, translators) might be included in references to the whole collection, while individual speakers who contribute narratives, songs, etc. might be credited only in the metadata for individual sessions. The entire team needs to understand the different contributions and what they involve in order to make such decisions – this might come about through workshops revolving around issues of consent. We recommend that the relative contributions of individual contributors are explicitly described in data collections.

In publications arising from language collections, each individual's contribution must be considered when determining co-authorship versus acknowledgement. The relative contributions of individual contributors should be explicitly described in the publication.

Research teams should do what they can to make credit by citation easy. Creators of collections should provide explicit and easy-to-find citation guidelines with the collection (with archives providing guidelines for citing whole deposits, as well as data and metadata at more granular levels; see for example the citation guidelines provided by AILLA at https://ailla.utexas.org/site/rights/citation). Users should cite examples by giving proper references, and researchers who make substantial use of particular collections for a publication should consider including the compilers as co-authors. Compilers of data collections can present the structure of their archival deposit in a journal publication (e.g., Salffner 2015; Caballero 2017; Oez 2018) as a citable reference to the collection. Archival resources can also be cross-referenced in collections such as Glottolog.[8]

6.2 Credit, control, and access restrictions Access restrictions were mentioned in §5.1, and we return to them now, first looking at access restrictions and the community, and then at access restrictions and the researcher. We continue to draw a line between community and researcher, although in reality such lines can blur.

6.2.1 Credit, control, access restrictions and the community Language documentation typically works with languages spoken by a small number of speakers. Due to the small size of the cohort, recordings can contain materials which might put these communities at risk of harm, from outside or from within. A text might cause harm by asserting the rights of a particular group to a contested piece of land or a favorable version of history. Other recordings contain highly personal information, and in small societies it may be impossible to anonymize speakers.

While funders may require public access, community members may require restrictions before information is provided. Sensitive materials archived with restrictions can at least be preserved. Some researchers find setting immediate restrictions may

[8]http://www.glottolog.org. (Accessed 21 May 2019).

lead to Public Access over time, as people decide that they want materials to be accessible.

Where not at odds with the community's views, we recommend using restricted access only with a clearly specified embargo period, after which the restrictions can be lifted. That date could possibly be in the far future, but it must not be undefined. For any materials requiring long-term restrictions, legal successors to depositors should be identified wherever feasible (this implies an ongoing relationship at least between archives and researchers).

6.2.2 Credit, control, access restrictions, and researchers Researchers may avoid making their data collections publicly available out of fear that others might use the data without proper attribution. Creators of research data have a recognized right to reasonable first use of data. It is therefore possible to restrict access to data collections/corpora for a defined period to enable primary compilers to work with their data before others do (cf. Berez-Kroeker & Henke 2018:362–364). However, embargo periods should not be perpetuated without limits. Archives should require justifications for extensions beyond a standard embargo period (see §4.2). The risk in not allowing material to be embargoed is that not all records will be archived and they will then potentially be lost. Once data is released, citation standards for data sources must be applied and checked/enforced by peers and peer review processes when it is observed that data is being reused (see §6.1).

7. Summary This paper discusses some of the challenges arising from the ideal of Open Access to collections that result from language documentation projects. These include challenges involving communities, legal matters, archiving, costs, data types, access types, and credit. This paper suggests some possible solutions, noting the importance of being aware that communities, data contexts, and technology all evolve over time. In all areas, we emphasize the need for learning what external forces there are that must be complied with, and for focusing on education, on working together, and on flexibility at all levels.

Acknowledgements This paper is the result of a 3-day workshop on "Open Access and Open Data of Endangered Languages Collections" held October 10–12, 2016, at the University of Cologne funded by the Volkswagenstiftung. It emerged as collaborative writing by 35 stakeholders and researchers working on different aspects of language documentation and Open Access. Thanks to two anonymous reviewers for their constructive comments.

References

Anderson, Jane & Molly Torsen. 2012. *Intellectual property and the safeguarding of traditional cultures: Legal issues and practical options for museums, libraries and archives*. Geneva, Switzerland: WIPO. http://www.wipo.int/edocs/pubdocs/en/tk/1023/wipo_pub_1023.pdf.

ATHENA. 2009. *ATHENA deliverables and documents: WP6 – Analysis of IPR (Intellectual Property Rights) issues and definition of possible solutions*. http://www.athenaeurope.eu/index.php?en/149/athena-deliverables-and-documents.

Austin, Peter K. 2010. Communities, ethics and rights in language documentation. In Peter K. Austin (ed.), *Language documentation and description*, vol. 7, 34–54. London: The Hans Rausing Endangered Languages Project.

Berez-Kroeker, Andrea L., Lauren Gawne, Susan Kung, Barbara F. Kelly, Tyler Heston, Gary Holton, Peter Pulsifer, David Beaver, Shobhana Chelliah, Stanley Dubinsky, Richard Meier, Nicholas Thieberger, Keren Rice, & Anthony Woodbury. 2018. Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics* 57(1). 1–18. doi:10.1515/ling-2017-0032.

Berez-Kroeker, Andrea L. & Ryan Henke. 2018. Language archiving. In Rehg, Kenneth & Lyle Campbell (eds.), *Oxford handbook of endangered languages*, 347–369. Oxford: Oxford University Press.

Bhattachary, Darren & Douglas Dalziel. 2012. Open data dialogue: Final report. *Research Councils UK*. https://www.ukri.org/files/legacy/documents/tnsbmrbrcukopendatareport-pdf/.

Bowern, Claire. 2010. Fieldwork and the IRB: A snapshot. *Language* 86(4). 897–905.

Brown, Michael F. 2003. *Who owns native culture?* Cambridge, MA: Harvard University Press.

Caballero, Gabriela. 2017. Choguita Rarámuri (Tarahumara) language description and documentation: A guide to the deposited collection and associated materials. *Language Documentation & Conservation* 11. 224–255. http://hdl.handle.net/10125/24734.

Choukri, Khalid, Stelios Piperidis, Prodromos Tsiavos, Tasos Patrikakos, Maria Gavrilidou, & John Hendrik Weitzmann. 2012. *META-SHARE: Licenses, legal, IPR and licensing issues*. Berlin, Germany: META-NET. http://www.elra.info/media/filer_public/2015/03/30/meta-net-d613.pdf.

Christen, Kimberly. 2012. Does information really want to be free? Indigenous knowledge systems and the questions of openness. *International Journal of Communication* 6. 2870–2893. https://ijoc.org/index.php/ijoc/article/view/1618.

Christen, Kimberly. 2015. Tribal archives, traditional knowledge, and local contexts: Why the "s" matters. *Journal of Western Archives* 6(1). Article 3. http://digitalcommons.usu.edu/westernarchives/vol6/iss1/3.

CLARIN (Common Language Resources and Technology Infrastructure). Licenses and CLARIN categories. https://www.clarin.eu/content/license-categories. (Accessed 10 April 2018).

Dwyer, Arienne M. 2006. Ethics and practicalities of cooperative fieldwork and analysis. In Gippert, Jost, Nikolaus. P. Himmelmann, & Ulrike Mosel (eds.), *Essentials of language documentation*, 31–66. Berlin: Walter de Gruyter.

Fluehr-Lobban, Carolyn. 1994. Informed consent in anthropological research: We are not exempt. *Human Organization* 53(1). 1–10. doi:10.17730/humo.53.1.178j-ngk9n57vq685.

Gawne, Lauren, Barbara F. Kelly, Andrea L. Berez-Kroeker, & Tyler Heston. 2017. Putting practice into words: Fieldwork methodology in grammatical descriptions. *Language Documentation & Conservation* 11. 157–89. http://hdl.handle.net/10125/24731.

Good, Jeff. 2018. Ethics in language documentation and revitalisation. In Rehg, Kenneth & Lyle Campbell (eds.), *Oxford handbook of endangered languages*, 419–440. Oxford: Oxford University Press.

Grinevald, Colette. 2006. Worrying about ethics and wondering about "informed consent": Fieldwork from an Americanist perspective. In Saxena, Anju & Lars Borin (eds.), *Lesser known languages in South Asia: Status and policies, case studies and applications of information technology* [TiLSM 175], 339–370. Berlin: Mouton de Gruyter.

Himmelmann, Nikolaus P. 1998. Documentary and descriptive linguistics. *Linguistics* 36(1). 161–196.

Janke, Terri. 1998. *Our culture, our future: Executive summary of report on Australian Indigenous Cultural and Intellectural Heritage Rights*. Canberra: AIATSIS (Australian Institute of Aboriginal and Torres Strait Islander Studies) and ATSIC (The Aboriginal and Torres Strait Islander Commission). https://www.wipo.int/export/sites/www/tk/en/databases/creative_heritage/docs/terry_janke_culture_future.pdf.

Khan, Mehtab. 2018. *Traditional knowledge and the commons: The open movement, listening, and learning*. Creative Commons [blog]. https://creativecommons.org/2018/09/18/traditional-knowledge-and-the-commons-the-open-movement-listening-and-learning/.

Klimpel, Paul. 2013. *Free knowledge based on Creative Commons Licenses: Consequences, risks and side-effects of the license module "non-commercial use only – NC"*. Berlin: Wikimedia Germany. https://openglam.org/files/2013/01/iRights_CC-NC_Guide_English.pdf.

Macmillan, Fiona. 2013. The protection of cultural heritage: Common heritage of humankind, national cultural "patrimony" or private property? *Northern Ireland Legal Quarterly* 64(3). 351–364. http://eprints.bbk.ac.uk/7289/1/7289.pdf.

MINERVA EC Working Group "Quality, Accessibility and Usability" (ed.). 2008. *Handbook on cultural web user interaction*. 1st edn. http://www.minervaeurope.org/publications/Handbookwebuserinteraction.pdf.

Næss, Åshild & Even Hovdhaugen. 2011. Language is power: The impact of fieldwork in community politics. In Haig, Geoffrey, Nicole Nau, Stefan Schnell, & Claudia Wegner (eds.), *Documenting endangered languages: Achievements and perspectives*, 291–304. Berlin: Mouton de Gruyter.

Newman, Paul. 2011. Copyright and other legal concerns. In Thieberger, Nicholas (ed.), *The Oxford handbook of linguistic fieldwork*, 430–456. Oxford, New York: Oxford University Press.

Nowviskie, Bethany. 2014. *Why, oh why, CC-BY?* Bethany Nowviskie [blog]. http://nowviskie.org/2011/why-oh-why-cc-by/.

OECD (Organisation for Economic Cooperation and Development). 2015. Making open science a reality. *Science, Technology & Industry Policy Papers* No. 25. Paris: OECD Publishing. doi:10.1787/5jrs2f963zs1-en.

Oez, Mikael. 2018. A guide to the documentation of the Beth Qustan dialect of Central Neo-Aramaic language Turoyo. *Language Documentation & Conservation* 12. 339–358. http://hdl.handle.net/10125/24773.

O'Meara, Carolyn & Jeff Good. 2010. Ethical issues in legacy language resources. *Language & Communication* 30. 162–170. doi:10.1016/j.langcom.2009.11.008.

Rice, Keren. 2006. Ethical issues in linguistic fieldwork: An overview. *Journal of Academic Ethics* 4(1–4). 123–155. doi:10.1007/s10805-006-9016-2.

Rice, Keren. 2011. Ethical issues in linguistic fieldwork. In Thieberger, Nicholas (ed.), *The Oxford handbook of linguistic fieldwork*, 407–429. Oxford, New York: Oxford University Press.

Robinson, Laura C. 2010. Informed consent among analog people in a digital world. *Language & Communication* 30. 186–191. doi:10.1016/j.langcom.2009.11.002.

Rundle, Hugh. 2014. *Creative commons, Open Access, and hypocrisy*. Information Flaneur: Hugh Rundle [blog]. https://www.hughrundle.net/2014/03/24/creative-commons-open-access-and-hypocrisy/.

Salffner, Sophie. 2015. A guide to the Ikaan language and culture documentation. *Language Documentation & Conservation* 9. 237–267. http://hdl.handle.net/10125/24639.

Schmidutz, Daniel, Lorna Ryan, Anje Müller Gjesdal, & Koenraad De Smedt. 2013. *Report about new IPR challenges: Identifying ethics and legal challenges of SSH Research*. Deliverable D6.2 of Data Service Infrastructure for the Social Sciences and Humanities (DASISH). http://dasish.eu/publications/projectreports/D6.1_final.pdf.

Selfe, Cynthia. L. & Gail E. Hawisher. 2004. *Literate lives in the Information Age: Narratives of literacy from the United States*. Mahwah, NJ: Lawrence Erlbaum Associates. doi:10.4324/9781410610768.

Singer, Ruth. 2014. *Open access and intimate fieldwork*. Endangered Languages and Cultures [blog]. http://www.paradisec.org.au/blog/2014/03/7940/.

Smith, Linda Tuhiwai. 1999. *Decolonizing methodologies: Research and Indigenous peoples*. London, New York: Zed Books; Dunedin, New Zealand: University of Otago Press.

Suber, Peter. Last revised 2009. Timeline of the Open Access Movement. https://legacy.earlham.edu/ peters/fos/timeline.htm. (Accessed 21 May 2019).

Thieberger, Nicholas. 2014. The cost of not archiving. Presented at the 3rd InNet conference, Budapest, Hungary, September 5–6. http://www.nthieberger.net/CostOfNotArchiving.pdf.

Thieberger, Nicholas & Simon Musgrave. 2007. Documentary linguistics and ethical issues. In Austin, Peter K. (ed.), *Language documentation and description*, vol. 4, 26–37. London: SOAS. http://www.elpublishing.org/PID/048.

Tyner, Kathleen R. 1998. *Literacy in a digital world: Teaching and learning in the age of information*. Mahwah, NJ: Lawrence Erlbaum Associates.

Urban, Jennifer M., Joe Karaganis, & Brianna Schofield. 2017. Notice and takedown in everyday practice. *UC Berkeley Public Law Research Paper No. 2755628*. doi:10.2139/ssrn.2755628.

van Driem, George. 2016. Endangered language research and the moral depravity of ethics protocols. *Language Documentation & Conservation* 10. 243–252. http://hdl.handle.net/10125/24693.

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

Whimp, Kathy & Mark Busse (eds.). 2000. *Protection of intellectual, biological and cultural property in Papua New Guinea*. Canberra: Asia Pacific Press.

Wilkins, David. 1992. Linguistic research under Aboriginal control. *Australian Journal of Linguistics* 12. 171–200.

WIPO (World International Property Organization). 2016a. *Traditional knowledge and intellectual property*. Background Brief No. 1. http://www.wipo.int/edocs/pubdocs/en/wipo_pub_tk_1.pdf.

WIPO (World International Property Organization). 2016b. *Documentation of traditional knowledge and traditional cultural expressions*. Background Brief No. 9. http://www.wipo.int/edocs/pubdocs/en/wipo_pub_tk_9.pdf.

Mandana Seyfeddinipur
ms123@soas.ac.uk