ARCHIVING AND LANGUAGE DOCUMENTATION

Patience Epps, Susan Smythe Kung, Denny Moore, Jorge Rosés Labrada, Zachary O'Hagan, Ana Paula Brandao

https://doi.org/10.1590/SciELOPreprints.4755

Submitted on: 2022-09-17
Posted on: 2022-09-21 (version 1)
(YYYY-MM-DD)

ARCHIVING AND LANGUAGE DOCUMENTATION

OS ACERVOS E A DOCUMENTAÇÃO LINGUÍSTICA

Experience report

Patience Epps

The University of Texas at Austin, Department of Linguistics, Austin, Texas, pattieepps@austin.utexas.edu, https://orcid.org/0000-0002-7429-7885

Susan Smythe Kung

The University of Texas at Austin, Archive of the Indigenous Languages of Latin America (AILLA), Austin, Texas, skung@austin.utexas.edu, https://orcid.org/0000-0002-3582-1613

Denny Moore

Museu Paraense Emílio Goeldi, Área de Linguística, Belém, Pará, dennymoore5@gmail.com, https://orcid.org/0000-0001-6249-1317

Zachary O'Hagan

University of California, California Language Archive & Survey of California and Other Indian Languages, Berkeley, California, zohagan@berkeley.edu, https://orcid.org/0000-0002-2720-2070

Jorge Rosés Labrada

University of Alberta, Department of Linguistics, Edmonton, Alberta, jrosesla@ualberta.ca, https://orcid.org/0000-0002-4454-7396

Ana Paula Brandão

Universidade Federal do Pará, Faculdades de Letras e Programa de Pós-graduação em Letras, Belém, Pará, apbrandao@ufpa.br, https://orcid.org/0000-0002-1635-9929.

ABSTRACT

As more and more of the world's languages become endangered, their documentation provides key resources for linguists and communities. Documentary linguists look to digital archives as an essential resource for ensuring the preservation, conservation, and access of the outcomes of their work. In this article, we consider the benefits and challenges associated with archiving in language documentation, relating to issues of preservation, conservation, access, ownership, and use of materials. We draw on our accumulated knowledge as scholars who are deeply involved in administering, contributing to, and using language archives, particularly relating to the indigenous languages of Latin America. We focus in particular on the relevance of language archiving in Brazil, and its significance for scholars, community members, and other stakeholders. Our discussion considers the steps that are needed to ensure the quality and longevity of resources; the principles and strategies by which archived materials may be made available; and ways in which language archives can inform ongoing work with indigenous languages. As we lay out here, language archives provide key resources for scholars and for communities who wish to revitalize, maintain, or simply remember their linguistic and cultural heritage.

KEYWORDS

Language documentation, Archiving, Indigenous languages of Latin America.

RESUMO

Enquanto que mais e mais línguas do mundo se tornam ameaçadas, sua documentação fornece recursos importantes para linguistas e comunidades. Os linguistas olham para os acervos digitais como um recurso essencial para garantir a preservação, conservação e acesso dos resultados de seu trabalho. Neste artigo, consideramos os benefícios e desafios associados ao arquivamento na documentação linguística, relacionados a questões de preservação, conservação, acesso, propriedade e uso de materiais. Baseamo-nos em nosso conhecimento acumulado como acadêmicos profundamente envolvidos na administração, contribuição e uso de acervos linguísticos, particularmente relacionados às línguas indígenas da América Latina. Nós nos concentramos em particular na relevância dos acervos linguísticos no Brasil e sua importância para acadêmicos, membros da comunidade e outras partes interessadas. Nossa discussão considera os passos necessários para garantir a qualidade e longevidade dos recursos; os princípios e estratégias pelos quais os materiais arquivados podem ser disponibilizados; e maneiras pelas quais os acervos linguísticos podem informar o trabalho em andamento com as línguas indígenas. Conforme apresentamos aqui, os acervos linguísticos fornecem recursos importantes para acadêmicos e comunidades que desejam revitalizar, manter ou simplesmente lembrar sua herança linguística e cultural.

PALAVRAS-CHAVE

Documentação linguística, Arquivamento, Línguas indígenas da América Latina.

Author contributions:

Patience Epps: Conceptualization, Resources, Visualization, Writing - Original Draft (§1, §7), Writing - Proofreading and Editing.

Susan Smythe Kung: Conceptualization, Resources, Visualization, Writing - Original Draft (§2), Writing - Proofreading and Editing.

Denny Moore: Conceptualization, Resources, Visualization, Writing - Original Draft (§3), Writing - Proofreading and Editing.

Zachary O'Hagan: Conceptualization, Resources, Visualization, Writing - Original Draft (§5), Writing - Proofreading and Editing.

Jorge Rosés Labrada: Conceptualization, Resources, Visualization, Writing - Original Draft (§4), Writing - Proofreading and Editing.

Ana Paula Brandão: Conceptualization, Resources, Visualization, Writing - Original Draft (§6), Writing - Proofreading and Editing.

Introduction[1]

As more and more of the world's languages become endangered, their documentation provides key resources for scholars and communities. Documentary materials provide an empirical basis to inform our knowledge about what is possible in human language, a register of diverse cultural and discursive traditions, and a tangible record of community heritage, offering future generations access to the voices of their parents and grandparents. Yet these materials tend to be fragile and ephemeral. Audio and video cassettes break down, notebooks mildew and fade, and even SD cards and hard drives are susceptible to fire, flood, and changing technologies – as underscored by tragic events like the 2018 fire in Brazil's Museu Nacional, in which countless precious recordings and manuscripts were lost. More and more, linguists and others look to digital archives as an essential resource in ensuring the preservation of and access to the outcomes of language documentation work.

In this article, we consider the benefits and challenges associated with archiving in language documentation. Our discussion draws on our accumulated knowledge as scholars who are deeply involved in administering, contributing to, and drawing on language archives, with an emphasis on the indigenous languages of Latin America. We focus in particular on the relevance of language archiving in Brazil, and its significance for scholars, community members, and other stakeholders.

We begin by sketching out the basics of archiving for language documentation initiatives – why archiving matters, what it offers, and how to go about it (§2). Toward the goal of ensuring the quality, longevity, and accessibility of resources, we consider contemporary best practices in digital curation, the differences between an established language archive and other online platforms, the benefits of archiving, and decisions regarding what to archive and when to do it. We also explore the question of deciding where to archive – that is, determining what archives are available and how they are set up and maintained – particularly with an eye

[1] This article follows a presentation in *Abralin ao Vivo* on July 11, 2020. The recorded presentation is available at https://aovivo.abralin.org/lives/archiving-and-language-documentation/.

to Brazilian indigenous languages and associated documentation projects; the archive of the Museu Paraense Emílio Goeldi provides an instructive case study (§3). Our discussion then considers archives and communities, focusing on the ethics of informed consent and issues of community access to documentation, informed by a case study from Jorge Emilio Rosés Labrada's work with Mako speakers in Venezuela (§4). The significance of archiving legacy materials – resulting from documentation carried out prior to the digital era, and frequently represented in fragile media with a limited lifespan – is addressed in §5. Finally, §6 returns to the Brazilian context with a detailed case study of documentation and archiving projects involving two indigenous languages of Brazil, Paresi-Haliti and Enewane Nawe, carried out by Ana Paula Brandão, which highlights many of the issues discussed in the previous sections. Final observations are offered in §7.

2. Archiving language documentation data

This section lays out principal considerations relating to the importance and process of archiving language documentation data. We provide a general assessment of the significance of this initiative (§2.1), followed by an overview of the types of digital repositories (§2.2) and the benefits of archiving for language documentation (§2.3). We explain how digital language archives differ from other online platforms (§2.4), and offer some advice on how and when language documentation data should be archived (§2.5). Throughout this article, we use the term *data* to refer to language documentation materials; that is, recorded examples and/or observations of spoken or signed language that can be processed, annotated, and analyzed (see e.g. GOOD, 2022). Primary data are the raw audio or video recordings or written observations of language, including narratives, oral histories, elicitation, conversations, interviews, and experimental protocols; secondary data are transcriptions, translations, morphological segmentations, glosses, and other types of annotation that require some level of preliminary analysis to create (KUNG, et al., 2020; HIMMELMANN, 2012; THIEBERGER and BEREZ, 2012). For accessible overviews of language archiving and its history, see also Henke and Berez-Kroeker (2016), Berez-Kroeker and Henke (2018), Kaplan and Lemov (2019), and Kung (2020).

2.1. The importance of archiving

In his groundbreaking article that defined the field of language documentation, Himmelmann (1998) includes archiving as one of the four key steps in creating what he calls "a language documentation" (p. 171); that is, a collection of transcribed and annotated audiovisual recordings with their accompanying metadata. *Metadata* are the supporting contextual, technical, and administrative documentation that helps to explain the data, including any keys (e.g., codes, orthographies) needed to understand, analyze, and reuse them (KUNG et al., 2020). Woodbury (2003) calls such a collection of primary data and the associated metadata a "*corpus*" and he includes archiving as one of six criteria that establish the overall quality of a documentary corpus.

Key reasons to archive language data are to ensure their longevity and accessibility. As we discuss below, digital repositories offer options for replicability and protection against the hazards of fire, flood, loss, mold, insects, etc. that threaten the conservation of physical materials, such as those seen in Figure 1.

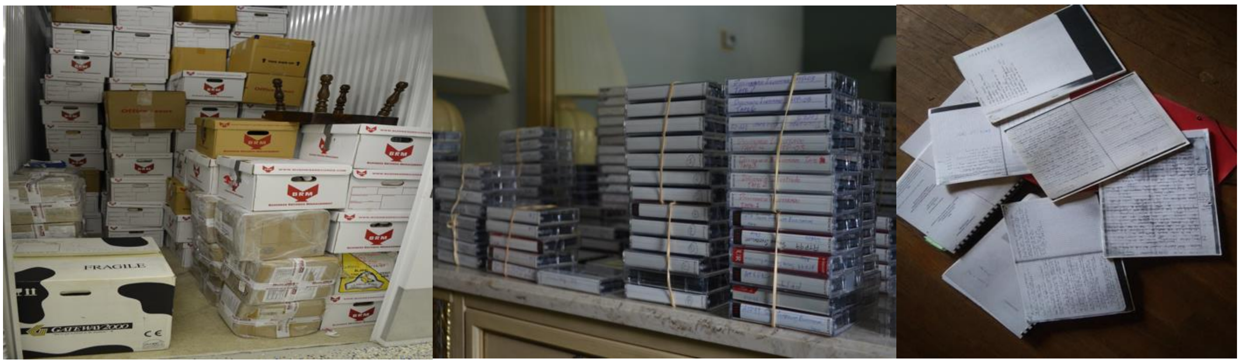


Figure 1. Physical copies of language documentation materials: boxes of papers, stacks of audio cassettes, handwritten and photocopied notebooks (ailla:257492, DSC_0001; ailla:257543, DSC_0052; P. Epps)

Himmelmann (2006) expounds on the importance of archiving the primary data that result from a research project in addition to publishing the analytical results of the project, that is the Boasian trilogy of a grammar (sketch), dictionary, and set of texts (WOODBURY, 2003). When the primary data are archived, those data can be reused for additional language documentation work, as well as other types of linguistic analysis. Further, archiving the primary and secondary data allows for the analytical output to be verifiable and reproducible (HIMMELMANN, 2006; BEREZ-KROEKER et al., 2018). Ultimately, archiving primary and secondary data in a digital repository where they are publicly accessible facilitates data reuse, and provides a stable means for citing the data (for recent guidance, see CONZETT and DE SMEDT, 2022), in the form of a persistent identifier. Persistent identifiers such as Digital Object Identifiers (DOIs), Handles, and Uniform Resource Identifiers (URIs) allow creators of the data to receive proper attribution for their work (BEREZ-KROEKER et al., 2018).

Language documentation materials are collected with great effort, time, and (often) money; however, they can represent something much more profound to the speakers: their culture, identity, and self-determination (UNDRIP, 2007; CARROLL et al., 2020). Thus, it is extremely important to ensure that speakers have access to the documentation. Ideally, copies of the documentation are left in the community from the outset, but otherwise a copy should be repatriated or returned to the community (KUNG 2020; VAPNARSKY, 2020; R. MILLER, 2021; see §4.3). Putting the data into a digital archive is a form of digital repatriation (KUNG, 2020), assuming communities have relatively unobstructed access to the data. While this is not the only way that data should be repatriated to the community of origin, it is *one* way to share these materials with the community while also preserving them for future generations of speakers. Nevertheless, it is still the case that many indigenous peoples and communities all over the world do not have (adequate) access to the internet. In these cases, a local, regional, or national archive, library, museum, school, governmental office, or some other location that is accessible to the speaker community should be identified, and a copy of the data should be deposited there (WILBUR, 2014) in addition to being deposited in a more broadly accessible digital repository, which will be better equipped to preserve the digital data for the long term (see below). Regardless of exactly where the data are archived, sharing and repatriating the primary and secondary data and resulting publications supports language maintenance, reclamation, and revitalization efforts by making materials available to speakers and their descendants. While we would hope that the data could be used for language education, maintenance, or revitalization efforts, it is impossible to foresee the actual or exact uses that the speakers and their descendants will make of the data (for recent examples see LUKANIEC, 2022; SPENCE, 2018; and VAPNARSKY, 2020). Nevertheless, the data contain the languages and cultural heritage of these speakers, and they deserve the right to decide how and when to use them (HOLTON et al., 2022).

2.2. Digital repositories

Throughout this article, we use the term *archive* to refer to a repository of materials that are saved and preserved by an organization or institution so that they can be reused in the present and the future. An archive can hold analog materials, digital materials (both born digital and digitized) or both. A *digital repository* is an archive of digital materials or records, and a *(digital) language archive* is an archive that specializes in linguistic cultural heritage and language materials or data. In the field of language documentation, the most quoted definition of an archive is from Johnson (2004, p. 142): “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”. Johnson’s definition does not mention the digital component, so we offer an updated definition from Trevor Owens, the Head of Digital Content Management at the US Library of Congress. This definition emphasizes the misconception that software designed for digital asset management is the same thing as a digital repository. From this point forward, we use the terms *archive* and *repository* interchangeably. Owens (2018, p. 4) writes: “*A repository is not a piece of software*… A repository is the sum of financial resources, hardware, staff time, and ongoing implementation of policies and planning to ensure long-term access to content. Any software system you use […] to preserve and provide access to digital content is by necessity temporary […] it likely will not last forever […] *Institutions make preservation possible*” (emphasis in the original). There are some recurring themes across these two definitions: institution, commitment, ongoing, long-term, and preservation. We will touch on each of these themes in this section, but the take-away here is that you cannot simply build a digital archive and forget about it. An archive is an *ongoing commitment* on the part of the institution or organization that decides to create it.

Fortunately, there are many institutions around the world that have made the commitment required to maintain digital archives. The Digital Endangered Languages and Musics Archives Network (DELAMAN, https://www.delaman.org/), is a network of member archives that specialize in language documentation data. Some of these archives are only digital, while others have physical holdings as well. Some specialize in a specific area of the world (e.g., the Archive of the Indigenous Languages of Latin America, AILLA, has a regional focus indicated in its title), others will accept material from anywhere in the world (e.g., the California Language Archive, CLA). Some archives will accept only data associated with a particular grant or funder (e.g., the Endangered Languages Archive, ELAR, mainly accepts materials collected with grants from the Endangered Languages Documentation Programme), while others primarily accept materials in or about a specific language (e.g., Standing Rock Sioux Tribe Language and Culture Institute). As part of the generalized increase in awareness of the endangered state of the majority of the world's languages, several Latin American institutions, including in Brazil, also committed to hosting local language archives; see Seifart et al. (2008) for an in-depth discussion of the early stages of this process. Unfortunately, many of these archives have not stood the test of time. Two repositories in Brazil, however, have continued to operate and are discussed here in §3. Other archives have also been created more recently, such as the *Archivo de Lenguas, Culturas y Memorias Históricas del Ecuador*, which has institutional support from FLACSO Ecuador (http://languages.flacso.edu.ec/).[2]

While the above-named archives cover a wide range of geographic locations and collecting requirements, they might not be a good fit for all language documentation projects. Alternative archives include other types of digital data repositories such as those managed by many universities. Linguistic datasets (but not primary data from language documentation) may be submitted to the Tromsø Repository of Language and Linguistics (TROLLing, https://site.uit.no/trolling/about/). Primary and secondary data, as well as datasets, may be archived in general data repositories such as Zenodo (https://zenodo.org/) and the Harvard DataVerse (https://dataverse.harvard.edu/). Additional possibilities can be found in the Registry of Research Data Repositories (https://www.re3data.org/).

2.3. Advantages of digital language archives

For researchers engaged in documentation projects, there are many advantages to archiving language documentation materials (data). These include simplified format migration, personal organization, accessibility, discoverability, graded access, rights management, collaboration, citability, and long-term digital preservation.

Anyone who has ever had to copy field recordings from an obsolete medium (e.g., cassette tapes) to another format, tried to open an old file in a new version of the program, or endeavored to find a program that can even open the old file, understands the headaches involved in *format migration*.[3] However, researchers who archive language documentation materials in a digital repository never need to worry about format migration again because it is part of the archival workflows that are overseen by the archive's personnel.

Archiving also facilitates personal organization. Shortly after returning from the field, most researchers still remember where all of the files they created are located and how they are organized. However, as time passes, memory fades, and it becomes much harder to remember where files are, how they are organized, how they relate to each other, and the relevant metadata (i.e., contextual information) that will allow for their future reuse, either by the researcher who created them or by others. The situation is amplified for digital files, which cannot easily be distinguished and, thus, will need to be viewed or opened to be identified. The solution is to

[2] FLACSO stands for *Facultad Latinoamericana de Ciencias Sociales*.

[3] See Han (2022) for a recent related discussion of the many ways data can be *transformed*.

archive language documentation data, as well as the associated metadata, as soon as possible after they are created. Once the data and metadata are archived in a trustworthy repository, they can always be found and accessed again. Researchers no longer need to remember where files are stored (e.g., on which external hard drive, laptop, or cloud storage system), and they no longer need to juggle the work of handling multiple copies and backups of the data. Archiving helps to ensure that the data and metadata are not lost, thrown out, or forgotten about.

Once the materials and metadata are archived, they are *accessible to* (i.e., they can be accessed by) their creators, as well as by collaborators, other researchers, members of the language community, and anyone else who might need access to the data. Furthermore, digital language archives are constructed in ways that support *discoverability* of materials, that is, the ability of users to find or discover materials they are looking for. Discoverability is relevant both to searches within the archive itself (via facet or targeted searching, advanced searches, etc.), and to searches carried out via broader internet platforms such as Google indexing and metadata harvesting by the Open Language Archives Community (OLAC).

While there are digital repositories that claim to be purely open access, all digital language archives have rules for how users may interact with the holdings, and most digital language archives have *graded access*, by which certain materials may be restricted to particular users and/or conditions of use (see also §2.5). Regarding the rules of use, all digital language archives have some sort of terms or conditions that the online visitors must agree to before they can access the media files. Many digital archives require a user to create a free account and log in before they can access the media files. However, the catalog information (i.e., the metadata) is usually publicly accessible, meaning that anyone who lands on the archive's web page can access and read it. Most language archives have some form of graded access to media files, though the way graded access works varies greatly between archives. AILLA uses numbered levels that indicate grades of access; other archives have specialized user roles and some use color coding to indicate who can access specific materials.

Most digital archives that specialize in language documentation data handle rights management similarly, though exact details vary between archives. In most cases, the original *rights holders*[4] retain all of their intellectual and/or cultural property rights. The rights holders give non-exclusive licenses to the archive and the archive's users; details of the licenses vary

[4] The rights holders are the persons who have the right to claim copyright over a creation or to claim the rights to the intellectual property, cultural property, moral rights, performance rights, or other rights inherent to a work or creation by virtue of having created it, group membership/affiliation, or inheritance.

between archives and according to specific copyright laws of the country where each archive is located. Commercial use of the data is *never* allowed.

Data archiving facilitates collaboration on many levels. For researchers who are engaged in remote collaboration with the speech community or other researchers, archiving data as they are created or as analyses are finished helps the entire project team stay organized. Also, once data are organized and archived, they can be discovered by other researchers, which in turn can lead to new opportunities for collaboration. Finally, archiving data helps to facilitate the reuse of the primary data for various research purposes, including a great deal of work that is being done in the areas of natural language processing and linguistic typology. Finally, archived data can be cited, and this is crucial for creating reproducible research (BEREZ-KROEKER et al., 2018). Some journals, such as *Language Documentation & Conservation*, request that the data sets associated with an article be archived and cited appropriately. Thus, researchers can and should cite their own archived data, and researchers who use archived data must cite them as well.

Finally, a trustworthy digital repository has an active plan for the *long-term digital preservation* of the digital media files and associated metadata. According to the Digital Preservation Coalition (DPC), *digital preservation* "[r]efers to the series of managed activities necessary to ensure continued access to digital materials for as long as necessary. [...] [It] refers to all of the actions required to maintain access to digital materials beyond the limits of media failure or technological and organizational change." *Long-term preservation* is the "continued access to digital materials, or at least to the information contained in them, indefinitely" (DPC, 2015). This means that digital preservation is much more than just backing up files. Digital preservation work includes migrating data and metadata from one format to another and from one software system to another as technology changes. It means ensuring accurate redundancy of data (meaning that there are duplicate copies stored on various media types and in multiple locations); and it means monitoring the health of all of the files in all of the locations on a regular basis. According to Owens (2018, p. 5), "preservation is the result of ongoing work of people and commitments of resources. The work is never finished […] It is not something that can be thought of as a one-time cost."

2.4 Differences between language archives and other online platforms

Now that we have discussed the advantages of archiving language documentation data, we want to contrast digital archives with other online platforms that are commonly confused with digital archives, such as social media sharing platforms, websites, and cloud-based file storage.

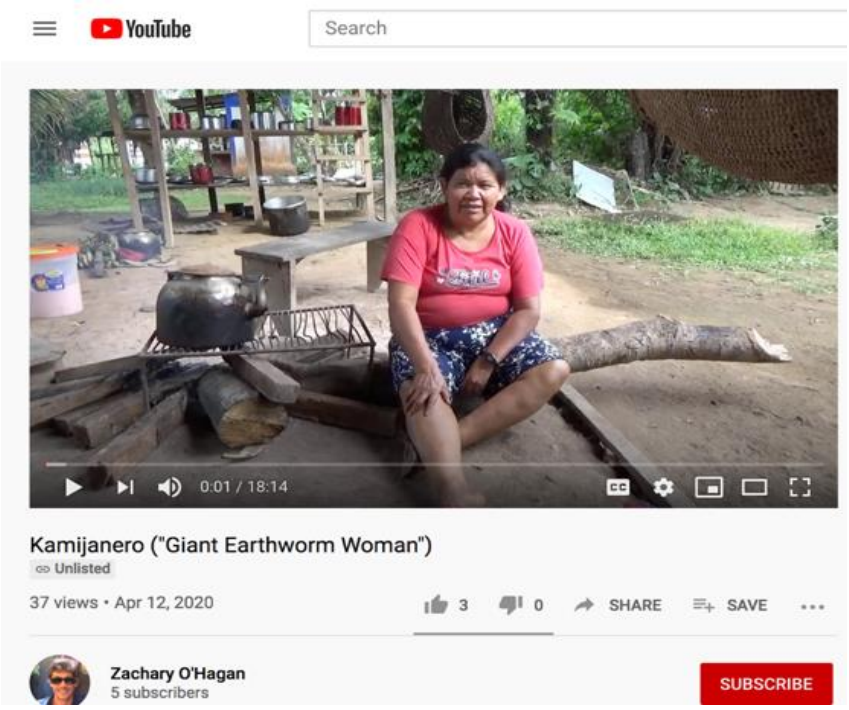
Social media sharing platforms, like YouTube, Vimeo and SoundCloud (see Figure 2), facilitate sharing video and audio files with communities, speakers, and others because their content is easily discoverable via online search engines. They facilitate rights management because they allow the person who uploads the file to choose between traditional copyright or the application of a Creative Commons license.[5] However, most people do not have a sufficient understanding of either traditional copyright or the application of Creative Commons licenses to make an informed decision. These platforms also have their own versions of graded access in that the content may be kept private or made public.

Websites also facilitate content sharing, and code can be added to make the pages discoverable by search engines. While limiting access to content is possible, it can be a technically challenging process. Traditional copyright automatically applies to websites, but the website developer can choose to apply Rights Statements[6] or open licenses, such as those managed by Creative Commons, to the webpages and/or the linked files.

File storage systems, like DropBox, Google Drive, OneDrive, and Box, are good for sharing files with your collaborators while you are still working on them, as well as for controlling access to the files, but they do not have built-in rights management, and they are not discoverable in web searches.

While all these options are great for sharing data, none of these systems is committed to the long-term digital preservation of files or data. While they all might offer some form of content backup, they do not promise long-term preservation. The terms of use of these platforms include their right to discontinue service and delete accounts and their contents. Moreover, some platforms, such as YouTube, retain copies of materials even when files are deleted by their owners/posters, or may be subject to automated downloads by other websites – an obvious problem where sensitive data are concerned (RICE, 2021).

[5] Creative Commons licenses are a type of open license that work in conjunction with traditional copyright; see https://creativecommons.org/ (Brazil: https://br.creativecommons.net/) for more information.
[6] https://rightsstatements.org/en/


Kamijanero ("Giant Earthworm Woman")

Unlisted

37 views • Apr 12, 2020

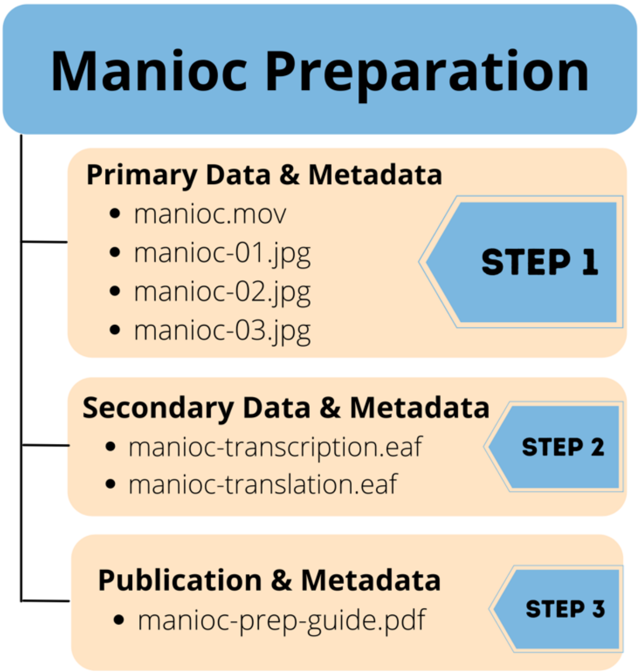
Zachary O'Hagan
5 subscribers

Figure 2. Video on YouTube of a story told in Caquinte by Antonina Salazar Torres (recorded by Zachary O'Hagan, permanently archived with the California Language Archive, DOI: http://dx.doi.org/doi:10.7297/X2Z60M7W)

2.5 How and when to archive language documentation data

Prior to the 21st century, the results of language documentation were typically archived at the end of a researcher's career or even after their death. However, with the advent of born-digital recording equipment, language archives, and data archiving requirements, most researchers today do not wait that long.

Today's motto – as frequently voiced by archivists – is "Archive early and archive often!". Many language archives recommend progressive or incremental archiving. Under this model, a researcher or research team submits primary data (e.g., audio and video recordings and photographs), always accompanied by the relevant metadata (such as names of participants, date, location, languages used, and descriptive information to contextualize why the media files were created), to a digital repository as soon as possible after the media files are created (see step 1 in Figure 3 below). Secondary data (e.g., annotations, transcriptions, interlinear glossed texts) and analyses or academic output (steps 2 and 3, respectively, in Figure 3) are added later after they are finalized (ROBINSON, 2006; NATHAN, 2013; KUNG et al., 2020). In Figure 3, the steps represent waves of archiving. Step 1, which involves archiving the primary data and relevant metadata, should be done as soon as the fieldwork or data collection phase is complete, or even while fieldwork or data collection are still in progress. Step 2, archiving the secondary data and relevant metadata, can be done iteratively as the transcriptions and translations are finalized. Step 3, archiving finalized analyses and academic output, along with relevant metadata, can also be done iteratively and well into the future. This approach recognizes that, for most researchers, or anyone else, there will never be a truly convenient time to archive language documentation data. The more time that passes and the more digital files and physical materials that accumulate between data collection and archiving, the harder and less convenient the archiving process will be, and the longer it will take. Moreover, since everything is fresh in a researcher's mind immediately after a period of fieldwork or data collection, it is much easier at that point to organize materials accurately and to make sure that the metadata are thorough.[7]


Manioc Preparation

Primary Data & Metadata
- manioc.mov
- manioc-01.jpg
- manioc-02.jpg
- manioc-03.jpg

STEP 1

Secondary Data & Metadata
- manioc-transcription.eaf
- manioc-translation.eaf

STEP 2

Publication & Metadata
- manioc-prep-guide.pdf

STEP 3

Figure 3. Progressive archiving (graphic by S. Kung, CC BY-SA 4.0)

While archiving is key to the long-term conservation, preservation, accessibility and discoverability of language documentation data, not all data are equally appropriate for archiving, or for archiving in the same way. Some materials may be culturally or personally sensitive to the point that they simply should not be archived at all. Other materials may require restricted access, as noted in §2.3 above, such that only certain people may engage with them. Sometimes such restrictions may pertain in different ways to, for example, an audio file, a video file, or a written transcription of the same event. Examples of potentially sensitive material include esoteric or protected knowledge that is not meant to be shared with particular people or groups of people (e.g., community outsiders, members of other clans, men vs. women),

[7] For more information on navigating the archiving process, see 'Archiving for the Future: Simple Steps for Archiving Language Documentation' at https://archivingforthefuture.teachable.com/.

personally damaging speech, and information that could endanger individuals or communities. Language documentation and archiving calls for open and ongoing communication among documenters, speakers/signers, and other community members in order to ensure a fully ethical and informed process, as we elaborate in §4 below.

Finally, as this section has explored, it is essential when archiving language documentation materials to choose a repository that will ensure their long-term preservation and accessibility. Many archives with the necessary infrastructure, including many of those in the DELAMAN consortium, serve an international community, and should not be viewed as proprietary with respect to whatever country might host them. Nonetheless, archiving initiatives are valuable at all levels – local, regional, national, and international. In some cases, it may be advisable to archive materials in more than one location, in order to meet local priorities while also leveraging higher-level infrastructure that offers the most reliable preservation and access (see §3.2 and §6 below).

3. Language archives in Brazil: the Museu Goeldi and other initiatives

In this section, we take a closer look at archiving initiatives within Brazil, and their relevance for indigenous language documentation within this country. In light of the large number of indigenous languages in Brazil and their precarious situation (MOORE, GALUCIO, and GABAS, 2008), language documentation is urgent, and it is widely supported among indigenous groups. In the survey of the languages of the state of Rondônia by the National Inventory of Linguistic Diversity, documentation was indicated as the second priority of indigenous people in relation to the language of the groups, with the correction of defective orthographies and production of correct written material being the first priority (GALUCIO, MOORE and VAN DER VOORT, 2018, p. 217). Many indigenous groups have young people trained in computer literacy, a fact that facilitates digital documentation. Potentially, a large number of recordings can be produced. Their permanent preservation and availability requires a huge storage capacity in digital files.

In what follows, we describe the collection of indigenous language materials housed at the Goeldi Museum (Belém, Brazil), as a concrete example of a language archive based in Brazil – how it began and was created, and what it contains (§3.1). We then provide a brief overview of other digital language archiving initiatives in Brazil (§3.2), of which the Museu do

Índio is currently the principal exemplar.[8] We also describe several language documentation projects and their outcomes as an illustration of the general need for documentation and archiving in the country, and discuss the potential of regional archives.

3.1. The Museu Goeldi archive

The Museu Paraense Emílio Goeldi is a research institute associated with the Ministry of Science, Technology and Innovation (MCTI). Located in Belém, in the state of Pará, the Goeldi focuses on research on the Amazon. Its Linguistics Area, part of the Human Sciences Division, has had a precarious history. In 1986, the Area had no recorder, no computer and only a few books. As is often the case in the country, there was little investment in infrastructure. However, there was a respected tradition of scientific collections at the Goeldi, and language documentation fit well into that tradition. Interns, fellows and visiting researchers fostered increasing levels of activity within the Linguistics Area, including documentation.

The infrastructure began to improve with the acquisition, through research projects, of good quality cassette tape recorders, as well as microphones, laptops and solar equipment. In 1996, the World Bank's Centers of Excellence program brought Digital Audio Tape (DAT) players, Hi-8 video recorders, a Hi-8 editing island and professional storage cabinets. On the basis of accumulated experience, Goeldi linguists, including fellows and affiliates, have successfully competed in international documentation programs that are favorable to projects in developing countries. In the early 2000s, projects from the Endangered Languages Documentation Programme (ELDP) and the Dokumentation Bedrohter Sprachen program (DOBES) provided more equipment and experience. Documentation and procurement of equipment (e.g. solid state recorders with flash memory) continued into the second half of the decade, with support from USAID, UNESCO and the Ambassador's Fund (US Embassy). MCTI supplied equipment in 2007, including a server as the basis for a digital repository. Documentation (and research) projects continued in the second half of the 2000s, with support from CNPq, DOBES and ELDP. This progress was threatened in 2007 by fulminating attacks by a small group of linguists, who opposed the ProDocLin Museum's linguistic documentation program,[9] international documentation programs and digital archives (MOORE and

[8] This presentation thus constitutes an update of the discussion about incipient archiving efforts in Brazil described in Seifart et al. (2008).

[9] http://prodoclin.museudoindio.gov.br/

GALUCIO, 2016, p. 40). Fortunately, this opposition did not attract support among Brazilian linguists and progress continued, both at the Goeldi Museum and in Brazil.

Through 2009, the activities of linguists based at the Goeldi resulted in a large set of diverse recording media: cassette tapes, DAT tapes, mini-disks, Hi-8 tapes, mini-DV tapes, etc. Due to time and budget constraints, this collection was simply placed in drawers, without systematization. From 2009 to 2014, linguist Ana Vilacy Galucio coordinated a project supported by the CFDD/Ministry of Justice to digitize and catalog the collection, following the practices of the best international archives. To streamline the process of cataloging and storing recordings and their metadata, scripts were created by Sebastian Drude and Rose Costa, reducing the time required by 70%. Thus the Goeldi's digital linguistic archive was created.

Subsequent years saw more equipment purchased and more documentation projects conducted. Throughout the evolution of the Goeldi Linguistics Area archive, technology has evolved and formats and physical media have changed. However, the conversion of all recordings to standardized digital form has mitigated problems of longevity and compatability. Some devices, such as microphones, do not become obsolete, and the various types of recorders (DAT, mini-DV, etc.) that are currently not used directly for documentation are maintained, since they can still be used to access legacy recordings and produce a signal that can be digitized. Original recordings are stored on their original media in professional storage cabinets .

Currently, recordings in digital form are stored on a 32TB Network Attached Storage (NAS). For security, in case of failure of one of the NAS hard drives, RAID-6 redundancy is implemented, reducing the storage capacity to 22TB. The contents of that NAS are copied to the new 96TB NAS, whose capacity is reduced to 72TB in light of redundancy (see Figure 4). To ensure their safety, in case of risks such as fire or lightning, the two NAS must be kept in different buildings. Audio recordings are saved in their original format, i.e. as .wav files. Video recordings are saved in their highest definition format and also in compressed (.mpg) format, which takes up much less space. These compressed files and audio files are cataloged and stored using Language Archiving Technology (LAT) software. Most of them are copied on the server of the Information and Communication Technology Sector of the Goeldi Museum.

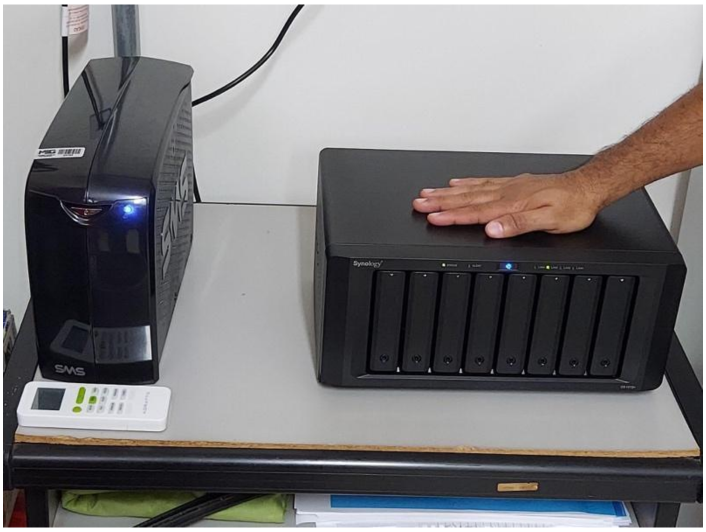


Figura 4. Network Attached Storage (NAS) de 96TB no acervo do Museu Goeldi

Currently, the Linguistics Area digital archive contains records of 80 indigenous languages. Of these, 73 are completely digitized, catalogued and stored in the LAT software. The LAT files take up 2.49 TB of space; the temporal duration of these recordings is 1,561 hours of audio and 474 hours of video. In addition to these files, the archive also contains 9TB of raw recordings, including High Definition video, which take up a lot of space. Edited works, mainly for community use, take up another 1.5 TB. Another 6 TB is occupied by photographs and loose files of the researchers. The metadata of the LAT files on the server can be found at http://arqling.museu-goeldi.br. The recordings are currently not available for download, pending reorganization of the archive and resolution of access issues.

To contribute to the development of indigenous language documentation in Brazil, the linguists at the Goeldi disseminate, through intensive training, knowledge of the technology and methodology involved, including suggestions for equipment. This training is also carried out among indigenous groups. Moreover, the archive offers digitization and storage services for legacy recordings (see §5 below). For example, the Goeldi team is currently working with an anthropologist to digitize his collection of 115 cassette tapes (both sides) of two Nambikwara dialects. The recordings, some from the 1970s, include music, a dictionary, conversations, and minimal tone pairs. The recordings are being digitized and, in collaboration with the anthropologist, will soon be catalogued and deposited in the digital linguistic archive. He will receive copies in .mp3 format, which take up little space and can be returned to Nambikwara communities. Cassette tapes can be deposited in the archive or returned to the researcher. People interested in this service should contact the Goeldi Linguistics Area (linguistica@museu-goeldi.br).

3.2. Other efforts and priorities for language archiving in Brazil

Alongside the Goeldi Museum, few other Brazilian institutions are involved in archiving initiatives. Currently, the principal effort is based at the Museu do Índio in Rio de Janeiro, a component of the Fundação Nacional do Índio. Over time, this museum has systematically increased its infrastructure and has developed impressive digital storage and backup capacities. The Museu do Índio has also been conducting active projects to document indigenous languages and cultures, following the best international practices. Its digital archive contains approximately 10 TB of linguistic documentation, including the languages of 25 indigenous groups. It also contains 9 TB of cultural documentation from 27 indigenous groups, which includes recordings of linguistic interest. Outputs of the 'Indigenous Sonorities' project (focusing on cultural preservation via audio documentation) for five indigenous groups occupy 1.69 TB. The total content of the file occupies approximately 50 TB. Copies of all recordings are returned to the communities involved, but Internet access is still being resolved.

Various well-organized linguistic documentation projects, led by Brazilian linguists, are ongoing and are producing significant amounts of recordings, steadily increasing the need for professional archiving. For example, the Documentation Center of the Federal University of Amapá in Oiapoque has documentation projects with the Karipuna and Galibi-Marworno (coordinated by Gelsama Mara Ferreira do Santos) and with the Palikur (coordinated by Elissandra Barros da Silva). Data is currently being stored on microcomputers and external hard drives. The project with the Galibi-Marworno has 13.5 GB of audio (.wav) and 403 GB of video (.mov, mp4), as well as 51 GB of photos and 15 GB of edited video. The Palikur project has 20GB of audio, 6TB of video (.mp4) and over 100,000 photographs in RAW format, for a total of 10TB.

Clearly, the demand for language documentation archives is growing in Brazil, as elsewhere, and is presently far above the current capacity. One possible solution to the need for more capacity is the creation of regional digital archives. These would not only increase the current capacity, but they would also have the advantage of relative proximity and visibility to the indigenous peoples of the region. This could facilitate both further recording and indigenous groups' access to those recordings. A natural place for archives of this type is indigenous training programs at universities, for example at the Universidade de Amazonas, the Universidade Federal de Goiás, or the Universidade Federal do Amapá in Oiapoque. In these programs, indigenous students are trained in computer literacy while maintaining contact with their respective groups, thus facilitating a productive cooperation between the archive and the groups in question. The infrastructure for an archive of this type can be bought at once or, as in the case of the Goeldi, built up, in stages, in tandem with research or documentation projects. The cost of a 48TB NAS (Network Attached Storage) was approximately R$20,000 in the middle of 2021 (with the dollar at R$5.80). For digital language documentation, a complete set of excellent portable semi-professional equipment (camcorder, tripod, digital audio recorder, laptop, three types of microphones, lamp, carrying case, batteries, cords, adapters) was approximately R$20,000. At least two kits would be needed to equip a regional documentation center and archive. In addition to infrastructure, training in documentation technology and methodology is required to ensure quality, which is always a challenge. One issue to be resolved is the need for an archiving software that is reasonably simple and user-friendly.

Ultimately, the archiving options presented in this section represent different levels of capacity regarding the major goals of preservation, conservation, and accessibility, in keeping with the observations presented at the end of §2 above. While smaller-scale archiving initiatives such as those proposed here may not meet DELAMAN standards with respect to some of these goals, they are certainly a step in the right direction, and in some cases they may offer more options for accessibility even where they are less developed in their capacity for long-term preservation. Moreover, as noted above, researchers should be aware of the possibility of archiving *both* at a local or regional level and also in a higher-infrastructure archive. This possibility is illustrated by the Paresi and Enawene Nawe case studies described in §6 below.

4. Archives and communities

This section considers ways in which members of a language community can be involved in archival projects, and the ethical questions of informed consent and access that accompany their involvement.

Community members may interact more or less directly with an archive. An example of quite direct interaction can be seen in the “Verdena Parker Collection of Hupa Sound Recordings and Films” (PARKER, n.d., held by California Language Archive). Verdena Parker, a native speaker of Hupa (Dene, aka Athabaskan; California), made sound recordings and films over many decades, culminating in her collaboration with linguists from the University of California, Berkeley on a documentation project, initiated in 2005 (see PARKER et al., 2005+). These recordings, archived in 2010, include texts, vocabulary, translations, and observations about life in Hoopa Valley, for use in revitalization programs.

However, the vast majority of archival projects involve at least some speakers who do not interact directly with the archive, and in many cases may not have prior familiarity with archiving or what it entails. The "Kawahiva Language Documentation Archive" (DOS SANTOS, 2017+) is one example; this collection, developed on an ongoing basis, consists of audio and video recordings of stories, conversations, songs, elicitation, meetings, field notes, and photographs. Two speakers record, transcribe, and send files monthly for archiving, "creat[ing] the sentiment that the archive belongs to them as well" (DOS SANTOS, p.c.), and depositor Wesley dos Santos has also created a video to explain to community members how to access the materials via the California Language Archive.

Whatever the affiliation of the person or people who lead a documentation and archiving initiative, they work under an ethical imperative to clearly communicate, explain, and obtain consent for these efforts from the speakers and other community members involved. Below, we explore these ethical requisites and offer some strategies to meet them (§4.1), illustrated via a case study from co-author Rosés Labrada's work with the Mako people of Venezuela (§4.2). At the end of this section (§4.3), we consider strategies for enabling community access to archival materials.

4.1. Ethics and informed consent

Discussions of ethics and ethical best practices have figured prominently in the language documentation literature. In these discussions, particular attention has been paid to collaboration (GLENN, 2009; LEONARD and HAYNES, 2010), community involvement and engagement (YAMADA, 2007; CZAYKOWSKA-HIGGINS, 2009; SAPIÉN, 2018; BISCHOFF and JANY, 2018), and the applicability of ethical models to different parts of the world (DOBRIN, 2008; HOLTON, 2009; PÉREZ BÁEZ, ROGERS and ROSÉS LABRADA, 2016). The intersection of archiving and ethics has also received considerable attention (e.g., Macri and Sarmento 2010 and Innes 2010, both of whom explored ethical issues related to archives). In these discussions, informed consent has played a key role, with multiple researchers questioning how 'informed' the consent really is (GRINEVALD, 2006; ROBINSON, 2010). An overview of some of the issues relating to informed consent may be found in Dwyer (2006, p. 43-48).

Informed consent is a *legal* – at least in many places, including Brazil – and *moral* obligation. It is, therefore, essential that speakers/signers understand what their participation entails and, crucially, what the consequences of digital archiving really are. One potential challenge to informed consent with respect to archiving is that of familiarity with archives and the Internet – and in some instances, computers – on the part of the community and/or the speakers. As Robinson (2010, p. 189) asks, "can we obtain *truly* informed consent [for archiving] if the consultant and the community have never seen a computer or heard of an archive or the Internet?". There are two additional issues here. Firstly, research is by default open-ended. We generally gather a corpus based on particular research questions but those questions, or rather the answers to them, may lead to new questions. Thus, when you gather consent at the beginning of a project, if the goals of the project change or new research questions arise, does the initial consent hold? This is particularly difficult to anticipate with archived data, which as we will see below can be reused by others. Secondly, the descendants and communities of tomorrow may need these materials for revitalization purposes, as described for example in Bomfim (2017). In what follows, we exemplify these issues and propose one approach, among many possible ones, to address some of these concerns.

4.2. Case study: Explaining archiving to community members

This case study reflects on the experience of co-author Rosés Labrada in explaining archiving to members of the Mako communities along the Ventuari River in Venezuela during his doctoral project (2012-2015) focused on the documentation and description of Mako, a Jodï-Sáliban language (ROSÉS LABRADA, 2015).

In 2011, Rosés Labrada undertook a trip to several Mako communities along the Ventuari River and its tributaries in order to gather community consent for the project as a whole. As part of this process and attending to local protocols, village meetings in several communities were held, which resulted in an invitation to return to work in two communities: Arena Blanca and San José de Yureva. However, no data were collected on that first trip. The documentation project thus started in July 2012, and Arena Blanca was selected as the site of the first field stay due to its composition as a largely homogenous Mako-speaking community.[10]

Initial preparation for a community-wide discussion about ethics and informed consent took place on July 14, 2012 in a meeting with two community members who were also the school teachers in Arena Blanca at the time and who had agreed to act as translators for the meeting. A clear outcome of the meeting with the schoolteachers was the realization that the concept of archiving was likely to be difficult for the outside linguist to explain in ways that

[10] San José de Yureva is a mixed Mako-Piaroa community where both Mako and Piaroa are spoken.

would be clearly understood due to the complexity of some of the concepts and technicalities behind them. Of significant concern was the fact that, at the time, there was no connectivity for the Internet or phones, there were no computers in the community, and both the teachers had limited experience with this technology.

The community-wide meeting on July 15 was well attended by many community adults, who engaged in discussion of the project and potential ethical issues posed by recording and archiving audiovisual materials. Because this was a meeting to seek permission to record, the meeting itself was not recorded. However, the permission seeking process advanced significantly during this meeting:

1. recording with both video and audio was accepted;
2. taking pictures was also accepted but with the caveat that pictures of unclothed children or topless women were to be avoided as the community had abandoned those traditional practices;
3. keeping the data beyond the duration of the project and sharing it outside the community was also accepted.

Nevertheless, Rosés Labrada sensed that the concept of archiving remained “fuzzy”, particularly in two areas: (1) the implications of the online sharing of materials that identify specific individuals and (2) the options for sharing or restricting access to those materials or the names of the individuals. Thus, the linguist took steps to make sure that the community really understood what was involved in the archiving process. His goal was to show them what the collection in the archive would look like, but the challenge came from the lack of Internet access in the community.

The solution adopted was to take screenshots of an initial archival deposit to show what the archive’s website looked like and what it would be like to navigate through it. In preparation, Rosés Labrada and one of the schoolteachers visited the AILLA and ELAR websites while in Puerto Ayacucho, the capital of Amazonas state, and the schoolteacher agreed to deposit one of his recorded stories to start a collection that was to be double-archived in both archives.[11] After the deposit of a folder with all the items corresponding to one single story, both archives supplied screenshots of what the collection and the deposit looked like.

[11] Both AILLA and ELAR initially agreed to host the data stemming from the research project. However, both archives moved later to a model that avoids “double-archiving” (i.e., duplication of the same collections in multiple DELAMAN repositories), as that means twice the amount of required digital storage space. The Mako collection is now hosted by ELAR (http://hdl.handle.net/2196/6bed9c49-c2dd-446d-b692-53c24cfbc916) and the initial deposit in the AILLA website (https://ailla.utexas.org/es/islandora/object/ailla%3A124494) redirects the visitor to the ELAR website.

A second community-wide meeting was held on November 4, 2012 in Arena Blanca after Rosés Labrada's return with these screenshots. The discussion that ensued was better informed in a number of aspects. First, one request was that originals stay in the community. This request allowed the linguist to provide further explanation about the context of digital documentation: that specific recordings were born digital, that originals and copies could be identical, and that copies could be easily made—which would not necessarily have been the case if we had recorded directly on cassettes or CDs/DVDs, which the community was more familiar with.

A second issue that came up was the question of access for community members and a distinction between those "who know" – meaning those who know how to use computers and the Internet – versus those "who don't know." The general consensus was that those who knew could potentially go to the Internet and use these archived materials but that for those who did not know, making copies in DVDs and CDs was needed. A third issue concerned the content of the recordings themselves; a community member expressed a preoccupation regarding recordings that may contain crude jokes or curse words and who could listen. For these, it did not seem to be a problem that people in other parts of the world could listen to these recordings. Rather, this particular community member was concerned about the potential reactions of Mako people from other nearby villages who might listen and not understand that these jokes were truly meant in jest. Finally, there was discussion of videos of traditional activities as being harmless, when compared to some other videos that the community was aware of, such as violent movies. This provided an opportunity to explain further that individual speakers had agency to restrict access both at the time of recording but also into the future, and that specific access provisions could be enabled and changed at any time in the future. Overall, this discussion reassured Rosés Labrada that community members in Arena Blanca had a clearer understanding – when compared to the initial meeting in July 2012 – of some of the ethical issues around informed consent, archiving and access.

This small case study illustrates three main points: 1) it is crucial that we explain archiving in ways that communities and individual participants can understand; 2) communities and participants should be able to revisit their decisions; and 3) we should try to anticipate future uses of the materials and make provisions where possible, while being fully aware that we cannot foresee the future. Ultimately, ethical choices are not necessarily defined in black and white, and what works in one specific context may not necessarily work in another (HOLTON, 2009; GASSER, 2017). However, as long as we aim to uphold principles of respect, reciprocity, responsibility, and relationships, as advocated by Rice (2006), we should be able to avoid some of the potential ethical pitfalls that could arise as part of the archiving process.

4.3. Enabling community access to archival materials

As the Mako case illustrates, having and maintaining access to language documentation materials tends to be a key priority for communities, not only for those people who are directly involved in documentation, but also – and sometimes even more so – for their descendants (e.g., DWYER, 2006, p. 59; VAPNARSKY, 2020; R. MILLER, 2021). Yet effectively returning language documentation materials to communities can be a complex and multi-stage process, and must take into account the different capacities community members have to access these materials. These capacities may be constrained not only by a lack of access to the internet, through which archived materials in many digital repositories can be viewed and downloaded, but also by the skills required for navigating online interfaces (as well as the languages in which these interfaces might be presented, such as English) and by basic computer literacy. For members of some communities, access even to static technologies such as CD and DVD players may be limited, as may be the capacity to read printed materials in the community language (or otherwise).

Enabling access may require creative solutions. One model that has met with considerable success is the “distributed digital audiovisual archive” or “jukebox archive” (BARWICK, 2004; BARWICK et al., 2005; O’MEARA and GONZÁLEZ GUADARRAMA, 2016). This initiative, originally piloted in Australia, involves setting up a computer in a community center or other neutral location where community members can easily access it, and does not require a connection to the internet. The jukebox computer contains language documentation materials – audio and video recordings, as well as (potentially) transcriptions, photographs, etc. – in easily movable file formats (such as .mp3), as well as the capacity to burn CDs/DVDs and to download files onto a flash drive, phone, or .mp3 player. Community members can thus take copies of materials home with them, and can also add to the jukebox by uploading their own. Another useful model relies on a local wifi transmitter called a ‘Raspberry Pi’, which is effective in contexts where people have smartphones and the associated skills, but have limited or no computer literacy (e.g. THIEBERGER, 2019).

Of course, ethical considerations are relevant to all these initiatives – and not only to the process of enabling and maintaining access to documentary materials, but also in considering how and whether community members wish to limit their access and use (see the sections above; also MACRI and SARMENTO, 2010; DEBENPORT, 2010). Some recorded material may be viewed as potentially damaging, dangerous, offensive, or otherwise private, and understandings of who should or should not have access to what material may involve complex and intersecting categories, associated with particular conceptions of outsider/insider, gender, clan, neighbor, relative, etc. – which may be much more complex than the blanket term 'community' implies. As pointed out above, decisions about access, like other facets of documentation and archiving, must be continually informed by ethical principles.

5. Legacy materials and the reach of archives

Archives have tended to increase their holdings with the proactive donations of living individuals (often academic researchers), or with bequeathments from the estates of deceased individuals. However, the reach of archives can be broadened and enriched beyond this model in a number of ways. As is becoming more common, speakers and other indigenous collaborators can contribute directly, as well as indirectly, to the archival process, as we addressed in §4. Moreover, as we explore here, researchers, community members, and archive staff can be proactive in locating and preserving legacy materials, a process that itself can involve collaboration among many stakeholders. In this section, we consider practical considerations in locating and archiving *legacy materials*, that is, written or audiovisual materials collected in the past, usually before digital recording and archiving methods were available (§5.1) and in making use of them (§5.2).

5.1. Valuing and archiving legacy materials

Researchers who work or have worked with speakers of indigenous languages are often in personal possession of rich documentary collections including field notes, sound recordings, photographs, and film, which may be stored in their offices or homes in sub-optimal or risky conditions (e.g., damage from humidity, insects, flooding). These legacy materials – even when they are not especially numerous – are of high linguistic, cultural, historical, and personal value for individuals and larger groups. For less well documented languages, legacy materials may constitute the only early records of language use, to say nothing of the other aspects of life that they often capture. Unlike widely spoken languages such as Portuguese, for which historical records are readily at hand, historical records of many indigenous languages often persist only by way of diligent efforts to archive legacy materials. As Austin (2017, p. 23, emphasis added) stresses, “For projects interested in documenting, describing or revitalizing languages, especially endangered languages, historically existing materials (whether digital or analog) like tape recordings made in earlier times or written materials collected years or even centuries ago may exist and may represent important sources of information, indeed, in some cases, *the only information* available.” Linguists have a professional responsibility to be aware of the existence of relevant legacy materials, and, when possible, work toward preserving them for and in collaboration with communities, for linguistics and related disciplines, and for posterity generally. This allows for the subsequent *philological* analysis of the documentary record, “recogniz[ing] the documentary filiation which characterizes all linguistic data as they are successively recorded, interpreted, and analyzed” (GODDARD, 1973, p. 727). At the same time, linguists have a responsibility to ensure that their engagement with legacy materials is carried out in an ethical fashion (O’MEARA and GOOD, 2010; see also §4 above).

The preservation of legacy materials involves certain key steps: locating the materials; appraising their physical condition; organizing them in a basic way; transporting them to an archive or other (temporary) safe location; carrying out any necessary conservation remediation; cataloging them; and, ideally, digitizing them (perhaps not necessarily in this order). Locating legacy materials is often one of the more difficult steps, requiring either inadvertently being made aware of their existence and location, or a good deal of sleuthing. In general, locating legacy materials is made easier by having a deep familiarity with the histories of research and other activities such as exploration in particular regions across many disciplines. More concretely, one should pay attention to descriptions, often in the methodology or similar section, of early publications (and unpublished works, when available).

For example, the late anthropologist Gerald Weiss (see O’HAGAN, 2021 for background), in a section of his PhD dissertation titled “Design and Method,” had the following to say about the output of his fieldwork in Ashaninka communities along the Tambo River of Peru from 1961 to 1964 (WEISS, 1969, p. 6).

> The techniques employed in the field for obtaining information were standard, adjusted only to the particular requirements of the information being gathered. A journal was kept; temperature, humidity and precipitation readings were recorded daily in camp; information obtained from observation and interrogation were accumulated on four-by-six slips or some other convenient form – in duplicate, with date, place and informant’s name indicated on each slip; specimens, photographs, and tape recordings were taken of everything possible.

Author O’Hagan was familiar with Weiss’s work in the context of his own fieldwork with speakers of the related Caquinte language, but in previous readings of his dissertation had skipped over the section “Design and Method.” From it, however, we learn of the existence – at least at a certain point in time – of field notes, index cards (cf. slips), biological specimens, photographs, and sound recordings. Gerald Weiss became a professor of anthropology at Florida Atlantic University; a Google search in March 2021 revealed that he was still affiliated with that institution as an emeritus professor. Furthermore, a phone call at the same time revealed that all the above-listed materials – and more from postdoctoral research – were kept at his home in Florida (see ANWAR, 2021 for more details).

The basic organization of legacy materials can be difficult if the objects in question (e.g., field notes, tapes) have little metadata associated with them, and if specialists who might be able to read or understand the spoken version of a particular language cannot be consulted. Additionally, some objects may be in a format that is not easily accessed, for example, as is often the case with the analog reel-to-reel tapes that were commonplace from the 1950s to the 1970s, an especially valuable period in the documentation of many languages. Depending on the specific situation, at this juncture it may be beneficial to collaborate with community members who can assist in the interpretation of the materials, or to first have them digitized, either by an external professional audio technician[12] or by an archive as part of a donation. Regardless, the main goal of this step of organization is to produce a basic inventory of objects (how many notebooks, tapes, etc.) before they are transported, so that there is a record that can be corroborated to ensure that all the materials arrive at their intended destination. If metadata is lacking, other descriptions of the physical objects can be given (e.g., “two tapes with red design on cover”).

The goal in the transportation of legacy materials is for the materials to be sent to a location that either (on a temporary basis) has relatively better storage conditions (e.g., one that is less humid) or facilitates the transfer of them to an archive, if not to an archive directly. The moment of transportation can be one of the most perilous for archival materials. Ideally, especially delicate materials like tapes can be transported personally, for example, in carry-on luggage.[13] For other materials, it is important to keep them dry (e.g., with plastic coverings) and separate from others that might damage them. If materials are to be shipped, make sure to use high-quality cardboard boxes (with ample tape and bubble wrap), and, when financially possible, opt for airmail over ground transportation, as this usually involves less repeated

[12] Older recording studios in urban areas often have the equipment and expertise to do such digitization.There are a number of software programs to digitize analog sound, such as Audacity or Sound Forge. If you cannot operate these and you have sufficient funding, you can consult a professional.

[13] At all costs, avoid exposure to magnets, which can severely damage analog tapes.

handling of boxes. In general, at this stage it is best to be in contact with an archive that can provide you with additional guidance. Archives will also be the best informed to advise you on cataloging and digitization. We emphasize that digitization of analog sound recordings and film is in most situations relatively urgent, as the original tape degrades over time, and in most situations the prior storage conditions of the materials are not at an archival standard. Humidity and water damage are especially common.

5.2. Utilizing legacy materials

Legacy materials often preserve the voices and knowledge of people who preceded other documentation projects by many decades, and who in some cases represent the last first-language speakers or semi-speakers. In such cases, legacy materials may be a key source of information for community efforts in language revitalization and in maintaining or reclaiming cultural heritage, as well as for scholars' efforts to understand the breadth and depth of human expression.

Once materials have been archived, many years may pass before they are utilized by others. For example, linguist Catherine Callaghan (1931-2019) made a series of recordings of Sarah Ballard speaking Bodega Miwok (Miwokan; California) in 1960, while the former was a graduate student in linguistics at the University of California, Berkeley (PhD, 1963). She donated these and other recordings to the California Language Archive[14] in 1979 (BALLARD and CALLAGHAN, n.d.). Forty years later, in 2019, linguist Andrew Cowell (Berkeley PhD in 1993) made time-aligned transcriptions of these recordings in ELAN, which he archived with the same repository (COWELL, 2019+).[15] The two collections are linked in the archive's digital catalog, and so can easily be related to each other when consulting one or the other. Similarly, linguist Gladwyn Kingsley Noble, Jr. (1923-1994) made recordings of speakers of Wapishana and Atorai (Arawak; Brazil, Guyana) during a single field trip to Guyana in 1965. After wending their way through the hands of different academics, the 13 tapes were donated to the California Language Archive around 2006 by Manjari Ohala (see GEORGE et al., 1965). The recordings of Atorai, formerly thought to no longer have any first-language speakers, were utilized for a preliminary phonological description of the language by O'Hagan (2018), which was elaborated on by E. Miller (2021) as part of an undergraduate honors thesis. In 2021, the

[14] At the time, the California Language Archive (CLA) was known as the Survey of California Indian Languages.
[15] In the model of incremental archiving, in 2021 Cowell added time-aligned transcriptions of Richard Applegate's recordings of Sarah Ballard (Ballard and Applegate 1974).

texts included in the recordings were translated by remaining speakers now resident in Wapishana communities, and may serve as a basis for a workshop dedicated to Atorai language later in 2022 (K. RYBKA, p.c.).

Finally, we stress that legacy materials can be utilized on an ongoing basis, and that usage requires interpretation – a creative process of understanding of "what those materials meant to their creators, what new meanings they might take on in the context in which they are being used, and what roles [contemporary actors] themselves as persons might play in the materials' circulation and reception" (DOBRIN and SCHWARTZ, 2021, p. 23). One especially productive example of ongoing work with legacy materials is the biennial Breath of Life Archival Institute for Indigenous California Languages (see GEHR, 2013) held at the University of California, Berkeley, organized by the Advocates for Indigenous California Language Survival (AICLS) in conjunction with the Survey of California and Other Indian Languages, which houses the California Language Archive. The Institute brings indigenous people to the Berkeley campus to collaborate with volunteer linguists in the interpretation of archival materials, many of which date back to the early 20th century. This model has been successfully expanded to other locales (BALDWIN, PÉREZ BÁEZ and HINTON, 2018) and it may be fruitful to consider whether a similar initiative in Brazil could increase the accessibility of existing language materials.

6. Case studies: Documentation and archiving of the Paresi-Haliti and Enawene Nawe languages

This section offers a pair of case studies illustrating language documentation and archiving in Brazil, through projects carried out in collaboration with the Paresi-Haliti and Enawene Nawe peoples and led by Ana Paula Brandão. After a brief introduction to these two languages, we describe the documentation projects (§6.1), and the resulting archival collections (§6.2).

Paresi (Glottocode pare1272, latitude -14.59 and longitude -57.41) is an indigenous language spoken by a people of the same name, whose population of approximately 3000 is distributed in several communities in the State of Mato Grosso, near the city of Cuiabá, on the tributaries of the Juruena River. The Enawene Nawe language (Glottolog code enaw1238, latitude -12.43 and longitude -58.98) is spoken by a smaller group of approximately 1000 people, who live in two communities (*Halataikwa* and *Kolinakwa*), in an Indigenous Territory located near the cities of Juína, in the State of Mato Grosso, and Vilhena, in the State of Rondônia.

Both languages belong to the Arawakan family (PAYNE, 1991; AIKHENVALD, 1999; RAMIREZ, 2001). Brandão, Carvalho and Pereira (2018) and Pereira (2018) present evidence that Enawene Nawe (hereinafter EN) and Paresi are very closely related, and together with the Saraveka language form a subgroup, which they term Juruena. In previous classifications of the Arawak family, only Fabre (2005) suggested a proximity between Paresi and Enenawe Nawe, whereas Payne (1991) grouped Paresi together with Waurá in a Central group, and Aikhenvald (1999) likewise classified Paresi together with the Xingu languages in a 'Paresi-Xingu' branch.

Considerable documentation and description now exists for the Paresi language. Ana Paula Brandão and Glauber Silva independently documented the language over several years. Among the main descriptive works for Paresi are Silva (2009, 2013) and Brandão (2010, 2014). Work with EN is quite recent, on the other hand; the only known descriptive works are Rezende (2003, 2013),[16] Brandão and Reis (2020), Reis (2020). In 2005, the project 'Sketch grammar, texts, and dictionary of Enawene Nawe (Arawak, Brazil)' was funded by the Endangered Languages Documentation Program (ELDP), but unfortunately the project was not completed. In 2019, Brandão received funding from the ELDP, through the Federal University of Pará, for the project 'Documentation of the Enawene Nawe language'.

6.1. The documentation projects

The Paresi language documentation project started in 2006 and was developed via Brandão's doctoral and postdoctoral research on the language. The EN language documentation project began in May 2019, with completion expected at the end of 2022. The goals of both projects were to organize a large corpus of audio and video recordings, spanning a variety of linguistic genres, and transcribed in the relevant indigenous language and translated into Portuguese.

The documentation of Paresi was initiated at the request of the Rio Formoso community, who were interested in recording their traditional culture. In the EN Halataikwa community, we also got in touch with a speaker of the language who invited us to visit them and present the project proposal to them. We obtained permission from the communities to record material

[16] Rezende (2013), a morphosyntactic description of the language, is not available to the academic community or to the community of speakers.

within a non-profit model. Communities benefited from the production of DVDs, CDs and USB drives for accessing traditional stories and songs, and from the training of indigenous teachers in linguistic documentation. The Paresi material also served as a basis for the elaboration of a reference grammar of the language, which was defended as a doctoral thesis (BRANDÃO, 2014).

The project participants included indigenous teachers and elders who were knowledgeable about traditional culture. Some teachers worked with recording, and others with transcription, translation, and metadata organization. One of the Paresi participants also worked with video editing. A key Paresi collaborator, Jurandir Zezokiware, participated in the documentation project with the EN community; his presence was very important in establishing a relationship of trust with the EN. He assisted during trainings and collected linguistic data. The project also involved UFPA undergraduate students, who visited the Paresi and Enawene Nawe communities to learn about fieldwork.

Both projects allowed for high quality equipment to be purchased for the Paresi and Enawene Nawe via financial support from the ELDP/SOAS. This equipment included a Zoom digital recorder, Shure head microphones, a digital video camera, an external microphone for the camcorder, a tripod, and other materials. Various cultural and speech events were documented throughout the projects. The Paresi communities selected traditional stories, songs, blessings, traditional festivals, indigenous games, etc. for documentation, and we also recorded different dialects or varieties of the Paresi language. The EN preferred to focus on recording traditional stories (e.g. their origin story, the account of the origin of cassava, a story about spirits, and others). More information about the two projects is provided in Brandão and Zezokiware (2018).

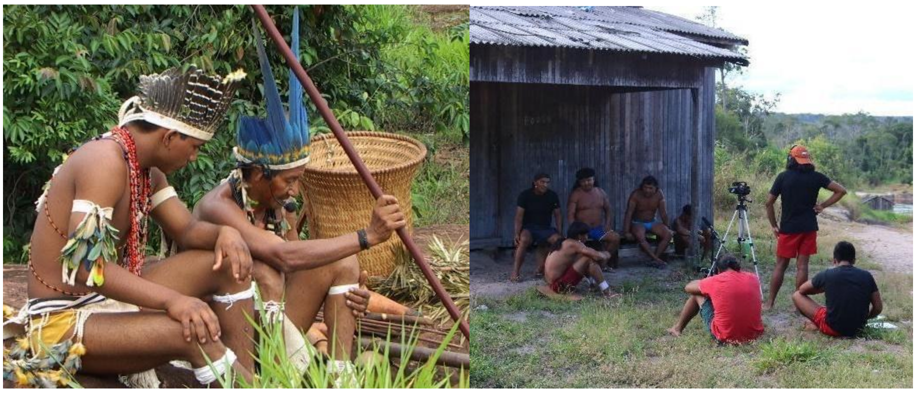


Figure 5. Paresi men wearing traditional clothing (PAB-200712-AP-RC-PontePedra80.JPG, photograph: Rose Costa); Enawene Nawe telling stories (UNK-20190500-AP-treinamentos-136.jpg, photograph: Ana Paula Brandão)

6.2. The Paresi and Enawene Nawe archival collections

The primary data – audio and video recordings – were recorded in .WAV and .MTS formats, respectively. Each session was recorded in audio and video. Later, the .MTS files were converted to .MP4 for archiving, to yield a more compact format for storage. The file names begin with the ISO code for the language (PAB for Paresi and UNK for EN), followed by the recording date (in YYYY/MM/DD format), the abbreviations of the names of the person who recorded and of the primary speaker, and the keyword of the session in which the file is included; for example: UNK-20200114-WE-YI-Kolito.MTS. These documentation projects generated digital collections involving more than 150 hours of Paresi recordings and 37 hours of Enawene Nawe recordings, which are organized into seven categories, as illustrated in Table 1.

Categories	Paresi	Enawene Nawe
Daily activities	9h	30 min
Ritual activities	13h	0
General elicitation	36h	30 min
Lexical elicitation	10h	4h
Music	15h	1h
Traditional stories	38h	24h
Non-traditional stories	30h	7h
Total	150h	37h

Table 1. Paresi and Enawene Nawe digital recording

The secondary data produced in these projects include annotations and metadata associated with the recordings. Annotations, consisting of transcriptions and translations into Portuguese, were made using the ELAN program (EUDICO Linguistic Annotator 2020) and Word. Lexicons for both languages were also developed, and several Paresi texts were interlinearized using the FLEx program (FieldWorks Language Explorer 2019) according to the analysis presented in the reference grammar. The transcriptions and translations of the texts were made by the speakers. Most of the transcriptions were produced during the period when the lead researcher was not in the communities and, in the case of Paresi, they were later reviewed with the speakers. Paresi speakers received training in the use of ELAN, while EN speakers had little interaction with this program and preferred to transcribe in Word using a computer or cell phone. We are still organizing the notes from the EN recordings, such that all material will be transferred to ELAN and can later be published in a digital collection.

Files in ELAN have at least three lines of annotations: the transcription, the translation, and the notes. More detailed information about the glosses of Paresi or EN morphemes (i.e. the interlinearization) is organized in the FLEx program, as it allows automatic insertion of glosses (once they are already entered in the database), unlike ELAN. Neither language has an established orthography that is used consistently in indigenous schools. Therefore, transcriptions were made in the orthographies that the speakers know and, in the case of Paresi, they were systematized according to the orthography that Brandão proposed for the language.

The metadata were initially organized in an Excel spreadsheet, then entered into the Lameta program (HALTON et al., 2021). Each session groups together an audio file, video file, and annotation in ELAN (.EAF format) or elicitations in PDF. IMDI files (ISLE Metadata Initiative) were also created for each session, and contain information about the recordings, such as the 'actors' (people involved), subject, content description, and keywords. There are also sets of photographs organized into the following categories: community, craft, school, daily life, researchers, training, people, and work.

All data from the Paresi project are stored in the archive of the Museu Paraense Emílio Goeldi, but are not yet available to the public. Twenty hours of material collected during the period of financial support from ELDP (2011-2012) were also stored in the archive associated with this institution, the Endangered Language Archive (ELAR).[17] Another thirty hours are stored at the Archive of the Indigenous Languages of the Americas (AILLA),[18] based at the University of Texas at Austin, with some materials available online. Part of the material collected in the Enawene Nawe project was organized in ELAR[19] and will also be made available in the Goeldi archive.

The archives referenced here employ a set of graded access codes, which indicate the different levels of access available for different files. ELAR has the following codes: a) O for free access, b) U for materials that can be accessed by creating an account in the collection; and c) S for materials that are restricted and can only be accessed with the depositor's permission. Files marked with the access code S may contain personal or sensitive information about the

[17] Paresi-Haliti collection at ELAR: http://hdl.handle.net/2196/00-0000-0000-000E-2C87-4

[18] Paresi-Haliti collection at AILLA: https://ailla.utexas.org/islandora/object/ailla:254756

[19] Enawene Nawe collection at ELAR: http://hdl.handle.net/2196/00-0000-0000-0012-D797-0

speakers, such as conversations and life stories, or may be currently under analysis by the depositor. Materials in the AILLA collection have similar access levels, indicated by numbers instead of letters. Some archives require users to complete a free registration on the website before any materials may be accessed.

These archives are available in English (ELAR) or in English and Spanish (AILLA), but they do not yet have access information in Portuguese,[20] which makes it more difficult for Paresi and Enenawe Nawe people to access them. In the future, we intend to make these collections available in regional digital archives near the communities, potentially housed in universities (e.g. the Federal University of Mato Grosso [UFMT]), intercultural colleges, and/or indigenous schools. Speakers could thereby have physical access to the server where the materials are stored, as well as easier access via the internet.

To conclude this section, the two documentation projects presented here have enabled the creation of Paresi and Enenawe Nawe archival collections, which are available both to the academic community and to indigenous community members. In the context of the Covid-19 virus pandemic, archival collections such as these have been particularly useful for researchers who were unable to visit indigenous communities. Importantly, these collections are also a way of safeguarding traditional knowledge, as has been made even more urgent through the loss of many indigenous elders during the Covid-19 pandemic. It is crucial that language documentation materials that have not yet been deposited in archives be archived to ensure their long-term preservation. An important next step is to make these collections increasingly accessible to indigenous peoples in Brazil. In this way, indigenous teachers will be able to use the materials in community classrooms and for academic work within indigenous colleges.

7. Conclusion

Brazil and its neighboring regions are home to some of the highest levels of linguistic diversity known around the world. Some 300 indigenous languages are spoken in South America, and more than half this number within Brazil (MOORE, 2007; GALUCIO, MOORE and VAN DER VOORT, 2018, p. 195) – a fact celebrated by UNESCO’s International Decade of Indigenous Languages, of which the inception coincides with the writing of this article. Yet this wealth of languages represents just a fraction of the number that must have existed in South America on the eve of European contact, and the processes of language shift and loss have

[20] However, a Portuguese language interface is planned for AILLA, to be available soon.

continued through the present day. Nearly 80 of the remaining 300 languages are now critically endangered (MOORE, 2007). The loss of indigenous languages in Brazil continues in spite of such initiatives as the Decreto n. 7.387 in 2010, which instituted the National Inventory of Linguistic Diversity (INDL), a government program to survey languages and declare them to be immaterial cultural patrimony (GALUCIO, MOORE and VAN DER VOORT, 2018). Even today, the great majority of Brazilian and other South American indigenous languages lack substantial description and documentation. Perhaps half have even a minimal record in archives.

As this article has explored, archiving is an essential component of language documentation. It is only through archiving, and via well-maintained digital repositories with adequate infrastructure and institutional commitment, that documentary materials are reliably preserved and made accessible for the long term. Fortunately, the number of robust documentation projects and archival collections has been growing rapidly in the last few decades, as exemplified by the Paresi and Enawene Nawe projects described in the final section of this article (see also MOORE and GALUCIO, 2016). But we must build our participation and investment in archiving if we are to maintain the outcomes of this documentary work. As we discuss here, this means contributing to, supporting, and expanding archiving initiatives on all levels – regional and national initiatives like the Museu Goeldi in Brazil, and archives with international scope such as AILLA. It also means supporting and expanding the possibilities for speaker and heritage communities to access these materials, both through, by, and in collaboration with archives. As linguists, we have a social and academic responsibility to archive the documentary materials that we produce, to support initiatives to archive legacy materials, to make materials as accessible as possible, and to work closely with communities to ensure an ethical process.

The COVID-19 pandemic has highlighted the urgency of these endeavors. Tragically, communities are losing elders, and languages are losing speakers. Documentary archives help to preserve their knowledge for the generations to come, and make it accessible to community members, scholars, and others into the future. Archives open new possibilities for research when fieldwork is impossible, offering alternative sources of data and analysis, and new pathways for investigation. And archives can provide key resources for communities who wish to revitalize, maintain, or simply remember their linguistic and cultural heritage.

ADDITIONAL INFORMATION

Competing interests

The authors declare that they have no competing interests.

Link to Preprint

https://preprints.scielo.org/index.php/scielo/preprint/view/4755

Research Preregistration and Standards

The authors confirm that they have reviewed the standards for research as registered in the Equator Network. The overview nature of this article does not constitute research on humans or animals. Since no primary research was done for this article, no preregistration in an independent, institutional registry was necessary. Prior ethics approvals were obtained for the studies referenced in this article, but no distinct ethics review for the work represented here was necessary.

Data Accessibility Statement

This article represents an overview study with no project-specific primary data. DOI links to relevant archival content have been included throughout the text where relevant.

Acknowledgements

We thank the organizers of *Abralin ao Vivo* for the opportunity to participate in their online forum (July 11, 2020). We would also like to thank the indigenous community members who have participated in the various projects cited here. While this study is an overview and did not receive project-specific funding, we are grateful to the entities that have generously supported the work that is reflected in the archival collections mentioned here. Rosés Labrada's doctoral project, reported on in Section 4, was funded by Canada's Social Sciences and Humanities Research Council (SSHRC) via a Vanier Canada Graduate Scholarship [Award No. 770-2012-0151]. Ana Paula Brandão's work on Paresi-Haliti and Enenawe-Nawe, reported on in Section 6, were funded by the National Science Foundation (BCS 1123943) and the Endangered Languages Documentation Programme (IGS0160, IPF0280).

References

AIKHENVALD, Alexandra. 1999. The Arawak language family. In: R. M. W. Dixon & A. Y. Aikhenvald (orgs.). **The Amazonian languages.** p. 65-106. Cambridge: Cambridge University Press.

ANWAR, Yasmin. Chance phone call keeps alive scholar's remarkable Amazonian legacy. In: **Berkeley News**, May 24, 2021. https://news.berkeley.edu/2021/05/24/chance-phone-call-keeps-alive-scholars-remarkable-amazonian-legacy/.

AUSTIN, Peter K. Language Documentation & Legacy Text Materials. **Asian and African Languages and Linguistics.** n. 11. p. 23-44. 2017.

BALDWIN, Daryl; PÉREZ BÁEZ, Gabriela; HINTON, Leanne. The Breath of Life Workshops and Institutes. In: HINTON, Leanne; HUSS, Leena; ROCHE, Gerald (orgs.). **The Routledge Handbook of Language Revitalization**, 1ed. , p. 188-196. London: Routledge, 2018.

BALLARD, Sarah; APPLEGATE, Richard B. **Bodega Miwok Sound Recordings 2014-05**, California Language Archive, Survey of California and Other Indian Languages, University of California, Berkeley, 1974. http://dx.doi.org/doi:10.7297/X2QN64XN.

BALLARD, Sarah; CALLAGHAN, Catherine A. n.d. **The Catherine A. Callaghan Collection of Bodega Miwok Sound Recordings, LA 6**, California Language Archive, Survey of California and Other Indian Languages, University of California, Berkeley, http://cla.berkeley.edu/collection/10086.

BARWICK, Linda. Turning It All Upside Down…Imagining a Distributed Digital Audiovisual Archive. **Literary and Linguistic Computing** Volume 19, N. 3, Setembro de 2004, p. 253–263. https://doi.org/10.1093/llc/19.3.253.

BARWICK, Linda; MARETT, Allan; WALSH, Michael; REID, Nicholas; FORD, Lysbeth. Communities of Interest: Issues in Establishing a Digital Resource on Murrinh-Patha Song at Wadeye (Port Keats), NT. **Literary and Linguistic Computing** Volume 20, N. 4, Novembro de 2005, p. 383–397. https://doi.org/10.1093/llc/fqi048.

BEREZ-KROEKER, Andrea L.; GAWNE, Lauren; KUNG, Susan Smythe; KELLY, Barbara F.; HESTON, Tyler; HOLTON, Gary; PULSIFER, Peter; BEAVER, David I.; CHELLIAH, Shobhana; DUBINSKY, Stanley; MEIER, Richard P.; THIEBERGER, Nick; RICE, Keren; WOODBURY, Anthony C. Reproducible research in linguistics: A position statement on data citation and attribution in our field. **Linguistics**. vol. 56, no. 1, 2018, pp. 1-18. https://doi.org/10.1515/ling-2017-0032.

BEREZ-KROEKER, Andrea L.; HENKE, Ryan. Language Archiving. In: REHG, Kenneth L.; CAMPBELL, Lyle (orgs.) **The Oxford Handbook of Endangered Languages**. pp. 347–69. Oxford: Oxford University Press, 2018. https://doi.org/10.1093/oxfordhb/9780190610029.013.18.

BISCHOFF, Shannon T.; JANY, Carmen (orgs.). **Insights from Practices in Community-Based Research: From Theory to Practice around the Globe.** Berlin & Boston: De Gruyter Mouton, 2018. https://doi.org/10.1515/9783110527018.

BOMFIM, Anari Braz. Patxohã: A Retomada Da Língua Do Povo Pataxó. **Revista LinguíStica**. Volume 13, n. 1 jan de 2017, p. 303-327. ISSN 2238-975X 1. https://revistas.ufrj.br/index.php/rl.

BRANDÃO, Ana Paula. **Verb morphology in Paresi-Haliti (Aruak)**. Qualifying paper, University of Texas at Austin, 2010.

BRANDÃO, Ana Paula. **A reference gramar of Paresi-haliti (Aruák)**. PhD dissertation, University of Texas at Austin, 2014.

BRANDÃO, Ana Paula; CARVALHO, Fernando; PEREIRA, Everton. Estudo histórico-comparativo preliminar do subgrupo Juruena (Aruák). Trabalho apresentado no **Congresso Internacional de Estudos Linguísticos e Literários na Amazônia** (VI CIELLA), Belém, 2018.

BRANDÃO, Ana Paula; REIS, Thainá de Lima. Gênero gramatical em Enawene Nawe? **Revista de Letras Norte@mento**. Dossiê Temático: Para a década das línguas indígenas, Sinop, v. 13, n. 33, p. 208-227, nov. 2020.

BRANDÃO, Ana Paula; ZEZOKIWARE, Jurandir. A documentação participativa: O caso das línguas Paresi e Enawene Nawe. **Revista Moara**. Edição 50. Agosto - Junho de 2018, Estudos Linguísticos. ISSN: 0104-0944

CARROLL, Stephanie Russo; GARBA, Ibrahim; FIGUEROA-RODRÍGUEZ, Oscar L.; HOLBROOK, Jarita; LOVETT, Raymond; MATERECHERA, Simeon; PARSONS, Mark; RASEROKA, Kay; RODRIGUEZ-LONEBEAR, Desi; ROWE, Robyn; SARA, Rodrigo; WALKER, Jennifer D.; ANDERSON, Jane; HUDSON, Maui. The care principles for Indigenous data governance. **Data Science Journal**. vol. 19. n. 1. p.1-12. 4 de novembro de 2020. https://doi.org/10.5334/dsj-2019-031.

CONZETT, Philipp; DE SMEDT, Koenradd. Guidance for citing linguistic data. In: BEREZ-KROEKER, Andrea L.; MCDONNELL, Bradley; KOLLER, Eve; COLLISTER, Lauren B. (orgs.). **The Open Handbook of Linguistic Data Management**. p. 143–155. Cambridge: The MIT Press, 2022. DOI: https://doi.org/10.7551/mitpress/12200.001.0001.

COWELL, Andrew. **Time-aligned Annotations of Bodega Miwok Sound Recordings**. **2019-18**. California Language Archive, Survey of California and Other Indian Languages, University of California, Berkeley. Disponível em: http://dx.doi.org/doi:10.7297/X2251GC0.

CZAYKOWSKA-HIGGINS, Ewa. Research Models, Community Engagement, and Linguistic Fieldwork: Reflections on Working within Canadian Indigenous Communities. **Language Documentation and Conservation**. Vol. 3. n. 1. p. 15-50. 2009.

DEBENPORT, Erin. The potential complexity of 'universal ownership': Cultural property, textual circulation, and linguistic fieldwork. **Language & Communication**. vol. 30, n. 3. p. 204-210. Julho de 2010.
DIGITAL PRESERVATION COALITION (DPC). Glossary. **Digital Preservation Handbook**. Ed. 2. 2015. Dispnível em: https://www.dpconline.org/handbook/glossary

DOBRIN, Lise M. From Linguistic Elicitation to Eliciting the Linguist: Lessons in Community Empowerment from Melanesia. **Languageb**. Vol. 84, n. 2. p.300-324. Junho de 2008.

DOBRIN, Lise M.; SCHWARTZ, Saul. The social lives of linguistic legacy materials. **Language Documentation and Description**. vol. 21. p. 1-36. 2021.

DWYER, Arienne M. Ethics and Practicalities of Cooperative Fieldwork and Analysis. In: GIPPERT, Jost; HIMMELMANN, Nikolaus P.; MOSEL, Ulrike (orgs.). **Essentials of Language Documentation**. p. 31-66. Berlin: Mouton de Gruyter, 2006.

ELAN (version 5.9). Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive. https://tla.mpi.nl/tools/tla-tools/elan/. 2020.

FABRE, Alain. **Diccionario etnolinguístico y guía bibliográfica de los pueblos indígenas sudamericanos, Arawak**. 2005. Disponível em:
http://www.ling.fi/Entradas%20diccionario/Dic=Arawak.pdf.

FIELDWORKS (FLEx) (version 8.3.12). 2019. SIL. https://software.sil.org/fieldworks/.

GALUCIO, Ana Vilacy; MOORE, Denny; VAN DER VOORT, Hein. 2018. O Patrimônio linguístico do Brasil: Novas perspectivas e abordagens no planejamento e gestão de uma política de diversidade linguística. **Revista do Patrimônio Histórico e Artístico Nacional**. n. 38. p. 194-219.

GASSER, Emily. The right to say yes: Language Documentation in West Papua. **Australian Journal of Linguistics**. vol. 37, n. 4. Ago 2017. p. 502-526. https://doi.org/10.1080/07268602.2017.1350131.

GEHR, Susan. **Breath of Life: Revitalizing California's Native Languages Through Archives**. MA thesis, San José State University, 2013. https://doi.org/10.31979/etd.yfva-e77q.

GEORGE, Christine; JOSEPH, Lawrence; XAVIER, Felix; NOBLE JR, Gladwyn K. **Linguistic Materials on Indigenous Languages of Guyana, 2018-03**, California Language Archive, Survey of California and Other Indian Languages, University of California, Berkeley, 1965. http://dx.doi.org/doi:10.7297/X2NC5ZCN.

GLENN, Akiemi. Five Dimensions of Collaboration: Toward a Critical Theory of Coordination and Interoperability in Language Documentation. **Language Documentation & Conservation**. vol. 3, n. 2. p. 149-160. Dez 2009.

GODDARD, Ives. Philological Approaches to the Study of North American Indian Languages: Documents and Documentation. In: SEBEO, Thomas A. (org.). **Current Trends in Linguistic:**, **Linguistics in North America**. vol. 10. p. 727-745. The Hague: Mouton, 1973.

GOOD, Jeff. The Scope of Linguistic Data. In: BEREZ-KROEKER, Andrea L.; MCDONNELL, Bradley; KOLLER, Eve; COLLISTER, Lauren B. (orgs.). **The Open Handbook of Linguistic Data Management**. p. 27-47. Cambridge, MA: The MIT Press, 2022.

GRINEVALD, Colette. Worrying about Ethics and Wondering about 'Informed Consent': Fieldwork from an Americanist Perspective. In: SAXENA, Anju; BORIN, Lars (orgs.). **Lesser-Known Languages of South Asia: Status and Policies, Case Studies and Applications of Information Technology**. p. 339-370. Berlin: De Gruyter Mouton, 2006. https://doi-org/10.1515/9783110197785.3.339.

HALTON, John; HOLTON, Gary; SEYFEDDINIPUR, Mandana; THIEBERGER, Nicholas. **Lameta [software]**. 202. Disponível em: https://github.com/onset/laMETA/releases.

HAN, Na-Rae. Transforming Data. In: BEREZ-KROEKER Andrea L. ; MCDONNELL, Bradley; KOLLER, Eve; COLLISTER, Lauren B. (orgs.) **The Open Handbook of Linguistic Data Management**. p. 73-87. Cambridge: The MIT Press, 2022. https://doi.org/10.7551/mitpress/12200.001.0001.

HENKE, Ryan; BEREZ-KROEKER, Andrea L. A Brief History of Archiving in Language Documentation, with an Annotated Bibliography. **Language Documentation & Conservation**. vol. 10. p. 411-457. 2016.

HIMMELMANN, Nikolaus P. Documentary and descriptive linguistics. **Linguistics**. vol. 36. p. 161-195. 1998.

HIMMELMANN, Nikolaus P. Language documentation: What is it and what is it good for? In: GIPPERT, Jost; HIMMELMANN, Nikolaus P.; MOSEL, Ulrike (orgs.). **Essentials of Language Documentation**. p. 1-30. Berlin: Mouton de Gruyter, 2006.

HIMMELMANN, Nikolaus P. Linguistic Data Types and the Interface between Language Documentation and Description. **Language Documentation & Conservation**. vol. 6. p. 187-207. 2012.

HOLTON, Gary. Relatively Ethical: A Comparison of Linguistic Research Paradigms in Alaska and Indonesia. **Language Documentation & Conservation**. vol. 3, n. 2. p. 161-75. 2009.

HOLTON, Gary; LEONARD, Wesley Y.; PULSIFER, Peter L. Indigenous Peoples, Ethics, and Linguistic Data. In: BEREZ-KROEKER Andrea L.; MCDONNELL, Bradley; KOLLER, Eve; COLLISTER, Lauren B. (orgs.). **The Open Handbook of Linguistic Data Management**. p. 49-60. Cambridge: The MIT Press, 2022. https://doi.org/10.7551/mitpress/12200.001.0001.

INNES, Pamela. Ethical Problems in Archival Research: Beyond Accessibility. **Language & Communication**. vol. 30, n. 3. p. 198–203. Jul 2010. https://doi.org/10.1016/j.langcom.2009.11.006.

JOHNSON, Heidi. Language documentation and archiving, or how to build a better corpus. In: AUSTIN, Peter K. (org.). **Language Documentation and Description**. vol. 2. p. 140-153. London: SOAS, 2004.

KAPLAN, Judith; LEMOV, Rebecca. 2019. Archiving Endangerment, Endangered Archives: Journeys through the Sound Archives of Americanist Anthropology and Linguistics, 1911–2016. **Technology and Culture.** 60 (2): S161–87. https://doi.org/10.1353/tech.2019.0067.

KUNG, Susan Smythe; SULLIVANT, Ryan; POJMAN, Elena; NIWAGABA; Alicia. **Archiving for the Future: Simple Steps for Archiving Language Documentation Collections [OER]**. 2020. Disponível em: https://archivingforthefuture.teachable.com/. CC BY-SA 4.0 international license.

KUNG, Susan Smythe. Data archiving, access, and repatriation. In: STANLAW, James (org.). **The International Encyclopedia of Linguistic Anthropology**. Wiley Online Library, 2020. https://doi.org/10.1002/9781118786093.iela0430.

LEONARD, Wesley Y.; HAYNES, Erin. Making 'Collaboration' Collaborative: An Examination of Perspectives That Frame Linguistic Field Research. **Language Documentation and Conservation**. vol. 4. p. 268-293. 2010.

LUKANIEC, Megan. Managing Data from Archival Documentation for Language Reclamation. In: BEREZ-KROEKER, Andrea L.; MCDONNELL, Bradley; KOLLER, Eve; COLLISTER, Lauren B. (orgs.). **The Open Handbook of Linguistic Data Management**. p. 315-325. Cambridge: The MIT Press, 2022. https://doi.org/10.7551/mitpress/12200.001.0001.

MACRI, Martha J.; SARMENTO; James. Respecting Privacy: Ethical and Pragmatic Considerations. **Language & Communication**. Vol. 30, n. 3. p. 92–97. Jul 2010. https://doi.org/10.1016/j.langcom.2009.11.005.

MILLER, Ellis R. 2021. **Phylogenetic Classification of the Negro-Roraima Subgroup**. BA thesis, University of California, Berkeley.

MILLER, Robert J. Introduction. In: LINK, Adrianna; SHELTON, Abigail; SPERO, Patrick (orgs.). **Indigenous Languages and the Promise of Archives**. p.1-24. Lincoln: University of Nebraska Press, 2021.

MOORE, Denny. Endangered languages of lowland tropical South America. In: BRENZINGER, Matthias (org.). **Language Diversity Endangered**. p. 29-58. Berlin: Mouton de Gruyter, 2007.

MOORE, Denny; GALUCIO, Ana Vilacy. Perspectives for the documentation of indigenous languages in Brazil. In: PÉREZ BÁEZ, Gabriela; ROGERS, Chris; ROSÉS LABRADA, Jorge Emilio (orgs). **Language Documentation and Revitalization in Latin American Contexts**. p. 29-58. Berlin & Boston: De Gruyter Mouton, 2016.

MOORE, Denny; GALUCIO, Ana Vilacy; GABAS JR, Nilson.. Desafio de documentar e preservar as línguas Amazônicas. **Scientific American Brasil**. n. 3. p. 36-43. 2008.

NATHAN, David. **Progressive archiving: Theoretical and practical implications for documentary linguistics**. Presentation at the International Conference on Language Documentation and Conservation, March 3, 2013. http://hdl.handle.net/10125/26115.

O'HAGAN, Zachary. **A Phonological Sketch of Atorai (Arawak, Guyana) Based on Unique Recordings**. Presentation at Fieldwork Forum, Berkeley, May 2 2018.

O'HAGAN, Zachary.. Obituario, Gerald Weiss (1932-2021). **Amazonía peruana** n. 34. p. 279-286. Jul 2021.

O'MEARA, Caroline; GONZÁLEZ GUADARRAMA, Octavio Alonso. Accessibility to Results and Primary Data of Research on Indigenous Languages of Mexico. In: PÉREZ BÁEZ, Gabriela; ROGERS, Chris; ROSÉS LABRADA, Jorge Emilio (orgs.). **Language Documentation and Revitalization in Latin American Contexts**. p. 59-80. Berlin & Boston: De Gruyter Mouton, 2016. https://doi.org/10.1515/9783110428902.

O'MEARA, Carolyn; GOOD, Jeff. Ethical issues in Legacy Language Resources. **Language & Communication**. Vol. 30, n. 3. p. 162-170. July 2010.

OWENS, Trevor. **The Theory and Craft of Digital Preservation**. Baltimore: John Hopkins University Press, 2018.

PARKER, Verdena. **The Verdena Parker Collection of Hupa Sound Recordings and Films, LA 256**. California Language Archive, Survey of California and Other Indian Languages, University of California, Berkeley, s.d. http://dx.doi.org/doi:10.7297/X29S1PBD.

PARKER, Verdena; CAMPBELL, Amy; ESCAMILLA, Ramón; NEWBOLD, Lindsey; SPENCE, Justin. **Materials of the Hupa Language Documentation Project, 2017-06**. California Language Archive, Survey of California and Other Indian Languages, University of California, Berkeley, 2005+. http://dx.doi.org/doi:10.7297/X22R3Q2G.

PAYNE, David L. A classification of Maipuran (Arawakan) languages based on shared lexical retentions. In: DERBYSHIRE, Desmond; PULLUM, Geoffrey (orgs.). **Handbook of Amazonian Languages**. vol. 3. p. 355-499. Berlin: Mouton de Gruyter, 1991.

PEREIRA, Everton. **Estudo histórico-comparativo preliminar das línguas Paresi e Enawene Nawe**. Trabalho de Conclusão de Curso. Universidade Federal do Pará, 2018.

PÉREZ BÁEZ, Gabriela (photographer). **Kaufman Collection acquisition trip photos** (DSC_0001.JPG). Archiving the Terrence Kaufman Collection. The Archive of the Indigenous Languages of Latin America, ailla.utexas.org. Access: Public. PID ailla:257492.

PÉREZ BÁEZ, Gabriela (photographer). **Kaufman Collection acquisition trip photos** (DSC_0052.JPG). Archiving the Terrence Kaufman Collection. The Archive of the Indigenous Languages of Latin America, ailla.utexas.org. Access: Public. PID ailla:257543.

PÉREZ BÁEZ, Gabriela; ROGERS, Chris; ROSÉS LABRADA, Jorge Emilio. **Language Documentation and Revitalization in Latin American Contexts**. Berlin & Boston: De Gruyter Mouton, 2016. https://doi.org/10.1515/9783110428902.

RAMIREZ, Henri. **Línguas Arawak da Amazônia setentrional: Comparação e descrição**. Manaus: Editora da Universidade do Amazonas, 2001.

REIS, Thainá de Lima. **Uma Análise Preliminar do Gênero Gramatical em Enawene Nawe (Aruák)**. Trabalho de Conclusão de Curso. Universidade Federal do Pará, 2020.

REZENDE, Ubiray. **Fonética e fonologia da língua Enawene-Nawe (Aruak): Uma primeira abordagem**. Dissertação de Mestrado. Universidade Federal do Rio de Janeiro, 2003.

REZENDE, Ubiray. **Aspectos da gramática da língua Enawene-Nawe (Aruak)**. Tese de Doutorado, Universidade Federal do Rio de Janeiro, 2013.

RICE, Alexander. Using YouTube as the Primary Transcription and Translation Platform for Remote Corpus Work. **Language Documentation & Conservation.** Vol. 15. p. 514-550. 2021.

ROBINSON, Laura C. Archiving directly from the field. In: BARWICK, Linda; THIEBERGER, Nicholas (orgs.). **Sustainable data from digital fieldwork**. p. 23-32. Sydney: University of Sydney Press, 2006.

ROBINSON, Laura C. 2010. Informed Consent among Analog People in a Digital World. **Language & Communication**. vol. 30, n. 3. Jul 2010. p. 186-191. https://doi.org/10.1016/j.langcom.2009.11.002.

ROSÉS LABRADA, Jorge Emilio. **The Mako Language: Vitality, Grammar and Classification**. PhD dissertation, University of Western Ontario & Université Lumière-Lyon, 2015. https://ir.lib.uwo.ca/etd/2851.

DOS SANTOS, Wesley. **Kawahiva Language Documentation Archive, 2019-06**. California Language Archive, Survey of California and Other Indian Languages, University of California, Berkeley, 2017+ http://dx.doi.org/doi:10.7297/X2P26W9H.

SAPIÉN, Racquel-María. 2018. Design and Implementation of Collaborative Language Documentation Projects. In: REHG, Kenneth L.; CAMPBELL, Lyle (orgs.). **The Oxford Handbook of Endangered Languages**. pp. 203-224. Oxford & New York: Oxford University Press, 2018.

SEIFART, Frank; DRUDE, Sebastian; FRANCHETTO, Bruna; GASCHÉ, Jürg; GOLLUSCIO, Lucía; MANRIQUE, Elizabeth. Language Documentation and Archives in South America. **Language Documentation and Conservation.** vol. 2. n. 1. Junho de 2008. p. 130-140.

SILVA, Glauber. **Fonologia da lingua Paresi-Haliti (Aruak)**. Dissertação de Mestrado, Universidade Federal do Rio de Janeiro, 2009.

SILVA, Glauber Romling da. **Morfossintaxe da língua Paresi-Haliti**. Tese de Doutorado, Universidade Federal do Rio de Janeiro, 2013.

SPENCE, Justin. Learning Languages through Archives. In: HINTON, Leanne; HUSS, Leena; ROCHE, Gerald (orgs.). **The Routledge Handbook of Language Revitalization**. p. 179-187. New York: Routledge, 2018.

THIEBERGER, Nicholas. Lost and Found: Linguists and musicologists at three Australian universities are working together to preserve rare recordings and make them accessible to communities across the Pacific and beyond. **The ACU Review**. Publicado em 06 de dezembro de 2019. Disponível em: https://www.acu.ac.uk/the-acu-review/lost-and-found/.

THIEBERGER, Nicholas; BEREZ, Andrea L. Linguistic data management. In: THIEBERGER, Nicholas (org.). **The Oxford Handbook of Linguistic Fieldwork**. p. 90-118. Oxford: Oxford University Press, 2012.

United Nations General Assembly. 2007. **United Nations Declaration on the Rights of Indigenous Peoples: Resolution adopted by the General Assembly**, 2 October 2007, A/RES/61/295. https://www.un.org/esa/socdev/unpfii/documents/DRIPS_en.pdf (English), https://www.un.org/esa/socdev/unpfii/documents/DRIPS_pt.pdf (Portuguese).

VAPNARSKY, Valentina. Circulation et virtualités des savoirs amérindiens à l'ère du numérique – From home base to database…and back? The circulation and virtualities of Amerindian knowledge in the digital era. **Journal de la Société des américanistes** Vol. 106, n. 2. p. 79-104. 2020.

WEISS, Gerald. **The Cosmology of the Campa Indians of Eastern Peru**. PhD dissertation, University of Michigan, 1969.

WILBUR, Joshua. Archiving for the Community: Engaging Local Archives in Language Documentation Projects. In: NATHAN, David; AUSTIN, Peter K. (orgs.). **Language**

Documentation and Description. Volume 12: Special Issue on Language Documentation and Archiving. pp. 85-102. London: SOAS, 2014.

WOODBURY, Anthony C. Defining documentary linguistics. In: AUSTIN, Peter K. (org.). **Language Documentation and Description**, vol 1. p. 35-51. London: SOAS, 2003.

YAMADA, Racquel María. Collaborative Linguistic Fieldwork: Practical Application of the Empowerment Model. **Language Documentation & Conservation**. vol. 1, n. 2. p. 257-282. Dezembro de 2007.