LANGUAGE DOCUMENTATION & CONSERVATION

Vol. 14 (2020), pp. 520–578
http://nflrc.hawaii.edu/ldc
http://hdl.handle.net/10125/24949
Revised Version Received: 13 Jul 2020

Archival description for language documentation collections

Ryan Sullivant
University of Texas at Austin

Users of digital language archives face a number of barriers when trying to discover and reuse the materials preserved in the digital collections created by current language documentation projects. These barriers include sparse descriptive metadata throughout many collections and the prevalence of audio-video materials that are impervious to text-based search. Users could more easily evaluate, navigate, and use such a collection if it contained a guide that contextualized it, summarized its contents, and helped users identify and locate items within it. This article will discuss the importance of thorough collection descriptions and finding aids by synthesizing guidelines and best practices for archival description created for traditional archives and adapting these to the structure and makeup of today's digital language documentation collections. To facilitate the iterative description of growing collections, the checklist of information to include is presented in three groups of descending priority.

1. Introduction[1] Advances in portable recording technology have made it easier than ever to produce a large volume of media in the course of scientific research, including language documentation projects, and more and more of this media is sent to data repositories for preservation. The proliferation of archived language data has paradoxically made it more difficult for some users to find relevant materials or easily evaluate collections. Each new digital deposit makes finding items through searching and browsing more difficult, as the scale of today's archives and collections now means that keyword searches may return hundreds or even thousands of hits for users to evaluate. At the same time, this digital deluge has increased repository backlogs, and as a result, some data repositories and digital language archives have begun to involve data creators more and more in the archiving process. While the presence of metadata throughout a collection can aid some kinds of discovery, users could more easily find and evaluate materials if they were able to consult some kind of summary document, but making this document (or tracking the information that would be contained in it) is another task that a data creator might not be prepared for. This paper aims to serve as a guide and checklist for the kinds of descriptive and contextual information that should be recorded and included in a description of a language

[1]This material is based upon work supported by the National Science Foundation under Grant No. BCS-1653380. Susan Smythe Kung, Jennifer Isasi, and May Helena Plumb all gave comments and advice that has improved this paper.

E-ISSN 1934-5275

data collection. It is hoped that the resulting guide will not be overly fitted to a repository's current software configuration. This is important since digital archives will periodically migrate their assets from one repository system to another, sometimes leading to changes in the appearance of the website, how assets are displayed, and how the site is searched and navigated.

The intended audience of this paper is anyone who plans to archive language data in a digital repository, as well as anyone who has archived or is the steward of a language data collection who may want to improve its visibility and reusability. This paper may also be useful to anyone who is preparing a review of other people's collections or is evaluating the completeness of a collection guide.

The remainder of this section will clarify some of the terms that will be used in this paper and discuss the challenges of using navigation and search features to discover language data. §2 will discuss archival description of digital language collections. §3 will present a checklist of information to include in a collection description divided into three categories of descending importance. A collection guide created using these guidelines can be found in Appendix A. §4 will conclude the paper.

Since the terminology around digital language archiving varies from repository to repository, a few terms should be clarified. *Data* refers to materials collected or created in the course of research. Language archives accept a broad range of materials, such as photographs, audio-visual recordings, and research reports that are not "data" in its narrowest sense of collections of observations or measurements, but are considered to be data for present purposes. A *language archive* is a repository that holds, preserves, and makes accessible data about languages and the cultures that speak them. A *digital archive* is a repository that provides access to its holdings over the internet. Most digital language archives have a multilevel structure in which data files are organized into digital *folders* which are grouped together to form *collections*. Collections can take many forms. The ARC Centre of Excellence for the Dynamics of Language's distinction between uncurated *assemblages* of data, heterogenous *language documentation collections*, and highly structured *linguistic corpora*, will be used throughout this paper.[2] Files, folders, and collections all exist as *digital objects* in most repository software. A user can navigate to a digital object and read the *catalog record* that displays the object's *metadata*. Metadata is information about an object. *Descriptive metadata* allows users to find and locate an object, provides contextual details about it, its production, its relationship to other objects, etc. *Technical metadata* provides information about the form, size, and specifications of an object. *Structural metadata* indicates where an object is located within a sequence, hierarchy, or file structure. *Preservation metadata* includes information about the preservation status of an object and information like checksums used for fixity tests to assure that files have not been corrupted. *Rights metadata* includes information about an object's copyright status, who holds these rights, and any relevant licenses (Riley 2017). In most digital language archives, only some kinds of metadata (usually descriptive and technical) are visible for all digital objects. Other kinds of metadata, especially preservation metadata, are stored but are not distributed or published, while others may

[2]Thieberger (2018) credits Jane Simpson for the formulation of this division.

be implicit (such as structural information about what folder or collection contains a given object). Throughout this paper, language archive staff will be referred to as *archivists*, individuals who deliver materials to the archive (and who likely collected and/or created them) will be called *depositors*, and all people who search or view a digital language archive's website to view, access, or download the repository's digital objects will be called *users*.

1.1 Discovery and navigation through metadata records Unless they are following a direct link from somewhere else, users discover materials in a digital collection through browsing and searching. A user browses by navigating a website's interface and reviewing the titles and metadata displayed in catalog records. A user searches by entering text in a search box on a website and then is offered a list of objects which all contain that text somewhere in their catalog records (and within text-based files if the system offers full-text search). The user will then browse this list, perhaps refining the search results by filtering or modifying their search text, adding boolean operators, etc. Thus, while most users are likely to begin looking for data with an internet or catalog search, most searches will nevertheless involve some browsing. Both discovery strategies have their advantages and challenges. Browsing can be difficult or time consuming in a collection with many subparts, especially if they are labeled in ways that do not indicate their contents. Searching may not return all the items a user may be after. Search tools rely on the presence of metadata records, but the presence of rich metadata at different levels of a collection is beneficial to some kinds of search and discovery and not others. Metadata in an object's catalog record can be very useful for users who are able to search a repository knowing (or guessing) what terms will make for a successful search, but will not help a user who does not know the right terms. Furthermore, a search term may not appear in the catalog record or in the object itself, as when Boyd (2013:100) notes that "a narrator might describe living under segregation for several hours, without ever using the word 'segregation'". This item-level metadata, however, will not be visible to a user browsing a collection. This uneven distribution of metadata can be seen through a catalog search of the Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC),[3] which allows users to search either collection catalog records or folder catalog records. On April 9, 2020, a search for "weaving" across the collection metadata records finds only two results, whereas the same search over folder metadata finds matches in 67 folders across 30 different collections.

Depositors have been advised to produce metadata for all objects in their collections for about as long as there has been literature about language archiving (Johnson 2004; Nathan & Austin 2004). Depositors have also been advised to create a metadata document or guide that will orient an archive user to a collection as a whole and allow for easy summarization of its contents (Conathan 2011; Austin 2013). Groups like the Linguistics Data Interest Group of the Research Data Alliance[4] are now offering guidance on citation practices for the discipline (Berez-Kroeker et al. 2018;

[3]http://www.paradisec.org.au/.

[4]https://www.rd-alliance.org/groups/linguistics-data-ig.

Andreassen et al. 2019) in line with and informed by international standards and principles such as the FORCE11 manifesto about improving scholarly communication (Bourne et al. 2011), and the FAIR Guiding Principles for machine-readable interoperable data management (Wilkinson et al. 2016). Still, in many disciplines, depositor-curated collections often fail to include complete metadata (Koshoffer et al. 2019), and many datasets even fail to include enough metadata to make them findable and accessible (Grant et al. 2019).

In some repositories, searches take place not over just the metadata listed in catalog records, but over all machine-readable text within the files themselves. This full-text search can greatly improve discovery and reuse of digital collections (Milligan 2013), and has already transformed disciplines like history which regularly interrogate these kinds of materials (Putnam 2016). Most digital language collections, however, will not benefit from these technologies given the predominance of audio-visual media (which lacks text) and the painstakingly slow process of producing transcriptions and translations for them (Jung & Himmelmann 2011:201).[5] Because of the difficulty of producing annotations, digital language collections often do not have any machine-readable transcriptions – let alone translations – for large portions of their audio-visual holdings. For example, at the Endangered Languages Archive (ELAR)[6] and the Archive of the Indigenous Languages of Latin America (AILLA)[7] (two repositories where I was able to straightforwardly determine the presence of annotations from public metadata), between 60% and 73% of audio-visual recordings do not have machine-readable annotations of any sort.[8] These large amounts of media are only discoverable through their metadata records, and are likely to remain that way.

1.2 Advantages of collection descriptions Instead of relying on searching and browsing, a user can find their way into a collection by consulting a collection description or finding aid. Collection descriptions help users find materials within a collection, evaluate its suitability for their purposes, and understand the intellectual and historical context in which it was created. Digital language archiving – as well as data archiving endeavors in other disciplines – is becoming a participatory practice, meaning that it is more often the depositor and not an archivist who curates and describes the materials in the collection, and sometimes even adds the objects to the collections themselves (Shilton & Srinivasan 2007; Huvila 2008; Theimer 2011; Linn 2014). This participatory turn alleviates pressure on archive staff due to the digital deluge driven by ever-expanding born-digital collections and archiving mandates, and gives

[5]Data creators working on some high-resource languages can speed up transcription with automatic speech recognition tools, but for many languages (and nearly all languages that are the focus of language documentation projects), these tools must be developed. While there have been some successes (see e.g., Ćavar et al. 2016; Maldonado et al. 2016), the transcription produced by these tools still has a way to go (Adams et al. 2018).

[6]https://elar.soas.ac.uk/.

[7]https://ailla.utexas.org/.

[8]At the time of writing, approximately 73% of folders containing audio-visual media in AILLA lacked a text-based file identified as an annotation file ("annotation", "transcription", "translation", etc.). A similar study of materials held at ELAR found that approximately 60% of folders containing audio-visual media did not also contain a text-based annotation. See Appendix B for data and interpretation.

depositors an opportunity to leverage their deep familiarity with their materials and knowledge of their context to produce high-quality descriptions. In fact, depositors have distinct advantages over archivists when it comes to writing certain aspects of collection descriptions, and it has been recognized by users that "many aspects of the recording situation are known to the depositor(s) […] [but] do not enter into the metadata for the archive" (Whalen & McDonough 2019:53).

Digital language collections and their metadata are replete with languages understood by relatively small numbers of people. This poses difficulties for archivists who may not understand a language well enough to understand titles, classify content, or identify names of individuals and places mentioned in the materials. Identifying and distinguishing languages can also be a challenge in multilingual collections. Even metadata recorded in a project's working language (often a relevant national or regional language) may prove difficult for archivists in the predominantly anglophone world of digital archiving. Conversely, since most metadata and the digital archives themselves are typically only available in one or two languages (usually just in English), collection guides written in a language relevant for a community can serve as an alternative portal into a collection, side-stepping potential navigation and discovery issues. Users do prefer non-English finding aids for some collections; Lule (2019) notes that the Spanish-language finding aid for the Gabriel García Márquez Collection at the Harry Ransom Center is accessed four times as often as its English-language counterpart.[9]

Recognizing and identifying the rich ethnographic content of many digital language collections also requires specialized area knowledge. Archivists, even those at language repositories with a regional scope, may not be able to easily and appropriately recognize cultural content and know the relevant cultural protocols for handling and accessing materials. Depositors on the other hand, can leverage their own cultural and contextual familiarity with the content of their collections and the peoples represented within them to highlight culturally relevant themes, genres, and practices, and help inform archivists of relevant cultural protocols.

2. Existing models for language collection description Standalone documents summarizing and describing digital language collections are uncommon. Some digital language archives that grew out of physical collections such as the Center for Native American and Indigenous Research at the American Philosophical Society[10] and the Alaska Native Language Archive at the University of Alaska Fairbanks,[11] do have finding aids that resemble those in physical archives, with notes about the scope and

[9]A reviewer points out that written guides will not meaningfully increase access to collection materials for communities that are not literate in any of their languages. However, making unwritten collection guides is outside the scope of this paper, and I know of no collection guide available in any unwritten modality, and it is unclear to me how such a guide could serve to lead users to items within a collection. About 50 ELAR collections have a 'showreel' video embedded on the collection's webpage, but the ones I have seen serve as exhibits for a few video recordings within the collection and do not describe the entirety of the collection's contents or the contexts of its creation.

[10]https://www.amphilsoc.org/library/CNAIR.

[11]https://www.uaf.edu/anla/.

content of collections and container lists with hyperlinks to objects that function as portals to the collections. More commonly, some descriptive information is found within the collection's catalog record, and other information is scattered across the catalog records of folder and file objects.

One type of model for a digital language collection description is the genre of corpora introduction and discussion articles in journals such as the *International Journal of Corpus Linguistics*. These descriptive articles are shaped by the kinds of collections they describe and their perceived audience of future users. One recent example is Love et al. (2017) which motivates the need for and describes the creation of an updated corpus of spoken English. Love and colleagues detail the structure of the corpus and its annotations as well as the demographic speaker metadata they tracked. Such a description works well for collections like these, but the typical language documentation collection differs significantly from a highly structured linguistic corpus in two key ways that ought to be reflected in their descriptions: participant identifiability and granularity of data access and reuse.

Speakers and signers are typically de-identified in corpora (as much as is possible), whereas many language documentation collections recognize and name individual participants and attribute their contributions to them. This has consequences for reuse: if people are identified, they can be discovered by users (including community or family members interested in their recordings), and can become the subjects of research themselves. Thus, participants other than the depositors should be included in collection descriptions, assuming there are no concerns regarding privacy, consent, or participant safety.

The other key difference between corpora and collections is the granularity of data access and reuse. Corpora are generally designed and built to be consulted and reused in aggregate: users are unlikely to listen to, for example, individual telephone switchboard recordings (Godfrey et al. 1992) to hear a particular participant's response to a conversation prompt. In contrast, the heterogeneous and opportunistically collected materials of language documentation collections can and often are consulted individually for their content or for non-linguistic research (Holton 2012). A language learner may want to listen to or read a particular version of a folktale, or someone studying the works of a renowned poet or artisan may find their recorded contributions to a language documentation collection to be relevant. Thus, folder-level, or even file-level, descriptive and contextual metadata is far more important for language documentation collections than for linguistic corpora.

The heterogeneity of language documentation collections also means that the access conditions of materials may vary throughout a collection. While linguistic corpora may have access and reuse conditions on the entirety of their data (e.g., Oard et al. 2015), they are unlikely to have the more granular access conditions that many digital language collections require. Unfortunately, in most digital language archives, information about access restrictions typically resides in folder- or file-level metadata records and is not readily visible to a user browsing a collection or reviewing search results. A collection description could use a number of different strategies to distinguish between public and restricted objects in a collection.

Other models for collection descriptions come from the genre of collection review articles in *Language Documentation & Conservation (LD&C)*. As of this writing, seven collection descriptions have been published (Schembri et al. 2013; Salffner 2015; Caballero 2017; Gawne 2018; Oez 2018; Franjieh 2019; Hildebrandt et al. 2019). Much like the corpus reviews discussed earlier, these article-length descriptions detail the intellectual motivations, structure, and production of the collections, and provide summary information about them. They are all distinct from each other in structure and content, and their usefulness for locating materials within each collection varies. It is hoped that guidelines such as those offered in §3 below will help depositors create thorough and complete guides to their collections and help reviewers evaluate collection descriptions (Thieberger et al. 2016). Both linguistic corpora and language collection description articles have appeared as peer-reviewed journal articles, but it is worth noting that a good and useful collection description need not be peer reviewed. Most of the information in a collection description ought to be present in the collection itself, and this information can be iteratively improved as the collection – and the depositor's understanding of it – grows.

The following section will offer a checklist of information to include in a collection description, with examples of the kinds of information that could be considered for each topic. Based on recommendations and guidelines for archival description practices and the creation of archival finding aids (Library of Congress 2008; Society of American Archivists 2013), the categories of information in this checklist have been sorted into three levels of priority: (i) *required information* that amounts to a minimum description, (ii) *recommended information* that helps the user understand the context of the collection's production and find materials within it, and (iii) *optional information* that may more deeply inform the user about the collection and materials related to it. Given the recent turn towards participatory archiving practices mentioned earlier, this checklist and the guidelines within it have been modified and adjusted for use by depositors rather than archivists. As such, most preoccupations about precisely how metadata should be formatted, the order of presentation information in the guide, what authority lists to use, and so forth, have been set aside, and no particular description format or template is mandated.

3. A checklist for describing digital language collections Digital language collections are *multilevel* since they consist of folders that contain media files, and metadata should be present in catalog records at all levels (collection, folder, and media file). Archival description can either be multilevel, describing each subpart of the collection in turn, or can simply describe the entire collection as a whole. This checklist focuses on the production of a guide that describes and summarizes the collection as a whole. Nevertheless, depositors should point out discrete subparts within their materials and include relevant information in folder and file catalog records.

Archival description is iterative, and descriptions should be amended and improved upon as data are added to the collection, as time allows, and as one's understanding of a collection becomes clearer. It is hoped that the first priority information below can be included during or soon after the initial curation of the collection and

updated following any major additions or rearrangements of materials. Subsequent descriptions can then be expanded by adding information from the second and possibly third priority categories below.

A collection guide created from these guidelines can be found in Appendix A. The reader may wish to consult the example collection guide as they read the checklist guidelines below.

3.1 First priority collection metadata The most important collection metadata is information used to discover a collection and materials within it and describe any limitations on users' access and use of the data. Much of this *required information* may already be visible on the public catalog record of the collection or repository web pages detailing collection policies. However, some of this information – especially information regarding access restrictions – is typically only found in lower-level folder and file metadata records that a user will not readily see while browsing the collection, making discovery and evaluation of the accessible portions of the collection tedious.

3.1.1 Basic collection information The *basic collection information* includes information needed to locate and cite the collection, such as any codes used to identify the collection, information needed to locate the repository, the title and date(s) of the collection, and the names and roles of its creators.

3.1.1.1 Collection identifier An *identifier* is a code used to find a collection within the repository where it is held. Identifiers should be persistent and unlikely to change in the future. For physical collections, this code will often be called an accession number. For digital collections, the names and forms of this identifier vary from repository to repository. Including an identifier will not only specifically recognize the collection, but will also help a user locate the collection even after major changes to the repository or the name of the collection. Table 1 gives examples of identifiers used by three major language repositories and one institutional data repository in 2019. For many repositories, this identifier is part of the collection's URL, whereas at others, like ELAR,[12] it is not. If possible, a direct link to the collection's public webpage should be included. A collection may also have an external persistent identifier such as a DOI or handle which can direct users directly to the collection. Many discipline-neutral data repositories, such as those using Dataverse software, will not have a local identifier as such, but may identify the collection in a URL (such as "indicating" in the URL in Table 1), and may or may not have external identifiers for entire collections.[13] Note that identifiers, and especially external identifiers, may not be available until some late stage of the collection's processing.

[12]The Endangered Language Archive. https://elar.soas.ac.uk/.

[13]This may vary depending on the configuration of the repository software and the organization schema used – external identifiers are provided for datasets (folders) in the Texas Data Repository Dataverse, but not for dataverses (collections).

Table 1. Identifiers for three language repositories[14]

Repository	Local identifier	URL	External identifier
PARADISEC	AA1	http://catalog.paradisec.org.au/-collections/AA1	DOI:10.4225/72/56E-7A7256C274
ELAR	0448	https://elar.soas.ac.uk/Collection/MPI1035101	—-
AILLA[15]	ailla:124459	https://ailla.utexas.org/islandora/object/ailla:124459	—-
Texas Data Repository Dataverse[16]	—-	https://dataverse.tdl.org/dataverse/indicating	[folder objects only]

3.1.1.2 Name and location of repository Include the *name of the repository* where the collection can be found. While this information could be inferred from a local identifier or the collection URL, an external identifier such as a DOI will offer no such hint about the repository's identity. A direct link to the repository's main webpage could also be included here, as in the following examples.

- *Endangered Languages Archive, SOAS University of London,* https://elar.soas.as.uk
- *Alaska Native Language Archive, University of Alaska Fairbanks*, https://www.uaf.edu/anla
- *Texas Data Repository Dataverse, Texas Data Repository,* https://dataverse.tdl.org/

3.1.1.3 Collection title Include the *title of the collection* as it appears in the repository. If the collection has or has had a different title – for example, if the collection title has been changed, or if the collection is also known by a translated title – these titles should be included as well.

- *A preliminary documentation of the Okiek language of Kenya*

(Oduor 2016)

- *Alan Walker's Sabu materials*

(Walker 1975)

[14]The collections cited in this table are Adelaar (1986), Lovestrand (2017), Pride & Pride (2007), and Mesh (2017).
[15]The Archive of the Indigenous Languages of Latin America. https://ailla.utexas.org.
[16]https://dataverse.tdl.org.

- *Chuj Oral Tradition Collection of Pedro Mateo Pedro and Jessica Coon*
  Spanish title: *Colección de Tradición Oral Chuj de Pedro Mateo Pedro y Jessica Coon*
  Chuj title: *Lolonel Sb'eyb'al Chuj*

(Coon & Mateo Pedro 2017)

3.1.1.4 Collection dates Two kinds of dates are important for collection descriptions: the dates of collection and creation of collection materials, and the dates when a collection was published or updated in the repository. *Publication* or *update dates* are needed to cite (the appropriate version of) the collection, and *collection* and *creation dates* are needed to inform users of the time span that a collection covers.

Give the year span including the earliest and latest date of creation (or revision) of the materials in the collection. These are *inclusive dates*. In most project-based language collections, most items will be created between the start and end dates of the project, however, a collection may include items that were created well before or after the period of significant active collection. In these cases, the *bulk* or *predominant dates* of the collection can be specified as well. For example, if a collection includes recordings made a generation prior, the dates of the original recordings would be in the inclusive date range but not the bulk date range. Since a collection description captures a collection at a moment in time, it is best to not use open ended spans like "2016–" or "2016 – ongoing" and instead end the date span at the current year. If there is a gap between periods of significant creation, multiple date spans can be indicated.[17]

- *2009–2011*
- *1970–2015, predominantly 2011–2015*
- *2006–2007, 2015–2018*

3.1.1.5 Names and roles of contributors List the *names of the individuals who contributed to the creation of the data in the collection*. In most dedicated language repositories, only the collectors and depositors are named in the collection's metadata record, and other contributors are identified only at the folder or file level, making their contributions less visible. Including only collectors and depositors in this section will be appropriate for some collections, whereas including more collaborators will be appropriate for others. This section could include only those people who worked on large parts of the collection materials – such as the project's transcription and translation team, the speakers and signers who provided significant portions of the recorded language use, or the project's videographer – and omit here any participants whose contributions are limited to a few files or folders, such as someone who contributed a single narrative or a research assistant who analysed a data set.

[17]All examples without citations are hypothetical.

The role individuals played in the collection can also be specified. As always, identifiable information should only be shared with permission, and requests for anonymity or deidentification should be honored.

Depositors:

- *Jessica Coon*
- *Juan Jesús Vázquez Álvarez*

Collectors:

- *Juan Jesús Vázquez Álvarez*
- *Nilda Patricia Guzmán López*
- *Ana Claudia Díaz Jiménez*
- *Juan Mario Mayo Hidalgo*
- *Patricia López Vázquez*
- *Estela Álvaro Díaz*
- *Adrián Sánchez Méndez*
- *Abigael Pérez Gómez*
- *Félix Ignacio López López*
- *Elmar Martínez*
- *Pedro Gutiérrez Sánchez*
- *Juan José Juárez Pérez*
- *Damián López Gutiérrez*
- *Uriel Martínez Pérez*
- *Matilde Vázquez Vázquez (transcriber and translator)*

(Sullivant 2019*b*)

3.1.2 Extent and scope of content A collection description should contain statements about the *scope* and the *extent* of the collection. In archival description, a note on the scope of a collection is a summary of the materials in the collection, the circumstances of their production, and what kinds of information they contain. The extent is the quantity and size of the aggregated materials in the collection. While physical archives may measure extent in the number of boxes that a collection is housed in or how much space papers take up on shelves, digital archives can provide extent information by counting the number of digital files or saying how much space they take up. Since language collections and corpora are frequently measured by

the duration of their audiovisual recordings and/or amount of annotated transcriptions, these figures are also useful in defining the extent of a collection. Scope and extent are distinct concepts that are often separated out in archival description, but are grouped together here since this information can easily be combined into a single narrative description. For example, the descriptive statement below by Applebaum (2010) provides the scope of the data without giving any information about their extent, and the next statement is a simple statement of the extent (i.e., number) of each of the file types within. These two types of information could be combined into a single narrative statement, as in the third example.

- *This deposit includes audio, text, images, ELAN, and metadata files, of elicited word lists glossed in English and Turkish, children stories, Kabardian mythological story (nart sagas), riddles, traditional recipes, conversations, and narratives.*
  (Applebaum 2010)

- *19 WAV, 19 EAF, 12 TXT, 10 JPG*

- *19 WAV audio recordings of elicited Kabardian word lists, children's stories, mythological stories (nart sagas), riddles, traditional recipes, conversations and narratives, 19 EAF files containing time-aligned Kabardian transcriptions glossed in Turkish and English, 12 TXT files containing Kabardian transcriptions with Turkish and English glosses, and 10 JPG images of research participants*

A statement like this will ideally list all file formats present in the collection, along with the number of each and/or their total size. This information is crucial for users to evaluate collections, but is often not found in a collection's metadata, and may be difficult for a user to determine given the structure of many digital language archives. At a minimum, this note would provide a summary of the kinds of materials present in the collection (audio recordings of narratives, video recordings of face-to-face conversations, interlinear glossed texts, etc.). A more detailed listing of content is a second-priority item (§3.2.1).

Given the centrality of annotations to the use of language collections, some comment ought to be made about what portion of audio-visual materials have been transcribed, translated, or morphologically analyzed. Since we cannot presume that all collection users will immediately recognize file formats commonly used by linguistic researchers but almost no one else – nor can we presume that they will all have a place in the linguists' digital tool box in the future – a very brief description of the kind of content in each file format may be given. For example, "WAV audio recordings", "XML documents in EAF format containing time-aligned transcriptions made with ELAN", or "PDF documents containing English translations". These statements may also be a suitable place to mention any processing or file conversions that were done (§3.3.5).

- *Audio-video data of Bininj Kunwok language recorded in free-recalls, interviews, conversations, commentaries some of it recorded with GoPro cameras, and geotagged.*
  (Cialone 2016)

- *The deposit comprises over 150 audio files, as well as written texts, elicitation materials and participant observation notes.*

  Genres include traditional narratives, procedural and route descriptions, personal stories and accounts of everyday events in which the researcher participated. Elicited materials include responses to the caused positions and cut and break video stimuli, frog stories, and men-and-tree games, which were used in the researcher's doctoral investigation into the Kubokota directional system.

  (Chambers 2007)

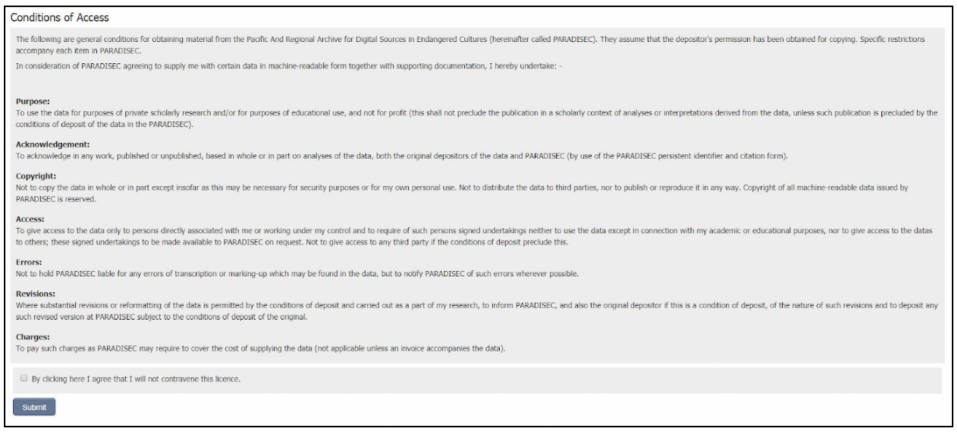
- *This collection contains twelve audio recordings of traditional narratives. Most are accompanied by transcriptions and translations in illustrated children's primers.*

  (Margery Peña 2002)

3.1.3 Access and use restrictions The collection description should alert the user to any *access* and *use restrictions* that may limit how they interact with the collection's materials. Note that access and use are distinct: access restrictions govern how a user may view, stream, or download files whereas use restrictions govern what the user can then do with the files. It may be the case that data are freely available but certain types of use – such as use with modification or commercial reuse – are prohibited.

Many repositories have some access and/or use restrictions that apply to all the materials they hold. Users may be notified of these conditions when creating a user account or while accessing media. For example, PARADISEC asks users to agree to the terms in Figure 1 when downloading a file. These terms include restrictions on reuse of items (use must be non-commercial, attributed, and some kinds of distribution are prohibited), and there are conditions that must be met to access the data – namely that a user must agree to respect the use restrictions.

Figure 1. PARADISEC Conditions of Access (as of January 23, 2020)



Many digital language repositories allow materials to be restricted beyond the general access restrictions. If any part of a collection has access or use conditions distinct from the general conditions of the repository, this should be made clear in the collection guide so that a user will not spend time locating files they may not be permitted to view or use for their intended purposes. Depositors may be useful here to specify reuse conditions with the increasingly common Creative Commons licenses[18] or Traditional Knowledge Labels[19] (Anderson 2012; Anderson & Christen 2013), but they are cautioned to carefully read these conditions to determine if they are appropriate. In cases where no special or additional conditions on access or use are necessary, a generic statement such as "subject to access and use conditions set by the repository" or "contact the repository for access and use restrictions" will suffice.

Some digital language archives have mechanisms in place to allow a user to request access to materials from the depositor themselves, but these messages are usually succinct and general. For example, when attempting to view files that are available on request, ELAR users (as of April 6, 2020) are greeted with the following message:

> "This file is only available upon request. Please contact the depositor for individual access rights. You will receive an email with the outcome of your request. Note that the contents of some resources are sensitive and that your request may not be granted for this particular file."

Users may be discouraged by the notion of having to bother a stranger to see materials, but may be encouraged to reach out if they encounter language in a collection description like "All reasonable requests for access by members of the community/academic researchers/students of the language will be accommodated".

- *Materials are publicly accessible subject to the repository's conditions on access and use.*
- *Materials have no special access restrictions except that materials published elsewhere are restricted until the time embargo specified by the publisher elapses.*
- *All materials in the collection are restricted. Users may request access to materials via the depositor's email. Requests for access from community members and students of the language are particularly welcome.*

3.1.4 Languages and scripts List what *languages and scripts* appear in the collection – both the languages that are the focus of the collection and any languages used in translation, annotation, or analysis. Most language collections often explicitly identify the language that is the focus of research in the collection's title, whereas other languages used in collections are less prominently identified. These other languages can include languages used in translations and glosses, or any languages that may be spoken or signed in recordings. Information about these languages may only appear

[18]https://creativecommons.org/.
[19]http://localcontexts.org/.

in catalog records for files or folders – if at all – and are not immediately visible to a user evaluating a collection.

A summary statement about the distribution of languages throughout the collection can be made instead of or in addition to a list of languages. This may be preferable if certain languages are confined to particular kinds of files. If the languages present in a collection are quite varied, as may be the case for large surveys or collections from multilingual communities, it would be enough to note what the languages present are, and refer the user to the pertinent folder or file-level metadata in the folder list in the scope and content note. It should not be assumed that English is the language used in the collection for notes, analysis, meta-documentation, and translation.

If more than one script or writing system is available for a language, or if the script does not enjoy wide technical support, these details can be given as well. Oftentimes, depositors devise their own orthographies for languages, either because they are among the first people to write a language, or else they find prevailing orthographies lacking. If a collection contains writing in a bespoke practical orthography, the user should ideally be referred to some document (either within the collection or published elsewhere) explaining its details.

- *Targeted towards users either from the Korean-speaking realm or with active involvement in Korean studies, transcribed annotations focus on providing an IPA transcription, a Korean script transcription of Jejuan, and a Korean translation into Korean script. An expanding subset of these annotations is being enriched with English translations.*

  (Kim 2018)

- *Audio recordings are predominantly in Baniwa and Portuguese. Photograph captions are in Portuguese and English. Documents are mostly in Baniwa, Portuguese and English. The Baniwa is written in a contemporary orthography (where aspirated consonants are indicated by <h>) apart from some earlier documents using the orthography of the Salesian missionaries (where aspirated consonants are indicated by carons <ˆ>). A few documents are in Spanish and French.*

3.1.5 Citation information Specify how the collection should be cited, or direct users to guidelines for citing the collection and its subparts. Including *citation information* in a collection description helps build a practice of citing archival data in linguistics, which is necessary since many subdisciplines of linguistics lag considerably behind best citation practices (Berez-Kroeker et al. 2018). If repositories provide this information themselves, then the user could just be pointed to where they may find the citation guidelines or sample citations. Note that repository citation guidelines or offered citations may not include elements considered essential by some organizations or recommendation guidelines, such as Andreassen et al. (2019) and DataCite Metadata Working Group (2019).

- *Users are requested to acknowledge the depositor, Mary Chambers, when citing resources from this deposit.*
  (Chambers 2007)

- *Users are directed to follow AILLA citation guidelines (https://ailla.utexas.org-/site/rights/citation) when citing materials in this collection.*

- *Cite as: Claudia Cialone (collector), 2016; Recordings of Bininj Kunwok bush culture, language, spatial navigation documentation (CC01), Digital collection managed by PARADISEC. [Open Access] DOI: 10.4225/72/56E979E6F4200*
  (Cialone 2016)

3.2 Second priority collection metadata The second set of information is *recommended* but not essential for a minimal or initial description. This information can help situate the collection in the history of an individual or a field, as well as help the user find and use materials within the collection. This information is largely present in the introduction and background sections of the *LD&C* collection articles (Schembri et al. 2013; Salffner 2015; Caballero 2017; Oez 2018; Gawne 2018; Franjieh 2019, Hildebrandt et al. 2019), but is not found in most collection's catalog records.

3.2.1 Detailed contents list Give a *list of all the contents of the collection*, or direct users to where such a list may be found. This is especially important for larger collections that contain a great number of files as repository software may have a limit on how many items can be displayed per page, meaning that the contents of large collections will never be visible all at once on a single page. The contents list should at least identify each folder in the collection with some of its basic information. Beyond the folder's name, information to consider including are the dates and location of creation, the languages involved, the topic of the material, participants involved, etc. The content of folders in each major subpart of the collection can be briefly described in general terms. For many language collections, this will be a statement about how it contains folders of audio-visual recordings and their annotations.

A detailed contents list makes a collection guide function as a finding aid, pointing users directly to the materials, especially if the document contains direct links to objects in the collection. While most corpora and collection review articles do not include detailed contents lists (perhaps for reasons of space or perhaps because the lists published in journal articles will likely become out-of-date) some do direct readers to contents lists or metadata spreadsheets within their collections (Salffner 2015; Caballero 2017). Folder lists are likely sufficient for most language documentation collections, though some collections might benefit from having a contents list that identifies all files in a collection, or all the files in a few of the collection's folders. It may be helpful to list individual files if folder contents are very heterogeneous, or if related files are scattered across folders (requiring a user to look in multiple folders to identify materials that belong together).

Since collectors and depositors often have little control over the display order of items in their collections, this contents list may be an opportunity to present the

collection in an alternative arrangement and order to what is visible in the repository. Whereas most repository software will display objects in alphabetical order according to title, a collection guide can create a contents list that orders items chronologically or groups together series of folders according to geographic or thematic criteria. For collections that are undergoing annotation, the folders which have been transcribed and/or glossed could be grouped together in the list. It may be useful to separate all public content from restricted content in this list.

Creating such a detailed contents list from scratch may be very difficult, and is probably unnecessary. Depositors who keep metadata registers in spreadsheets or relational databases may be able to export them into an archivable format and include a column or field giving the object's identifier or a link to it in the repository. Such a metadata register could either be included in the collection and referenced in the collection guide or could be used to generate a contents list that can be added to the guide itself.

Examples of detailed inventories are not given here due to space considerations, but readers are encouraged to consult some of the following online contents lists and inventories to see different approaches to the presentation of a contents list.

- *Recordings of Native American languages (Mss.Rec.184), American Philosophical Society.* https://search.amphilsoc.org/collections/view?docId=ead/Mss.Rec.-184-ead.xml
- *Series SAW2S002 – Solomon Islands Materials. PARADISEC.* https://www.paradisec.org.au/fieldnotes/SAW2/SAW2S002.htm
- *Guide (in English) to the Iskonawa Language Collection of José Antonio Mazzotti, Roberto Zariquiey, Rodolfo Cerrón Palomino and Carolina Rodríguez Alzza. Archive of the Indigenous Languages of Latin America.* https://ailla.utexas.org/islandora/object/ailla:256955

3.2.2 Biographical sketches and project history A language documentation collection provides a record of a culture in a particular place and time, and a complete description of the collection should include *background information* about the community represented in the collection. The amount of detail that is appropriate will vary, but should be enough to contextualize the materials present in the collection. Part of this background information can include the motivation for the creation of the collection and previous documentation and description efforts.

It is important to remember that one of the ways that collections have value is as evidence of their creators and the contexts they were created in, and *biographical sketches* are one way that this kind of information can be given. Future users may like to have some information about the creators beyond their names, affiliations, and languages spoken. Biographical sketches can either be in narrative form or in the form of a chronology of major events in the person's life, especially regarding education, employment, and life events which help contextualize the collection within the creator's work.

Apart from descriptions of the legacy collections (for example, in the Background Notes of finding aids in the American Philosophical Society collections), often very little attention is paid to the depositors of project-based collections. This may be in part because depositors, who are often not members of the language community of the collection, aim to center the language and its users rather than themselves. Many language documentation projects gather demographic information from contributors – including depositors and research personnel – and after taking privacy, safety, and personal preferences into account, this may be included in a collection's metadata or in a document within the collection. Biographical information about contributors is usually absent from collection description articles, though Oez (2018) is a notable exception. His discussion of why and how he chose the contributors to his collection helps provide a very full picture of the history of his project.[20] Hinton (2005:25) suggests that even more detailed biographical information ought to be collected for the speakers and signers present in a collection.

> *A medical doctor by training, Jaime de Angulo (1887–1950) became interested in anthropology and linguistics in the 1910s. He joined the Department of Anthropology at the University of California, Berkeley from 1920–1922, during which time he conducted linguistic research on Pomo and Achumawi (Pit River). De Angulo held a research appointment in Mexico's Department of Anthropology from 1922–1923 and documented several indigenous languages of Mexico. On returning to the United States, de Angulo resumed his research on the languages of California in collaboration with Lucy S. Freeland, publishing grammatical descriptions and texts for Achumawi, Karuk, Miwok, Pomo, and Northern Paiute in the 1920s and early 1930s. De Angulo was also a well-known poet and literary figure in California, and a collection of poems and short stories was published under the title Indians in Overalls (1950).*
>
> (Benson et al. n.d.)

3.2.3 Arrangement and collection conventions Provide a statement describing the *arrangement of items* in the folders of the collection and any information useful for understanding their organization, especially regarding how related files may be identified. One of the most daunting tasks an archive user faces is understanding how the materials in the collection are organized. Are data arranged in strict chronological order, alphabetical order, by language, genre, etc.? Collection reviews do an exemplary job of providing this information under headings such as "collection conventions" or "data processing". Users need to know about distinctive series of folders within a collection. For example, a certain range of folders contains recordings of natural conversation and another contains interviews about a particular cultural practice, yet another set of folders contains short utterances recorded for a talking dictionary, etc. This is the area where the organization of media files into folders may be explained.

[20]Specifically, recording was done with a group of Turoyo speakers in Germany rather than in Turoyo-speaking communities of Turkey due to political instability in that country during his project's time frame.

This may be as simple as noting that the collection generally follows the common language documentation collection arrangement where each folder contains audio-visual recordings, annotations, and ancillary materials from a single recording session, or if data are grouped by speaker or signer, or if data are grouped by experimental protocol, etc.

Collection conventions should also include *descriptions of the annotations* present in the collection. Both corpora and collection description articles pay special attention to the contents of annotation files, though this same level of description is generally absent from most collection metadata. Users could be given details about annotations, whether or how translations were created, and what level of morphological analysis is present.

This is also a suitable place to explain any *file naming conventions*, which are especially useful given many linguists' practice of producing materials with semantically rich file names with codes indicating information such as the participants involved, language variety, date, location, topic of the recording, etc.

> *The file name of each document is a unique identifier with continuous numbering and an abbreviation encoding the type of document involved, e.g. el45, where "el" = elicitation, "tx" = text, "co" = conversation, "in" = interview, "tr" = free Spanish translation, "en" = elicitation notes, "te" = language teaching, and "mu" = music.*
>
> *[…] the collection includes written annotations carried out by the documentation project members listed in (1) above, containing minimally a broad phonetic transcription and a Spanish free translation. In addition to this, some annotations additionally contain English free translations, morphological glosses, and fine phonetic transcriptions. Most annotations include a commentary specifying recording circumstances, issues in transcription and translation of particular segments, comments by speakers, and relevant grammatical or cultural information associated with any piece of data. All annotations also provide relevant metadata describing the content, associated media file(s), and recording circumstances of the document annotated.*
>
> *Most transcriptions use the Americanist convention of transcription, though some files contain a transcription using certain symbols from a Spanish-based orthographic convention (e.g., <ch> for the alveopalatal affricate and <rr> for the alveolar trill). Transcriptions also encode pauses as "…" and grammatical or general comments in angled brackets. Stress is marked with an acute accent diacritic. Tones are not marked.*
>
> (Caballero 2017:235–236)

3.3 Third priority collection metadata The last group of information to include is *optional information*. While this information is less crucial for most collections, some of this information may be helpful for users. For many language collections featuring only born-digital materials, there will be little or nothing to include under many of

these categories, especially those relating to physical materials' access or conservation.

3.3.1 Physical and technical access Provide information about any procedures, equipment, or software needed to access the materials. *Physical access* information includes, for example, the hours and policies for viewing (or requesting to view) objects in an archive reading room. This information does not apply to an online-only digital collection that does not close, but may be relevant for collections that have analog components (such as undigitized analog media, specimens, or realia) or digital components that for privacy or copyright reasons may only be accessed on-site. *Technical access* considerations for digital language collections include notes about what software applications (and if applicable, which versions) are needed to use a collection's files. For most digital language collections using enduring and open file formats, users can be expected to have access to the common applications needed to view or play most documents and audiovisual materials. Some digital language archives describe some of the linguist-specific file formats and applications in information pages or FAQs, but depositors at general data repositories should consider including brief statements about what applications are used to read less common file types. For example, explaining that EAF files are opened with the ELAN application and aligns XML transcriptions to audiovisual files or how to open a Toolbox database or Flex project may be helpful for some users.

> *Trilingual dictionary, Yakkha, Nepali, English, contains information on word class, source language, semantic category, and examples (Yakkha-English only) for roughly one third of the entries. The Lexique Pro file is an executable file to be installed on a hard drive first. It is for users who prefer the Yakkha data in Devanagari script. I am grateful to Kamala Linkha, Manmaya Jimi, Magman Linkha and Mohan Khamyahang for their help, comments and suggestions. All remaining mistakes are my own. The dictionary only runs on Windows systems. To run Lexique Pro under Mac OS or Linux, you need a Windows emulator. For more information on Lexique Pro, visit www.lexiquepro.com.*
>
> (Schackow 2014)

3.3.2 Keywords Some digital language archives such as ELAR and TLA allow depositors to tag the materials in their collections with *keywords*. This is meant to facilitate discovery of related materials within and across collections, but in practice depositor-generated tags tend not to be useful additions to collection metadata (Grant et al. 2019), and they frequently have a very long-tail distribution that does not aid discovery.[21] Nevertheless, if a depositor purposefully assigns keywords to organizing series of data within a collection, or to display the prevalence of themes within a

[21]In two digital language archives that have depositor-generated keyword tags, The Language Archive and ELAR, the median keyword was used twice, and 47.8 % of TLA and 40.3% of ELAR keywords are unique. See Appendix C for data and analysis.

collection (Oez 2018:353), then this can be explained, as done by Franjieh (2018) in Figure 2.

Figure 2. Organization of genre and sub-genre keywords in Franjieh (2018)

Deposit contents
Currently there are over 20 hours of video and 50 hours of audio recording for the North Ambrym language. For the Fanbyak language there are over five hours of video and 22 hours of audio recordings.

The deposit is split up into genre and sub-genre (topic in the ELAR menu).

Genre	Sub-Genre / Topic
narrative	traditional narrative, personal narrative, sand drawing, string figure, oratory, report, exposition, narrative stimuli
dialogue	sociolinguistic survey, language use survey, meeting, discussion
music	instrument, song, stringband
observational filming	traditional ceremony, traditional work
elicitation	possessive classifier, wordlist, grammatical structure, paradigm, elicitation stimuli, species identification
literacy materials	no sub-genre
academic materials	no sub-genre
legacy materials	no sub-genre
field notebooks	no sub-genre
genre guides	no sub-genre

For more information about the different genres and sub-genres please read the genre guides (located in the genre section). A paper has been submitted to Language Documentation and Conservation journal which gives a fuller overview of this deposit. This will be added to the academic materials once published.

3.3.3 Accruals State if any *accruals* are expected, that is, whether or not the collection is expected to grow, and what the frequency and content of any additions is forecasted to be. This information will help a user understand the expanding nature of many language collections, especially in light of the adoption of progressive archiving practices and some funders' timely archiving mandates. For a user who is interested in using the collection's materials but requires certain kinds of annotations or derivatives – for example, transcriptions or translations – it could be helpful to know if these may be added in the future or are unlikely to be produced. This information is included in many collection description articles either in sections such as "Next steps" and "Work in progress and future directions" or is mentioned in passing, but is often not found in collection metadata.

> *Over two hours of these recordings have been translated into Bislama and English, with the remainder translated into Bislama only at this stage. None of these recordings have been glossed at this stage, however these will be added over time, along with English translations where needed.*
> (Dewar 2018)

> *The transcriptions are largely based on the spelling conventions suggested in Baxter and de Silva (2005). They are still a work in progress and will be amended periodically as they are checked by other native speaker consultants. More data will be added in 2013.*
> (Pillai 2011)

> *A Kichwa translation of* Juncal *is in preparation and may be added at a later date.*
>
> (Sullivant 2019a)

3.3.4 Custodial history The *custodial history* refers to the succession of people or institutions that owned or held materials before they were in the possession of the depositor. Language collections are increasingly born-digital and very recent, so there will often be very little to say about a collection's custodial history. A statement that will work for most collections being deposited today could be "The materials were created by the depositor (in collaboration with others), and given to the archive in such-and-such year." More detailed statements may be in order if the collection includes materials created by people other than the depositor. In cases like these, it can be useful to point out how the depositor received the materials from their creators. Future users and archivists would like to know the circumstances in which materials were given and what permissions were granted, as well as the rights information of the materials: who holds the copyright? If the materials are thought to be orphaned works (copyrighted items whose creator is unknown or whose known creator cannot be reached), then it will be useful to know how this was determined. This kind of information is useful for archive users to know the intellectual history of materials that otherwise might be anomalous or difficult to explain. This custodial history also helps draw out unknown or unexpected connections between individuals that might be informative, and should be noted when appropriate.

> *Two of the audio cassettes given to AILLA for this collection (Kane 2018) were made by Joel Sherzer (1970a,b) years before her research began. While Sherzer was living in a Guna-speaking community of Panama, he recorded a couple word lists in open-reel tape from an Emberá man who was visiting the village, and later transferred these recordings to audio cassettes. When Kane was beginning to study Emberá in the 1980s, Sherzer gave her these audio cassettes for her own use.*

3.3.5 Appraisal, processing, conservation, and migration *Appraisal* is the process by which someone decides which materials merit preservation and inclusion in a curated collection. Traditionally, appraisal was done by archivists who reviewed an entire assemblage and judged what merited preservation. Today, appraisal is mostly done by the depositors themselves when selecting which materials they will deliver to a repository. Even projects that aim to archive "everything" will ultimately exclude some materials. In most cases, only the latest version of annotation and analysis files are delivered to the repository, and much administrative material and ephemera associated with the project is excluded: correspondence, personal notes, administrative paperwork, informal photographs, and audio level recording tests. If a depositor decides to exclude sets of materials for whatever reason, these kinds of decisions can be explained here.

Processing and migration includes the digitization of analog materials and the migration of digital files from one format to another. A depositor may not be aware of

the full technical details of how materials were processed since these steps may have been taken by archivists or a third-party vendor, but even statements identifying what the original format of the materials was can be useful since these will imply certain kinds of processing. If the files themselves were altered – for example, if the signal of digital audio files was modified – this should be noted since such filtering and processing may affect the suitability of the recordings from some kinds of analysis or reuse. Migration notes would explain how proprietary formats were converted into open enduring formats for preservation as well as any format conversation undertaken in the collector's pre-depositing work – for example, if uncompressed MTS video files were converted to MP4 in the field to save storage space.

Conservation includes any actions taken to improve or preserve physical items. While unlikely to apply to born-digital language collections, conservation notes could include pre-digitization steps taken to repair broken cassette tapes or clean photographs, for example:

> *During digitization, staff discovered that an ant had eaten through the tape of one audio cassette. The tape was spliced and digitized resulting in two files per side of tape, one before and one after the splice. It is unclear exactly how much audio signal was lost.*
>
> *AILLA staff renamed files to replace spaces in file names with underscores, as spaces are not permitted by the repository's software. PDF documents were migrated to the archival PDF/A format. The PDF file containing the Danish version of* Juncal *was created by concatenating and rotating single-page scans apparently made for the purposes of translating the text, as no pages containing only images are included. Some photographs were excluded.*
>
> (Sullivant 2019a)

3.3.6 Bibliographies, finding aids, and published Materials A collection description could reference documents that offer additional information about the collection, collector, or the themes of the collection. This could include a *bibliography* of resources that are useful for studying the language and culture, or for understanding the collector's work. It can be useful to include references to published books and articles by the collector as well as others that are based on data in the collection or otherwise relate to them, especially if a digital copy cannot be included in a collection for (e.g.) copyright reasons.

A collection description could include references to other *finding aids or descriptions* that cover the collection. This could either be an earlier version of the collection description or a reference guide that includes the language collection.

> *Published Materials*
> *Anthropology Resource Center. 1981. The Yanomami Indian Park: A Call for Action. ARC: Boston, 25 pp. booklet.*
> *Cornelio, José Marcellino and Robin M. Wright. 1999. Waferinaipe Ianheke. A Sabedoria dos Nossos Antepassados. ACIRA & FOIRN: São*

Gabriel da Cachoeira.
Wright, Robin M. 1998. Cosmos, Self and History in Baniwa Religion. For Those Unborn. Austin: University of Texas Press.
Wright, Robin M. 2005. História Indígena e do Indigenismo no Alto Rio Negro. Campinas: Mercado de Letras & Instituto Socioambiental.
Wright, Robin M. 2013. Mysteries of the Jaguar Shamans of the Northwest Amazon. University of Nebraska Press, Omaha.

3.3.7 Related materials elsewhere The finding aid can also point the user to *related materials* that are located in other repositories. Users could be referred to other archival collections (digital or not) that include materials about the languages, people, or cultures represented in the collection. Similarly, if any materials collected or created by the depositor have been sent to other repositories this can be indicated here as well. For example, the audio-visual and annotation files produced during a project may be in a digital language archive, but the thesis based on those materials may be housed in a different institutional repository.

> *Some of the photographs in this collection, as well as other photographs by Fock and Krener, may be found in the digital collections of the Danish National Museum (https://samlinger.natmus.dk/). In Danish […] Materials related to Niels Fock's 1950s research in British Guiana can be found in the Waiwai Collection of Niels Fock at AILLA (ailla:124389).*
>
> (Sullivant 2019a)

3.3.8 Alternative available forms (Originals, copies, realia) While language collections are increasingly born-digital, some language collectors still produce some physical materials and occasionally may gather analog legacy materials: many linguistic researchers still work in notebooks (Bowern 2015), ethnobotanical projects collect physical specimens, and a collector may be given tapes recorded by others. Even though their digital surrogates will be accessed and used far more frequently than the objects themselves, a user may seek out the physical objects for purposes of exhibition or, particularly in the case of specimens, additional research. If these objects are donated to a physical collection after they have been digitized and ingested into a digital archive (or already belong to a physical collection), then the physical items would be considered *alternative available forms* of the digital files, and their existence should be noted and enough information given to locate them within their collection (title, call number, series name, folder label, etc.).

> *The Danish National Museum also holds a small physical collection of Cañari artifacts (chiefly clothing and implements) [Fock and Krener] collected.*
>
> (Sullivant 2019a)

3.3.9 Notes on the production of the guide Any notes about *sources and tools used to produce the finding aid* can be listed. This can include any sources that were consulted to provide contextual information about the collection, create biographies of contributors, or locate related archival materials. For many collections based on ongoing or active projects, and especially for descriptions made by the collectors themselves, there is unlikely to be external sources consulted to create the guide. Some of the kinds of resources that may have been used when making a guide can include collection descriptions and guides that have served as a model for the description. If a bilingual dictionary or resource on orthography was used to present a normalized spelling of terms within the collection's titles and descriptions, this can also be made explicit. This is also where details about who created the guide and any updates to the guide can be made.

> *A bilingual dictionary (Consuelo Yánez Cossío. 2007. Léxico Ampliado: Quichua-español, español-quichua. Editorial Abya Yala: Quito) was consulted during processing.*
>
> (Sullivant 2019a)

4. Conclusion It is hoped that this paper and its checklist and guidelines will help people describe their own digital language documentation collections and understand the many different kinds of information that such a description may include. This paper may help reviewers evaluate collection descriptions for completeness by pointing out what critical bits of information might be missing from a basic description. At the same time, it may also make it easier to recognize descriptions that go above and beyond in their detail and completeness.

A few final words of encouragement. Collection descriptions do not need to be peer reviewed to be useful, nor must they only exist as fully fleshed-out descriptions with detailed content lists. Descriptions can be included in the collection metadata records (the "front page" of the collection) or could be added to the collection itself as a conspicuous document, for example, as a readme file in a data repository. Even partial description is better than none, and a collection described only with the first level of collection information above will be better described than many language documentation collections. Descriptions can always be improved upon and amended as data are added to the collection, as time allows, or as one's understanding of a collection becomes clearer. The first priority information can be included during or soon after the initial curation of the collection and updated following any major additions or rearrangements of data – for example, if newly-collected data is added to the collection or if transcriptions are added to folders. Subsequent descriptions can then expand on the description by adding information from the second and possibly third priority categories. New versions of collection descriptions can be identified by a numbering scheme, which will minimally either make use of a version number and/or the date of the guide's updating. It may also be useful to follow a semantic numbering system distinguishing between major (v1.0 → v2.0) and minor updates (v1.0 → v1.1).

References

Adams, Oliver, Trevor Cohn, Graham Neubig, Hilaria Cruz, Steven Bird, & Alexis Michaud. 2018. Evaluating phonemic transcription of low-resource tonal languages for language documentation. *Proceedings of LREC 2018: 11th edition of the Language Resources and Evaluation Conference*, Miyazaki (Japan), 7–12 May 2018.

Adelaar, Alexander (collector). 1986. Recordings of Selako (Indonesia) (AA1). Digital collection managed by PARADISEC [Open Access]. http://catalog.paradisec.org.au/repository/AA1.

Anderson, Jane. 2012. Options for the future protection of GRTKTCEs: The traditional knowledge licenses and labels initiative. *Journal of the World Intellectual Property Organization* 4(1). 66–75.

Anderson, Jane & Kim Christen. 2013. "Chuck a copyright on it": Dilemmas of digital return and the possibilities for Traditional Knowledge licenses and labels. *Museum Anthropology Review* 7(1–2). 105–126. https://scholarworks.iu.edu/journals/index.php/mar/article/view/2169.

Andreassen, Helene N., Andrea L. Berez-Kroeker, Lauren Collister, Philipp Conzett, Christopher Cox, Koenraad De Smedt, Bradley McDonnell, & the Research Data Alliance Linguistic Data Interest Group. 2019. Tromsø recommendations for citation of research data in linguistics. *Research Data Alliance*. doi:10.15497/RDA00040.

Applebaum, Ayla Bozkurt. 2010. Documentation and analysis of Kabardian as spoken in Turkey. *Endangered Languages Archive*. SOAS University of London. https://elar.soas.ac.uk/Collection/MPI122731.

Austin, Peter K. 2013. Language documentation and meta-documentation. In Jones, Mari C. & Sarah Ogilvie (eds.), *Keeping languages alive: Documentation, pedagogy and revitalization*, 3–16. Cambridge: Cambridge University Press.

Benson, William Ralgonal, Maria Rosa Gutierrez, L.S. Freeland, & Jaime de Angulo. n.d. *Jaime de Angulo Papers on Indigenous Languages of California and Mexico, Angulo: Survey of California and Other Indian Languages*. University of California, Berkeley. doi: 10.7297/X2MW2F24.

Berez-Kroeker, Andrea, Helene N Andreassen, Lauren Gawne, Gary Holton, Susan Smythe Kung, Peter Pulsifer, Lauren B. Collister, The Data Citation and Attribution in Linguistics Group, & the Linguistics Data Interest Group. 2018. The Austin Principles of Data Citation in Linguistics. Version 1.0). http://site.uit.no/linguistics-datacitation/austinprinciples/.

Berez-Kroeker, Andrea, Lauren Gawne, Susan Smythe Kung, Barbara F. Kelly, Tyler Heston, Gary Holton, Peter Pulsifer, David I. Beaver, Shobhana Chelliah, Stanley Dubinsky, Richard P. Meier, Nick Thieberger, Keren Rice, & Anthony C. Woodbury. 2018. Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics: An Interdisciplinary Journal of the Language Sciences* 56.1. doi: 10.1515/ling-2017-0032.

Bowern, Claire. 2015. *Linguistic fieldwork: A practical guide*. New York: Palgrave Macmillian. Bourne, Phil E., Tim Clark, Robert Dale, Anita de Waard, Ivan Herman, Eduard Hovy, & David Shotton. 2011. FORCE11 Manifesto: Improving future research communication and e-scholarship. https://www.force11.org/about/manifesto.

Boyd, Doug. 2013. OHMS: Enhancing access to oral history for free. *The Oral History Review* 40(1). 95–106. doi: https://doi.org/10.1093/ohr/oht031.

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

Ćavar, Małgorzata E., Damir Ćavar, & Hilaria Cruz. 2016. Endangered language documentation: Bootstrapping a Chatino speech corpus, forced aligner, ASR. Presented at the Language Resources & Evaluation Conference, Portorož, Slovenia, May 2016.

Chambers, Mary. 2007. Documentation of Kubokota. *Endangered Languages Archive*. London: SOAS. https://elar.soas.ac.uk/Collection/MPI99492.

Cialone, Claudia. 2016. Recordings of Bininj Kunwok bush culture, language, spatial navigation documentation (CC01). Digital collection managed by PARADISEC. [Open Access]. doi:10.4225/72/56E979E6F4200.

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

Coon, Jessica & Pedro Mateo Pedro. 2017. Chuj Oral Tradition Collection of Pedro Mateo Pedro and Jessica Coon. *The Archive of the Indigenous Languages of Latin America*. [Public Access]. https://ailla.utexas.org/islandora/object/ailla:254994.

DataCite Metadata Working Group. 2019. DataCite metadata schema documentation for the publication and citation of research data. Version 4.3. DataCite e.V. doi: 10.14454/7xq3-zf69.

Dewar, Amy. 2018. Documentation of Fakamae, a Polynesian outlier of Vanuatu. *Endangered Languages Archive*. London: SOAS. http://elar.soas.ac.uk/deposit/0487.

Franjieh, Michael. 2018. The languages of northern Ambrym, Vanuatu: An archive of linguistic and cultural material from the North Ambrym and Fanbyak languages. *Endangered Languages Archive*. London: SOAS. https://elar.soas.ac.uk/Collection/MPI1143013.

Franjieh, Michael. 2019. The languages of northern Ambrym, Vanuatu: A guide to the deposited materials in ELAR. *Language Documentation & Conservation* 13. 83–111. http://hdl.handle.net/10125/24849.

Gawne, Lauren. 2018. A guide to the Syuba (Kagate) Language Documentation Corpus. *Language Documentation & Conservation* 12. 204–234. http://hdl.handle.net/10125/24768.

Godfrey, John J., Edward C. Holliman, & Jane McDaniel. 1992. SWITCHBOARD: Telephone speech corpus for research and development. Proceedings of the IEEE International *Conference on Acoustics, Speech and Signal Processing* 1. 517–520.

Grant, Rebecca, Graham Smith, & Iain Hrynaszkiewicz. 2019. Assessing metadata and curation quality: A case study from the development of a third-party curation service at Springer Nature. *bioRxiv* 530691. doi:10.1101/530691.

Hildebrandt, Kristine, Tanner Burge-Buckley, & Jacob Sebok. 2019. Language documentation in the aftermath of the 2015 Nepal earthquakes: A guide to two archives and a web exhibit. *Language Documentation & Conservation* 13. 618–651. http://hdl.handle.net/10125/24914.

Hinton, Leanne. 2005. What to preserve: A viewpoint from linguistics. In *Native language preservation a reference guide for establishing archives and repositories*, 24–26. Department of Health and Human Services and Administration for Native Americans. http://www.aihec.org/ourstories/docs/NativeLanguagePreservationReferenceGuide.pdf.

Holton, Gary. 2012. Language archives: They're not just for linguists any more. In Seifart, Frank, Geoffrey Haig, Nikolaus P. Himmelmann, Dagmar Jung, Anna Margetts, & Paul Trilsbeek (eds), *Potentials of language documentation: Methods, analyses, and utilization*, 111–117. Honolulu: University of Hawai'i Press.

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

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

Jung, Dagmar & Nikolaus P. Himmelmann. 2011. Retelling data: Working on transcription. In Haig, Geoffrey L.J., Nicole Nau, Stefan Schnell, & Claudia Wegener (eds), *Documenting endangered languages: Achievements and perspectives*. Berlin: De Gruyter Mouton.

Kane, Stephanie. 2018. Emberá Collection of Stephanie Kane. *The Archive of the Indigenous Languages of Latin America*. [Public Access]. https://ailla.utexas.org/islandora/object/ailla:124412.

Kim, Soung-U. 2018. A multi-modal documentation of Jejuan conversations. *Endangered Languages Archive*. London: SOAS. https://elar.soas.ac.uk/deposit/0351.

Koshoffer, Amy, Amy E. Neeser, Linda Newman, & Lisa R. Johnston. 2019. Giving datasets context: A comparison study of institutional repositories that apply varying degrees of curation. 13(1). 15–34. http://dx.doi.org/10.2218/ijdc.v13i1.632.

Library of Congress. 2008. Library of Congress Encoded Archival Description Best Practices. https://www.loc.gov/rr/ead/lcp/lcp.pdf.

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

Love, Robbie, Claire Dembry, Andrew Hardie, Vaclav Brezina, & Tony McEnery. 2017. The spoken BNC2014: Designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics* 22(3). 319–344. doi:10.1075/ijcl.22.3.02lov.

Lovestrand, Joseph. 2017. Recording and archiving Barayin (Jalkiya) language data. *Endangered Languages Archive*. London: SOAS. https://elar.soas.ac.uk/Collection/MPI1035101.

Lule, Irene. 2019. Using our words: Description at the Harry Ransom Center. Panel: Transforming the archive: Increasing inclusivity through language. *Annual Meeting of the Society of American Archivists*. Austin, TX.

Maldonado, Diego Manuel, Rodrigo Villalba Barrientos, & Diego P. Pinto-Roa. 2016. Eñe'e: Sistema de reconocimiento automático del habla en guaraní. In *Simposio Argentino de Inteligencia Artificial (ASAI 2016)- JAIIO* 45.

Margery Peña, Enrique. 2002. Bocotá Collection of Enrique Margery Peña. *The Archive of the Indigenous Languages of Latin America*. [Public Access]. https://ailla.utexas.org/islandora/object/ailla:124420.

Mesh, Kate. 2017. Points of comparison: What indicating gestures tell us about the origins of signs in San Juan Quiahije Chatino Sign Language Dataverse. *Texas Data Repository Dataverse*. https://dataverse.tdl.org/dataverse/indicating.

Milligan, Ian. 2013. Illusionary order: Online databases, optical character recognition, and Canadian history, 1997-2010. *Canadian Historical Review* 94(4). 540–569.

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

Oard, Douglas, William Webber, David Kirsch, & Sergey Golitsynskiy. 2015. Avocado Research Email Collection. LDC2015T03. DVD. Philadelphia: Linguistic Data Consortium.

Oduor, Jane Akinyi Ngala. 2016. A preliminary documentation of the Okiek language of Kenya. *Endangered Languages Archive*. London: SOAS. https://elar.soas.ac.uk/-Collection/MPI1032020.

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

Pillai, Stefanie. 2011. Malacca Portuguese Creole: A Portuguese-based Creole. *Endangered Languages Archive*. London: SOAS. https://elar.soas.ac.uk/Collection/MPI130545.

Pride, Kitty & Leslie Pride. 2007. Chatino Collection of Leslie and Kitty Pride. *The Archive of the Indigenous Languages of Latin America*. [Public Access]. https://ailla.utexas.org/islandora/object/ailla:124459.

Putnam, Lara. 2016. The transnational and the text-searchable: Digitized sources and the shadows they cast. *The American Historical Review* 121(2). 377–402.

Riley, Jenn. 2017. Understanding metadata: What is metadata and what is it for? Baltimore: National Information Standards Organization. https://groups.niso.org/apps/group_public/download.php/17446/Understanding%20Metadata.pdf.

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

Schackow, Diana. 2014. Yakkha-Nepali-English dictionary. Documentation and grammatical description of Yakkha, Nepal. *Endangered Languages Archive*. London: SOAS. https://elar.soas.ac.uk/Record/MPI186554.

Schembri, Adam, Jordan Fenlon, Ramas Rentelis, Sally Reynolds, & Kearsy Cormier. 2013. Building the British Sign Language Corpus. *Language Documentation & Conservation* 7. 136–154. http://hdl.handle.net/10125/4592.

Sherzer, Joel. 1970a. Emberá words. Kuna Collection of Joel Sherzer. *The Archive of the Indigenous Languages of Latin America*. [Public Access]. https://ailla.utexas.org/islandora/object/ailla:255733.

Sherzer, Joel. 1970b. More Emberá words. Kuna Collection of Joel Sherzer. *The Archive of the Indigenous Languages of Latin America*. [Public Access]. https://ailla.utexas.org/islandora/object/ailla:255736.

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

Society of American Archivists. 2013. Describing archives: A content standard, 2nd edn. http://files.archivists.org/pubs/DACS2E-2013 v0315.pdf.

Sullivant, J. Ryan. 2019a. Niels Fock & Eva Krener Photographic Collection of a Cañari Village, Juncal, Cañar, Ecuador Finding Aid. *Archive of the Indigenous Languages of Latin America*. https://ailla.utexas.org/islandora/object/ailla:263281.

Sullivant, J. Ryan. 2019b. Chol Collection of Juan Jesús Vázquez Álvarez and Jessica Coon Finding Aid. *Archive of the Indigenous Languages of Latin America*. https://ailla.utexas.org/islandora/object/263305.

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

Thieberger, Nick. 2018. Texts and more texts: Corpora in the CoEDL. [Blog Post]. http://www.paradisec.org.au/blog/2018/02/texts-and-more-texts-corpora-in-the-coedl/.

Thieberger, Nick, Anna Margetts, Stephen Morey, & Simon Musgrave. 2016. Assessing annotated corpora as research output. *Australian Journal of Linguistics* 36. 1–21. doi:10.1080/07268602.2016.1109428.

Walker, Alan. 1975. Alan Walker's Sabu materials (AW2). Digital collection managed by PARADISEC. [Open Access]. doi:10.4225/72/56E97998EC9A4.

Whalen, D.H. & Joyce McDonough. 2019. Under-researched languages: Phonetic results from language archives. In Katz, William F. & Peter F. Assmann (eds.), *Routledge handbook of phonetics*. London: Routledge.

Wilkinson, Mark D., Michel Dumontier, Ijsbrand Jan Aalberberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C. 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene

van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, & Barend Mons. 2016. The FAIR Guiding Principles for scientific data management and stewardship. Scientific data, 3.

Ryan Sullivant
sullivant@austin.utexas.edu
orcid.org/0000-0002-3968-7693

Appendices

A. Finding Aid for the Chatino Collection of Ryan Sullivant

A finding aid following these guidelines has been prepared for the Chatino Collection of Ryan Sullivant, and is included in the supplementary materials as Finding Aid for Chatino Collection of Ryan Sullivant [25 pages].

B. Prevalence of text annotations of audio-visual media in AILLA and ELAR

A search of AILLA metadata was made using a non-public interface on November 11, 2018 to produce a list of all media objects along with their media type, original medium, and the identifier of the resource where they are found. A comma-separated values file of the underlying data supporting these figures is available as a 7MB CSV file (MediaTypes-AILLA-20181120.csv) in the supplementary materials. Of the 13,179 resources containing an audio-visual primary recording, 27.3% (3,593) also contained a text-based file with some kind of annotation of the recording (of the types "commentary", "transcription", "translation", "transcription & translation", "annotation", "guide", "interlinearization", "interpretation", or "context").

A search of ELAR folder metadata using the public interface was performed on November 20, 2018. 31 of ELAR's 508 collections are assigned the status "In-process" and are associated with folder objects. These folders may have been incomplete at the time of this study and could ultimately be curated with more annotation files than what is reflected in the figures below. ELAR's folder objects are associated with one or more "type" attributes specifying the kinds of media files they contain. To avoid double-counting the 10,327 bundles that contain both audio and video types, each of these facets were applied in turn, and the counts of the remaining attributes are reported in Table 2 below.

These figures show that 39.7% (7,724/19,451) of folders containing video files and 40.4% (25,237/62,449) of folders containing audio files also contained a file that is likely to contain some kind of text-based annotation (ELAN, Praat, Transcriber, Toolbox, FLEx, or Annotation).

C. User-generated genre labels in ELAR and TLA

As shown in Figure 3, an empty search of ELAR's catalog returns the entirety of items in its catalog. The user may click "more …" below the list of the six most frequent genre labels to find the entire list of genre labels and their frequency.

A list was made of all the genre labels and their frequencies on April 17, 2019. This list was copied into a 35 KB CSV spreadsheet (GenreLabels-ELAR-20190417.csv) available in the supplementary materials. Figure 4 shows the genre labels ordered from most to least frequent in a cumulative frequency plot. There are 1,579 distinct

labels (some differing only slightly). "Discourse" is the most commonly-used label with 6,515 uses (11% of all labels used), and the 11 most frequent labels account for about half of all labels used. Most labels are used much less often: the median label is used twice and 40.3% of labels (636/1,579) are unique.

Table 2. Counts of ELAR folder types by different facets

Type	Annotation	Count (bundle facet)	Count (bundle and video facet)	Count (bundle and audio facet)
Audio	no	62449	10327	62449
Video	no	19451	19451	10327
ELAN	yes	15473	7485	13358
Document	no	14236	3058	10762
Image	no	10511	1432	2281
Settings	no	6735	3738	6061
Other	no	2475	444	1591
Praat	yes	1029	121	612
Transcriber	yes	816	65	722
Lexical database	no	784	268	676
Toolbox	yes	153	30	148
FLEx	yes	105	21	67
Annotation	yes	3	2	3
–	no	2	1	1
Geographic data	no	1		
Total files		134223		
Total annotations			7724	25237
Total non-annotations			8941	21372

Figure 3. (left) ELAR search results for an empty string; (right) ELAR genre label counts

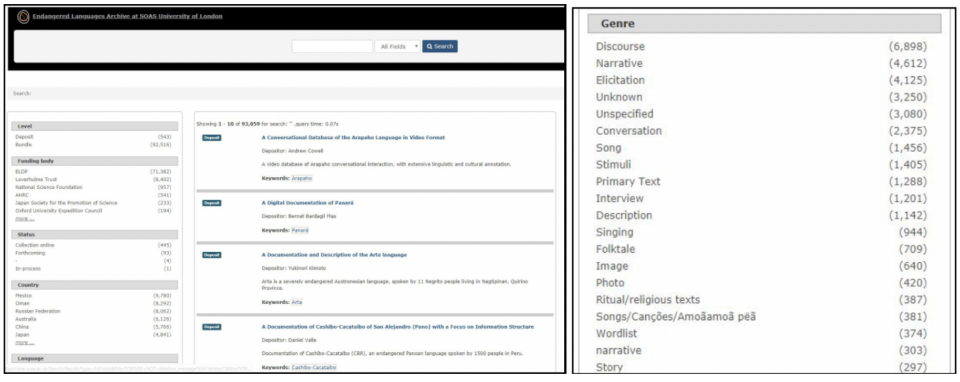
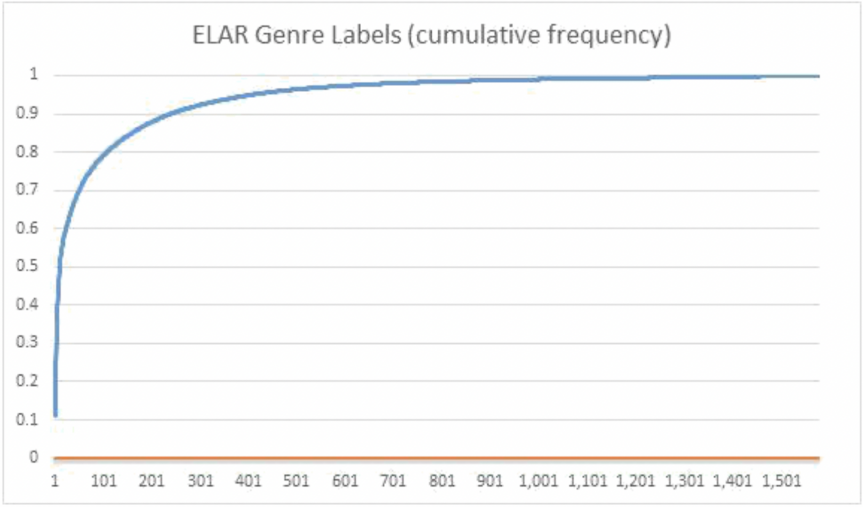
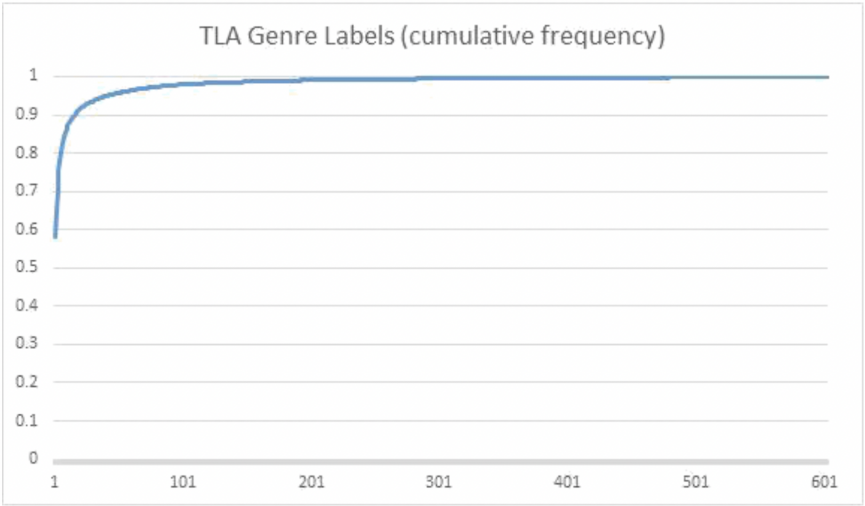


Figure 4. Cumulative frequency of ELAR genre labels



A review of genre labels at The Language Archive on February 6, 2019 was undertaken, and the labels and their frequencies are available in a 14 KB CSV spreadsheet (GenreLabels-TLA-20190206.csv). Figure 5 plots the genre labels ordered from most to least frequent with the frequencies of each plotted on a cumulative frequency plot. The most common label is "Unspecified", accounting for 42.4% of all labels used (42,963/101,402). If Unspecified labels are excluded, there are 604 labels used a total of 58,439 times. The most common specified label is "Discourse" accounting for 58% (33,921/58,439) of all label uses. The next most frequent label "Stimuli" accounts for 7.1% (4,121/58,439) of all labels used. The median label is used twice, and 47.8% (289/604) of labels are unique.

Figure 5. Cumulative frequency of TLA genre labels



Guide to the Chatino Collection of Ryan Sullivant

This collection guide is intended to be used to illustrate the checklist categories in the article "Archival Description for Language Documentation Collections". Therefore, the headings appear with the section numbers of that article and are divided into sections according to the three grades of priority with one exception: due to its length, the detailed contents list is placed at the end of this collection guide. Collection guides do not need to follow the order or style of presentation here–this is meant to be a checklist, not a collection guide template. Explanatory comments are given in square brackets.

Priority 1 items: Required information

Collection identifier (3.1.1.1)
Identifier: ailla:242619
URL: https://ailla.utexas.org/islandora/object/ailla:242619

Name and location of repository (3.1.1.2)
The Archive of the Indigenous Languages of Latin America (https://ailla.utexas.org)

Collection title (3.1.1.3)
The Chatino Collection of Ryan Sullivant
Spanish title: La Colección del Chatino de Ryan Sullivant

Collection dates (3.1.1.4)
Created: 2011–2018, bulk 2011–2012
Published: 2016
Last updated: 2018

Names and roles of contributors (3.1.1.5)
J. Ryan Sullivant, collector, depositor and transcriber
Flavia Mateo Mejía, speaker and translator
Celiflora Cortés Jiménez, speaker and translator
Cecilia Mejía López, translator
Modesta Martínez Mateo, translator
Adolfo Santiago Pérez, speaker
Florencia Mejía Cortés, speaker
Benita Gregorio Habana, speaker
Cenobia Jiménez Santiago, speaker
Elpidio Pérez Jiménez, speaker
Salomón Mejía Jiménez, speaker

Extent and scope of content (3.1.2)
The collection consists of

- 80 WAV audio recordings (about 20.5 hours) of spoken Tataltepec Chatino and Spanish
- 40 EAF documents containing Tataltepec Chatino transcriptions and Spanish translations of about 9.25 hours of audio recordings
- 31 ZIP containers folders containing 2871 WAV audio snippets in Tataltepec Chatino
- 12 TXT documents
- 10 PDF documents
- 3 CSV spreadsheets

Tataltepec Chatino audio recordings, transcriptions, and translations were created in Tataltepec de Valdés, Oaxaca, Mexico 2011–2012. The collected audio recordings are monologic narratives in Tataltepec Chatino and dialogues between a married couple. A number of folders contain recordings of the elicitations which underlie Sullivant's analysis of the grammar and lexicon of Tataltepec Chatino. Other documents, including reports and analytical papers (on Tataltepec Chatino as well as other languages) were created in Austin, Texas, United States between 2012–2018 as he completed his doctoral studies in linguistics at the University of Texas and began independent study of other Otomanguean languages of southwestern Oaxaca, Mexico.

The collected narratives and procedural texts focus on Tataltepec Chatino mythology, culture, procedural texts, and community and personal histories of the 20th and early 21st centuries. One document is a reproduction and rough translation of an 1829 travel diary that contains early transcriptions of a few words in Ixtapan Eastern Chatino.

Access and use restrictions (3.1.3)
To access most collection materials, one must create an AILLA user account and agree with AILLA's Conditions for Use of Archive Resources (https://ailla.utexas.org/site/rights/use_conditions). A few files containing published material were restricted with a temporary embargo in accordance with agreements between the author and the publisher. These files automatically revert to public access when the temporary embargo period expires (https://ailla.utexas.org/site/depositors/access).

Languages and scripts (3.1.4)
Audio recordings all contain Tataltepec Chatino (ISO 639-3:cta, Glottocode: tata1258) and frequently also contain Spanish. EAF transcription/translation files contain Tataltepec Chatino and Spanish, with occasional notes in English. Documents are in English or Spanish but may contain text in German, Teojomulco Chatino (Glottocode: teoj1234), Coatec Zapotec (ISO 639-3: zps, zpx; Glottocode: coat1242),

Tututepec Mixtec (ISO 639-3: mtu, Glottocode: tutu1243), Coatecas Altas Zapotec (ISO 639-3: zca, Glottocode: coat1244), and Ixtapan Eastern Chatino (ISO 639-3: cta).

Citation information (3.1.5)
This collection and its subparts should be cited in accordance with AILLA's Citation Guidelines (https://ailla.utexas.org/site/rights/citation).

Priority 2 items: Recommended information

Detailed contents list (3.2.1)
Found at the end of this document. Jump to the detailed contents list.

Biographical sketches and project history (3.2.2)
The production of most of this collection was supported by a Doctoral Dissertation Research Improvement Grant (BCS1065082) through the Documenting Endangered Languages Program of the National Science Foundation and is a continuation of work begun under the auspices of the Chatino Language Documentation Project (2006-2010) at the University of Texas (itself funded through grant MDP0153 from the Endangered Language Documentation Programme), through which Sullivant first began travelling to Tataltepec de Valdés and documenting Tataltepec Chatino.

The Chatino language of Tataltepec de Valdés, *Chá'knyá*, is one of three modern spoken Chatino languages of southwestern Oaxaca state. At the time of collection, this language boasted an estimated 500 speakers, nearly all of whom were bilingual in Spanish to greater or lesser degrees. Very few children were learning the language. For linguists at the time, Tataltepec Chatino was notable for having intricate systems of lexical tone, grammatical tone, and a complex system of verb inflection involving prefixes, as well as tone and vowel ablaut, and these features figure prominently in elicitation sessions led by Sullivant.

Flavia Mateo Mejía, Sullivant's main consultant and Tataltepec Chatino teacher, is a retired schoolteacher who learned Zenzontepec Chatino while posted to towns in that region. As a young woman, she also interacted with Leslie and Kitty Pride, missionary linguists from the Summer Institute of Linguistics who were long-time residents of Tataltepec.

Following the completion of the fieldwork grant, Sullivant returned to Austin to complete his dissertation. Since then he has produced some manuscripts based on studies of published writings of other Chatino, Zapotec, and Mixtec languages that are or were spoken in the area surrounding Tataltepec de Valdés.

[An alternate way to present biographical information could be as a chronology.]

J. Ryan Sullivant

2006	B.A. Spanish & B.S. Linguistics, Tulane University New Orleans
2009–2012	Linguistic fieldwork in Chatino communities of Oaxaca, Mexico
2015	Ph.D. Linguistics, University of Texas at Austin
2016–2019	Curator, Archive of the Indigenous Languages of Latin America

Arrangement and collection conventions (3.2.3)
Media in this collection are organized into 59 folders which are displayed alphabetically on AILLA's website. Some series of related materials have been created by prepending their titles with a prefix.

- 31 folders containing monologic narratives have titles beginning with "Text:" followed by either the title of the narration or the name of the speaker.
- 5 folders containing dialogues between two speakers have titles beginning with "Dialogue:" followed by the name of the speakers.
- 3 folders containing grammatical elicitation of verb forms have titles beginning with 'Verb inflection".
- 2 folders containing elicitations of translations of published sentences in studies of other Chatino languages appear begin with "Sentences:" followed by the name of their source.

The remaining 18 folders appear with a descriptive title (in the case of folders containing lexicogrammatical elicitations) or the title of the folder's main document. Text and Dialogue folders contain WAV recordings and their corresponding EAF files containing their transcriptions and translations into Spanish, if available.

Audio recording and annotation file names begin with a six-digit date code (e.g. "120330" = March 30, 2012), followed by a code for what kind of object the file is (e.g. "TXT" is a text, "INT" are recordings of interrogatives, "KIN" are kinship terms, etc.), then a code indicating the speaker being recorded. In zipped folders, many files are named according to the Spanish gloss of the Tataltepec Chatino word(s) recorded.

Tataltepec Chatino transcriptions in EAF files largely follows the practical orthography set out in Sullivant's dissertation (https://repositories.lib.utexas.edu/handle/2152/31493), with one systematic exception: the eight different tone sequences are indicated not with diacritics but with a short label separated from the end of the word by a slash (e.g. 'Chatino' is written *Cháʔknyá* in the dissertation orthography but would appear in the transcriptions as *chaʔ/a knya/a*). The short labels refer to impressionistic descriptions in Spanish of the tone.

Table 1. Tataltepec Chatino tone representation with diacritics and labels. S is a superhigh tone and tones in parentheses are underlyingly unlinked and only appear in certain contexts.

Tones	Diacritics	Labels	Tones	Diacritics	Labels
/∅/	CaCa	/ri	/(S)L/	⁰CaCà	/bm
/H/	CaCá	/a	/HL/	CaCā	/ar
/L/	CaCà	/bi	/SL/	CăCà; Câ	/ab
/(S)/	⁰CaCa	/rm	/S/	CaCă	/al

Priority 3 items: Optional information

Physical and technical access (3.3.1)
This is a born-digital collection in a digital repository; there are no physical materials to consult. The PDF, CSV, TXT, WAV and ZIP files in this collection may be read using commonly-available programs. The collection also contains materials in another less common format: the time-aligned transcriptions and translations are stored in XML text files with the extension EAF. These are meant to be read with the Eudico Linguistic Annotator (Elan). This open-source application may be downloaded from the Max Planck Institute for Psycholinguistics' The Language Archive (https://tla.mpi.nl/tools/tla-tools/elan/).

[No special equipment is needed to access these digital materials. As the collection is only available through a website, there is no need to discuss physical access to the materials.]

Keywords (3.3.2)
The AILLA repository that holds this collection does not have a keyword tagging function.

Accruals (3.3.3)
Documents may be infrequently added to this collection. No new audio-visual recordings, transcriptions or translations are expected. The digital photographs separated during appraisal may be added later pending additional description and clearance.

Custodial history (3.3.4)
The materials in this collection were created by Ryan Sullivant in collaboration with other named collaborators. Sullivant subsequently submitted the materials to AILLA. One folder contains the text of an 1829 travel diary by Eduard Mühlenpfordt (d. 1853) published as "Ausflug an die Ufer der Südsee in Frühjahre 1829" in the magazine *Das Ausland* in 1839. The depositor considers this text to be in the public

domain.

Appraisal, processing, conservation and migration (3.3.5)
Appraisal
Materials were appraised for their relevance, utility and merit. Duplicate backup copies, non-final versions, and audio and digital files of low quality and/or potential research value were excluded from the archival collection. A small collection of notebooks was deemed unsuitable for digitization and preservation due to poor penmanship, writing on both sides of sheets, and poor cataloging of their contents. A collection of digital photographs was not ingested since a significant amount of metadata recovery would be necessary to make them discoverable, and not everyone pictured in the photographs had granted permission for their image to be archived. The paper records used to produce the statistics in the folder "Language Vitality" were excluded due to privacy concerns.

Processing
Some subcollections of numerous very short audio files are contained within uncompressed ZIP containers. Files were renamed immediately prior to arrangement to normalize naming conventions and greatly shorten earlier verbose names. These original names may remain in the portions of EAF files indicating linked files.

Conservation
[Nothing to report as all materials included in the collection were born-digital.]

Migration
Materials were delivered to AILLA via an external hard drive. Files originally in Microsoft Word or Excel formats were converted to archival PDF/A or CSV formats, respectively, to facilitate long-term access.

Bibliographies, finding aids and published materials (3.3.6)
Two metadata spreadsheets, one for folders and one for media files, are publicly available within the collection:

- Metadata-Sullivant-resources.csv
  (http://ailla.utexas.org/islandora/object/ailla:252371)
- Metadata-Sullivant-media.csv
  (http://ailla.utexas.org/islandora/object/ailla:252372)

The recordings and annotation of Tataltepec Chatino in this collection are referenced in Sullivant's 2015 dissertation *The phonology and inflectional morphology of Cháʔknyá, Tataltepec de Valdés Chatino, a Zapotecan language*, which is also included in this collection. The materials in the folder "Language vitality" are referenced in Stéphanie Villard and J. Ryan Sullivant (2016), "Language Documentation

in two communities with high migration rates" in *Language Documentation and Revitalization in Latin American Contexts*. Some documents in the folder "Reintroducing Teojomulco Chatino" have been published as an article and online-only appendix (Sullivant, J. Ryan. 2016. Reintroducing Teojomulco Chatino. *International Journal of American Linguistics*. 82:4, 393–423).

[A bibliography of published works on Tataltepec Chatino or the Chatino people could also be included here. Any other descriptions of this collection could be included here if they existed.]

Related materials elsewhere (3.3.7)
Related archival materials at AILLA.

- Chatino Collection of Leslie and Kitty Pride (https://ailla.utexas.org/islandora/object/ailla:124459)
- Chatino Language Documentation Project Collection (https://ailla.utexas.org/islandora/object/ailla:124384)
- The Survey of Zapotec and Chatino Languages Collection (https://ailla.utexas.org/islandora/object/ailla:243980)
- Project for the Documentation of the Languages of MesoAmerica Collection (https://ailla.utexas.org/islandora/object/ailla:124491)
- The Zacatepec Chatino Documentation Project (https://ailla.utexas.org/islandora/object/ailla:124506)
- Chatino Documentation of Hilaria Cruz (https://ailla.utexas.org/islandora/object/ailla:124513)
- Chatino Landscape Collection (https://ailla.utexas.org/islandora/object/ailla:252384)
- Teotepec Chatino Collection of Justin McIntosh (https://ailla.utexas.org/islandora/object/ailla:124517)
- The Works of Thomas Cedric Smith Stark (https://ailla.utexas.org/islandora/object/ailla:124510)

Related archival materials at the Endangered Languages Archive:

- Documentation of Chatino (https://elar.soas.ac.uk/Collection/MPI113663)
- Documentation of Zacatepec Chatino Language (https://elar.soas.ac.uk/Collection/MPI185420)
- Documentation of Zenzontepec Chatino Language and Culture (https://elar.soas.ac.uk/Collection/MPI142111)

- Documenting Chatino Sign Language
  (https://elar.soas.ac.uk/Collection/MPI1031992)
- Documenting Sign Language Structure and Language Socialization in the San Juan Quiahije Chatino Signing Community
  (https://elar.soas.ac.uk/Collection/MPI1235744)
- Gesture, Speech and Sign in Chatino Communities
  (https://elar.soas.ac.uk/Collection/MPI1053087)
- Teotepec Chatino Language Documentation through History and Culture: An Integrated Approach
  (https://elar.soas.ac.uk/Collection/MPI43304)

Related archival materials at the American Philosophical Society:

- Franz Boas Anthropometric Data and Early Field Notebooks (Mss.B.B61.5)
  (https://diglib.amphilsoc.org/islandora/object/text:141696)
- Franz Boas Collection of Materials for American Linguistics 1882–1958
  (Mss.497.3.B63c)

Alternative available forms (originals, copies, realia) (3.3.8)
As this is a born-digital collection, there are no original materials. There are no copies of any unpublished materials in other repositories.

Notes on the production of the guide (3.3.9)
Description based on *Describing Archives: A Content Standard, Second Edition*, with AILLA controlled vocabularies. Written by Ryan Sullivant, the collector and depositor of the collection, in 2019.

Detailed content list (3.2.1)
The following list provides some information about each folder in the collection. The code beginning with "ailla:" is the identifier of each folder and is a hyperlink to the folder itself. Texts and dialogs have been transcribed and translated into Spanish unless otherwise indicated.

–Tataltepec Chatino Text–

ailla:243384	Text: Acapulco
Description:	Florencia Mejía talks about her trip to Acapulco, a boat ride to see statues of saints in the waters offshore, how a daring cliff diver retrieved a coin someone had thrown into the water, the heat in Acapulco.
Date:	2012-03-14
Genre:	Narrative
Contributor(s):	Florencia Mejía Cortés (speaker), Celiflora Cortés Jiménez (translator), Flavia Mateo Mejía (translator)
ailla:243321	Text: Adolfo Santiago 1
Description:	Adolfo Santiago talks about why he didn't go to school, and instead was sent to work as a hired hand for other people as a boy, picking corn, building corn cribs, how they grew chilies so that they'd have enough money to buy a donkey to haul their things, how he married his first wife who died leaving behind a small child, and his personal hardships since.
Date:	2011-07-23
Genre:	Narrative
Contributor(s):	Adolfo Santiago Pérez (speaker), Modesta Martínez Mateo (translator)
ailla:243324	Text: Adolfo Santiago 2
Description:	Adolfo Santiago talks about old methods of building a house, getting food, planting corn and keeping grackles away, other foodstuffs that are planted, plowing fields, changes in clothing and how clothing is made, selling chilies and tomatoes, old hairstyles, the flute and drum.
Date:	2011-07-29
Genre:	Narrative
Contributor(s):	Adolfo Santiago Pérez (speaker), Modesta Martínez Mateo (translator)
ailla:243340	Text: Benita Gregorio 1
Description:	Benita Gregorio talks about storing tortillas in a tray made from a gourd, drinking water from a gourd cup, daily routines, tending maize fields, working as a hired hand for other people.
Date:	2011-07-13
Genre:	Narrative
Contributor(s):	Benita Gregorio Habana (speaker), Cecilia Mejía López (translator)
ailla:243343	Text: Benita Gregorio 2
Description:	Benita Gregorio talks about storing tortillas in a tray made from a gourd, drinking water from a gourd cup, daily routines, tending maize fields, working as a hired hand for other people.

Continued from previous page

Date:	2011-07-18
Genre:	Narrative
Contributor(s):	Benita Gregorio Habana (speaker), Celiflora Cortés Jiménez (translator), Modesta Martínez Mateo (translator)
ailla:243351	Text: Cenobia Jiménez 1
Description:	Cenobia Jiménez speaks of her life in Tataltepec, the struggle to provide for oneself. How it was that she married her husband. How her children didn't learn Chatino. Her mother's work weaving. The upcoming feasts of St. James, which involves horse races, and All Saints', which is one of the largest feasts in the community.
Date:	2012-07-24
Genre:	Narrative
Contributor(s):	Cenobia Jiménez Santiago (speaker), Modesta Martínez Mateo (translator)
ailla:243354	Text: Cenobia Jiménez 2
Description:	Cenobia Jiménez speaks. This recording has not been transcribed or translated.
Date:	2012-07-30
Genre:	Narrative
Contributor(s):	Cenobia Jiménez Santiago (speaker)
ailla:243348	Text: Earthquake
Description:	Celiflora Cortés tells of an earthquake that rocked the town some eleven years prior.
Date:	2011-07-27
Genre:	Narrative
Contributor(s):	Celiflora Cortés Jiménez (speaker), Modesta Martínez Mateo (translator)
ailla:243357	Text: Elpidio Pérez 1
Description:	Elpidio Pérez talks about 1. how life used to be in Tataltepec and how life used to be rougher. 2. how his children couldn't go to school but learned to work in the fields.
Date:	2011-07-08
Genre:	Narrative
Contributor(s):	Elpidio Pérez Jiménez (speaker), Flavia Mateo Mejía (translator)
ailla:243362	Text: Elpidio Pérez 2
Description:	Elpidio Pérez talks about 1. salt 2. the Río Verde and the animal life in it 4. getting around the region before and after the road to Tataltepec from the coast was opened.
Date:	2011-07-11
Genre:	Narrative
Contributor(s):	Elpidio Pérez Jiménez (speaker), Celiflora Cortés Jiménez (translator)

Continued from previous page

ailla:243369	Text: Elpidio Pérez 3
Description:	Elpidio Pérez talks about 1. the customs surrounding the *mayordomía* traditions and the *cargo* positions in the town's civil-religious hierarchy 2. some positions in the civil-religious hierarchy 3. his own service in the cargo system 4. how he and others had to walk to Oaxaca for municipal business
Date:	2011-07-12
Genre:	Narrative
Contributor(s):	Elpidio Pérez Jiménez (speaker), Modesta Martínez Mateo (translator)
ailla:243300	Text: Feasts
Description:	Adolfo Santiago talks about the cycle of feasts, beginning with the upcoming feast on the 15th of October, which is followed in two weeks by the feast of All Saints', the feasts of the *mayordomos*, Christmas, Easter, St. Isidore, St. Joseph, St. Anthony.
Date:	2011-07-23
Genre:	Narrative
Contributor(s):	Adolfo Santiago Pérez (speaker), Modesta Martínez Mateo (translator)
ailla:243393	Text: Florencia Mejía 1
Description:	Florencia Mejía talks about her life as a child, and how things have changed since then. Her children in Mexico state, harvesting chilies and beans, picking coffee in Peñas Negras and El Zapote, her work as a potter making clay griddles, irrigation.
Date:	2011-07-20
Genre:	Narrative
Contributor(s):	Florencia Mejía Cortés (speaker), Flavia Mateo Mejía (translator)
ailla:243396	Text: Florencia Mejía 2
Description:	Florencia Mejía talks about old traditions associated with betrothal requests, how marriages used to be, her children in Mexico state, her thoughts on life in Mexico City.
Date:	2011-07-27
Genre:	Narrative
Contributor(s):	Florencia Mejía Cortés (speaker), Celiflora Cortés Jiménez (translator)
ailla:243399	Text: Florencia Mejía 3
Description:	Florencia Mejía talks about how music used to be played in the community, the authorities burning candles at various points where crosses have been put on Cerro Vidrio, Cerro Chinche, Cerro Agua Fría, Arroyo Arriba and Cerro de la Iglesia Vieja, *mayordomía* traditions, the feast of All Saints', her work as a potter making clay griddles which are strong

Continued from previous page

	enough that they don't break when dogs lay on them, making new copal censers for the All Saints' feast, other people who are skilled potters, the problems of eating out of metal or plastic pots, bringing gourd cups to the feast of the Virgin in Juquila.
Date:	2011-08-01
Genre:	Narrative
Contributor(s):	Florencia Mejía Cortés (speaker), Celiflora Cortés Jiménez (translator)
ailla:243402	Text: Florencia Mejía 4
Description:	Florencia Mejía talks about preparing a chicken for cooking different kinds of dishes that can be prepared, cooking *barbacoa* in an earth oven, how a whole beef is prepared when slaughtered for a feast, keeping animals and insects away from food.
Date:	2012-02-13
Genre:	Narrative
Contributor(s):	Florencia Mejía Cortés (speaker), Modesta Martínez Mateo (translator)
ailla:243405	Text: Florencia Mejía 5
Description:	Florencia Mejía talks about traditions about childbirth, planting the placenta near a spring, placing a candle near where the placenta was planted, with a candle placed three days later where the child was born, how these traditions have changed with the decrease in home births, illnesses due to marital infidelity, feeding an infant to the Holy Fire.
Date:	2012-02-13
Genre:	Narrative
Contributor(s):	Florencia Mejía Cortés (speaker), Modesta Martínez Mateo (translator)
ailla:243408	Text: Florencia Mejía 6
Description:	Florencia Mejía talks about an earthquake which destroyed Tataltepec's previous church, and how a new one was built. How people came together to outfit the new church. Wearing *soyates* (tight cloth belts) to avoid pain. Changing customs.
Date:	2012-04-09
Genre:	Narrative
Contributor(s):	Florencia Mejía Cortés (speaker), Celiflora Cortés Jiménez (translator)
ailla:243303	Text: Lime at the River
Description:	Adolfo Santiago talks about getting lime from burning rocks found near the river, and using lime to gather crustaceans living in the river.
Date:	2012-04-11
Genre:	Narrative
Contributor(s):	Adolfo Santiago Pérez (speaker)

Continued from previous page

ailla:242655	Text: Noah's ark
Description:	Adolfo Santiago tells the tale of Noah's ark.
Date:	2012-04-11
Genre:	Narrative
Contributor(s):	Adolfo Santiago Pérez (speaker)
ailla:243306	Text: Rabbit tricks coyote into swallowing a coyol fruit
Description:	Adolfo Santiago tells how the rabbit tricked the coyote into swallowing a *coyol* fruit.
Date:	2012-03-11
Genre:	Myth
Contributor(s):	Adolfo Santiago Pérez (speaker), Celiflora Cortés Jiménez (translator)
ailla:243411	Text: Salomón Mejía 1
Description:	Salomón Mejía tells of his early life, working with his grandfather and father tending cattle, working in the cargo system, having to hunt iguanas for the feast of the *carnestolenda*, his work on the Council of Elders, illnesses, his Catholic faith.
Date:	2011-07-28
Genre:	Narrative
Contributor(s):	Salomón Mejía López (speaker), Modesta Martínez Mateo (translator)
ailla:243414	Text: Salomón Mejía 2
Description:	Salomón Mejía tells part of the story of Jesus's life, and how various animals have the form they have because God punished them for mocking Jesus. The importance of work, invitations to help out at weddings and special events, questions about what will become of the Chatino recordings, other peoples' negative opinion of Chatino.
Date:	2011-08-02
Genre:	Narrative
Contributor(s):	Salomón Mejía López (speaker), Flavia Mateo Mejía (translator)
ailla:243381	Text: The founding of Tataltepec
Description:	Flavia Mateo tells a story of the founding of Tataltepec by a rich merchant woman from the Valley of Oaxaca named María who passed by the site of Tataltepec and noticed that it was good land, and suggested that the Chatinos of Cerro Agua Fría settle the plain. This story also explains the splitting up of the Tataltepec-Chatino-speaking people from the Zenzontepec-Chatino-speaking people to the north, since those people remained in the north having decided not to settle in the area of Tataltepec. This María is also the person who suggests that the Chatinos plant prickly pear cactuses to raise cochineal. It is because of this that Our Lady of the

Continued from previous page

	Assumption was chosen as the patroness of the town when the church was first built. [Tataltepec was known as Santa María Asunción Tataltepec before being re-named in honor of revolutionary hero Antonio de Valdés]. Flavia Mateo was told this story by her father Leonor Mateo.
Date:	2011-07-20
Genre:	Narrative
Contributor(s):	Flavia Mateo Mejía (speaker), Celiflora Cortés Jiménez (translator)
ailla:243387	Text: The gossipy magpie
Description:	Florencia Mejía tells a tale about the gossipy magpie.
Date:	2012-03-09
Genre:	Narrative
Contributor(s):	Florencia Mejía Cortés (speaker), Modesta Martínez Mateo (translator)
ailla:243309	Text: The man who switched places with a buzzard
Description:	Adolfo Santiago tells a tale about the man who switched places with the buzzard.
Date:	2012-03-11
Genre:	Narrative
Contributor(s):	Adolfo Santiago Pérez (speaker), Celiflora Cortés Jiménez (translator)
ailla:243378	Text: The north wind and the sun
Description:	Flavia Mateo offers three versions of the tale of the North Wind and the Sun. These recordings were made at the request of Ryan Sullivant, who intended to use one of these passages for a journal series that frequently uses this text as a sample. This is a translation of a Spanish telling of the tale and is not a traditional Chatino or Mesoamerican tale.
Date:	2012-06-28
Genre:	Narrative
Contributor(s):	Flavia Mateo Mejía (speaker and translator)
ailla:243312	Text: The Sun and the Moon
Description:	Adolfo Santiago tells a tale of the twins the Sun and the Moon.
Date:	2012-02-23
Genre:	Narrative
Contributor(s):	Adolfo Santiago Pérez (speaker), Modesta Martínez Mateo (translator)
ailla:243315	Text: The tale of Saint James
Description:	Adolfo Santiago tells a tale of St. James who attended a party disguised as a poor man and as a rich man to see the difference in how he would be treated.
Date:	2012-03-14
Genre:	Narrative

Continued from previous page

Contributor(s):	Adolfo Santiago Pérez (speaker), Celiflora Cortés Jiménez (translator)
ailla:243390	Text: The two compadres
Description:	Florencia Mejía tells a brief tale of two compadres.
Date:	2012-03-09
Genre:	Narrative
Contributor(s):	Florencia Mejía Cortés (speaker), Modesta Martínez Mateo (translator)
ailla:243318	Text: Why a rabbit appears on the face of the moon
Description:	Adolfo Santiago tells why a rabbit appears on the face of the moon.
Date:	2012-03-11
Genre:	Narrative
Contributor(s):	Adolfo Santiago Pérez (speaker), Celiflora Cortés Jiménez (translator)

–Tataltepec Chatino Dialogues–

ailla:243327	Dialogue: Adolfo Santiago and Florencia Mejía 1
Description:	Adolfo Santiago and Florencia Mejía have a free-ranging conversation. This recording has not yet been transcribed nor translated.
Date:	2012-03-30
Genre:	Conversation
Contributor(s):	Adolfo Santiago Pérez (speaker), Florencia Mejía Cortés (speaker)
ailla:243329	Dialogue: Adolfo Santiago and Florencia Mejía 2
Description:	Adolfo Santiago and Florencia Mejía talk about the people in town who would play music over speakers back when very few people had sound systems. Other people in town who would play musical instruments. Making soap. Making clothes. Making cane sugar. Viewing a solar eclipse from its reflection on the water in a gourd cup. An earthquake a number of years back. The dangerous times when armies of bandits roamed the countryside.
Date:	2012-04-11
Genre:	Conversation
Contributor(s):	Adolfo Santiago Pérez (speaker), Florencia Mejía Cortés (speaker), Celiflora Cortés Jiménez (translator)

Continued from previous page

ailla:243332	Dialogue: Adolfo Santiago and Florencia Mejía 3
Description:	Adolfo Santiago and Florencia Mejía talk about illnesses that have affected Tataltepec, criminal punishments, animals, the late Cayetana who made remedies, and other topics in this free conversation.
Date:	2012-06-25
Genre:	Conversation
Contributor(s):	Adolfo Santiago Pérez (speaker), Florencia Mejía Cortés (speaker), Modesta Martínez Mateo (translator)
ailla:243335	Dialogue: Adolfo Santiago and Florencia Mejía 4
Description:	Adolfo Santiago and Florencia Mejía talk about the sacred hallucinogenic San Juan mushroom, and its uses, funerary rites, the Reed Mat Bull (*wata yaka* or *toro petate*), *tepache*, animals found in the countryside, and other topics in this free-ranging conversation.
Date:	2012-07-01
Genre:	Conversation
Contributor(s):	Florencia Mejía Cortés (speaker), Adolfo Santiago Pérez (speaker), Modesta Martínez Mateo (translator)
ailla:243338	Dialogue: Adolfo Santiago and Florencia Mejía 5
Description:	Adolfo Santiago and Florencia Mejía have a free-ranging conversation. This recording has not yet been transcribed nor translated.
Date:	2012-07-09
Genre:	Conversation
Contributor(s):	Adolfo Santiago Pérez (speaker), Florencia Mejía Cortés (speaker)

–Tataltepec Chatino Lexical and Grammatical Elicitation–

ailla:243488	Adjective Reduplication
Description:	There is a marginal process of adjective reduplication in Chá?knyá. To indicate an intense degree of an adjective's meaning, an adjective is prefixed by a partial reduplication and suffixed with the particle *ka* /ka$^{(S)}$/. The form of the reduplicant minimally involves the reduplicand's initial consonant and a vowel, which for some speakers tends to be influenced by the vowel of the stem. For some speakers for some adjectives, the reduplication is total. This process appears marginal and most likely non-productive, and there is a large amount of disagreement between speakers about exactly what form the reduplicant will take and even if reduplication is possible for a given adjective.

Continued from previous page

Date:	2012-04-15
Genre:	Elicitation
Contributor(s):	Celiflora Cortés Jiménez (speaker), Flavia Mateo Mejía (speaker)
ailla:243602	Commands
Description:	Attempts were made to elicit imperatives, jussives, and hortatives in Cháʔknyá. These commands in Tataltepec Chatino appear to be based only on the potential mood form of the verb, unlike what has been found in Zenzontepec Chatino and Zacatepec Eastern Chatino where imperatives are variously formed with potential mood markers, completive aspect markers, and (in Zenzontepec Chatino at least) unique imperative markers.
Date:	2012-04-14
Genre:	Elicitation
Contributor(s):	Celiflora Cortés Jiménez (speaker), Flavia Mateo Mejía (speaker), Cecilia Mejía López (speaker)
ailla:243606	Inalienably possessed noun inflection
Description:	All Chatino languages have two kinds of person inflection strategies for nouns and relational nouns/prepositions/case markers. Alienably-possessed nouns indicate their possessor by following the noun with the possessor, which is preceded by the marker *jiʔín*. Inalienably-possessed nouns and a small set of terms indicating spatial relationships (most of which can be shown to derive from inalienably-possessed nouns) indicate their possessor by placing it immediately after the term, or in the case of a first-person singular or second-person singular possessor through a tone and/or vowel ablaut. These files are recordings of many of these words uninflected for possessor, inflected for a second-person singular possessor, and inflected for a first-person singular possessor.
Date:	2012-03-24
Genre:	Elicitation
Contributor(s):	Flavia Mateo Mejía (speaker)
ailla:243608	Interrogatives
Description:	A few of the interrogatives used to form content (wh-) questions in Cháʔknyá were elicited. When an object of a verb (which is preceded by the marker *jiʔín* in certain circumstances) is questioned, a typically Mesoamerican pied-piping with inversion occurs, and *jiʔín* follows the interrogative instead of preceding it or remaining in situ.
Date:	2011-07-22
Genre:	Elicitation
Contributor(s):	Flavia Mateo Mejía (speaker)

Continued from previous page

ailla:243615	Kinship and affinity terms
Description:	The Cháʔknyá community has a rich set of kinship and affinity terms, consisting mostly of native terms, but also showing evidence of Spanish influence in the expansion of *compadrazgo* affinity terms. The system would appear to be a Hawaiian system, with little lexical differentiation between lineal and collateral relatives. As in many languages, many of these terms are inalienably possessed, and some of the recordings here may include these terms elicited in an unpossessed form along with forms inflected for first-person singular and second-person singular forms.
Date:	2012-03-22
Genre:	Elicitation
Contributor(s):	Celiflora Cortés Jiménez (speaker), Flavia Mateo Mejía (speaker)
ailla:243490	Lexicon
Description:	A large collection of most non-verbal lexemes in Cháʔknyá. In order to determine the tones associated with a given word, after a word is given in citation form, it is then repeated preceded and the followed by a word (typically of tone set /(S)L/) whose tones allow for identifying both the presence of an unlinked superhigh tone on a stem, and the presence of a low or high tone which might have otherwise been misheard.
Date:	2011-07-13
Genre:	Elicitation
Contributor(s):	Flavia Mateo Mejía (speaker), Celiflora Cortés Jiménez (speaker), Modesta Martínez Mateo (speaker)
ailla:242845	Personal names
Description:	Some personal names in Cháʔknyá. Native naming traditions have largely been replaced by Spanish naming traditions among Tataltepec's Chatinos. A person's Cháʔknyá name is an adaptation of their Spanish name (which is legally required), though very commonly the form adapted into Cháʔknyá is not the full Spanish name but the hypocoristic form (e.g. Ch. *Lenchù* < Sp. *Lencho* < Sp. *Lorenzo*). Speakers appear to prefer monosyllabic names or disyllabic names derived from the stressed trochee of the Spanish name (i.e. Ch. *Teyà* < Sp. *Eleuteria* /e.leu.'te.ɾja/). Because of this, multiple Spanish names may correspond to a single Chatino name (e.g. *Stinù* corresponds to both *Justino* and *Faustino*.) Names, being foreign loans, typically belong to tone class /(S)L/, though other tone classes are also attested. For discussions of naming strategies in other Chatino languages, see: Cruz, Emiliana. 2011. Phonology, tone and the functions of tone in San Juan Quiahije Chatino. Ph.D. dissertation, The University of Texas at Austin. http://hdl.handle.net/2152/ETD-UT-2011-08-4280. Pp. 262–268.

Continued from previous page

Date:	2012-07-26
Genre:	Elicitation
Contributor(s):	Flavia Mateo Mejía (speaker)
ailla:243644	Toponyms
Description:	Names of locations in and around Tataltepec de Valdés.
Date:	2011-08-05
Genre:	Elicitation
Contributor(s):	Flavia Mateo Mejía (speaker), Celiflora Cortés Jiménez (speaker)

–Tataltepec Chatino Verb Inflection Elicitation–

ailla:243592	Verb inflection
Description:	A verb in Chá?knyá is obligatorily inflected for an aspect and/or mood category. All verbs are inflected for at least four categories, traditionally referred to as Completive, Progressive, Habitual, and Potential. A few verbs historically (and possibly synchronically) also inflect for a Stative aspect, and possibly an Imperative mood. Aspect/mood inflection is indicated in two ways: by an aspect/mood prefix (which may be null or a modification of the verb stem's initial consonant) and by a tone alternation. The files in this resource are various verbs elicited for each of the four common aspect/mood categories. Spanish preterites are used to elicit Completive Aspect verbs, the present progressive for Progressive Aspect verbs, the present tense for Habitual Aspect verbs, and either the morphological or periphrastic Spanish future for Potential Mood verbs. Additionally, temporal adverbs such as *laká* 'yesterday', *jwaniì* 'now', *lkaa tzaan* 'every day', and *lakeè* 'tomorrow' are elicited before the verb as a prompt for the speaker. To correctly identify the tones of the verb it is usually repeated preceded and the followed by a word (typically of tone set /(S)L/) whose tones allow for identifying both the presence of an unlinked superhigh tone on a stem, and the presence of a low or high tone which might have otherwise been misheard. Contents of zipped folders:
Date:	2011-07-06
Genre:	Elicitation
Contributor(s):	Flavia Mateo Mejía (speaker), Celiflora Cortés Jiménez (speaker)

Continued from previous page

ailla:243646	Verb inflection: 1st person singular
Description:	In addition to being inflected for (at least) one of four aspect and/or mood categories (Completive, Progressive, Habitual, and Potential). A verb in Cháʔknyá can also be inflected to indicate a first-person singular or second-person singular subject. First-person singular verbs will feature a nasalized non-high final, stressed vowel (nasalizing and lowering an underlying vowel as necessary) and most likely will also have undergone a tone replacement. The tone of a first-person verb is informed, but not entirely determined, by the tone of its corresponding verb stem inflected for aspect/mood but not person. The files in this resource are various verbs elicited in the first-person singular for each of the four common aspect/mood categories. Spanish preterits are used to elicit Completive Aspect verbs, the present progressive for Progressive Aspect verbs, the present tense for Habitual Aspect verbs, and either the morphological or periphrastic Spanish future for Potential Mood verbs. Additionally, temporal adverbs such as *laká* 'yesterday', *jwaniì* 'now', *lkaa tzaan* 'every day', and *lakeè* 'tomorrow' are elicited before the verb as a prompt for the speaker. To correctly identify the tones of the verb it is usually repeated preceded and the followed by a word (typically of tone set /(S)L/) whose tones allow for identifying both the presence of an unlinked superhigh tone on a stem, and the presence of a low or high tone which might have otherwise been misheard.
Date:	2011-07-06
Genre:	Elicitation
Contributor(s):	Flavia Mateo Mejía (speaker), Celiflora Cortés Jiménez (speaker)
ailla:243655	Verb inflection: 2nd person singular
Description:	In addition to being inflected for (at least) one of four aspect and/or mood categories (Completive, Progressive, Habitual, and Potential). A verb in Cháʔknyá can also be inflected to indicate a first-person singular or second-person singular subject. second-person singular verbs undergo a tone replacement. The tone of a second-person verb is almost entirely determined by the tone of its corresponding verb stem inflected for aspect/mood but not person. The files in this resource are various verbs elicited in the second-person singular for each of the four common aspect/mood categories. Spanish preterites are used to elicit Completive Aspect verbs, the present progressive for Progressive Aspect verbs, the present tense for Habitual Aspect verbs, and either the morphological or periphrastic Spanish future for Potential Mood verbs. Additionally, temporal adverbs such as *laká* 'yesterday', *jwaniì* 'now', *lkaa tzaan* 'every day', and *lakeè*

Continued from previous page

	'tomorrow' are elicited before the verb as a prompt for the speaker. To correctly identify the tones of the verb it is usually repeated preceded and the followed by a word (typically of tone set /(S)L/) whose tones allow for identifying both the presence of an unlinked superhigh tone on a stem, and the presence of a low or high tone which might have otherwise been misheard.
Date:	2011-07-11
Genre:	Elicitation
Contributor(s):	Flavia Mateo Mejía (speaker)

–Sentence Translation into Tataltepec Chatino–

ailla:243641	Sentences: Wardle
Description:	Elicitation of the example sentences from a work on another Chatino language. Wardle, Ed. 1976. Aspects of the structure of discourse in Nopala Chatino. http://www.sil.org/resources/archives/31009
Date:	2012-06-18
Genre:	Elicitation
Contributor(s):	Flavia Mateo Mejía (speaker)
ailla:243634	Sentences: Smith-Stark et al
Description:	Elicitation of the example sentences from a work on complementation strategies in another Chatino language. Smith-Stark, Thomas, Hilaria Cruz and Emiliana Cruz. 2008. Complementación en el chatino de San Juan Quiahije. Proceedings of the Conference on Indigenous Languages of Latin America-III. Organized by the Center for Indigenous Languages of Latin America (CILLA), Teresa Lozano Long Institute of Latin American Studies at the University of Texas at Austin.
Date:	2011-07-24
Genre:	Elicitation
Contributor(s):	Flavia Mateo Mejía (speaker)

–Documents based on Field Work on Tataltepec Chatino–

ailla:257327	A brief overview of tone in Cháknyá, Tataltepec de Valdés Chatino
Description:	Abstract of the paper: Chá?knyá, Tataltepec de Valdés Chatino (ISO 639-3: cta) has a tonal system wherein four tones (three level and one contour) are arranged in sequences of zero to two tones. Synchronically, the tone system is unusual in having a typologically unexpected set of levels (Low, High and Superhigh rather than Low, Mid and High) and the presence of a floating tone which never appears outside of the stem it is associated with. Diachronically, the tone system show signs of being a simplified version of the Proto-Coastal Chatino system which underwent a tonal apheresis, leading to both the fusion of certain etymological sets and a high frequency of singleton tone sequences rather than sequences formed of pairs of tones.
Date:	2016-04-29
Genre:	Document
ailla:242652	Language vitality
Description:	An attempt to identify the size of the speech community in Tataltepec de Valdés. Previous speaker counts were problematic as census data both fails to distinguish among Chatino languages and are reported by *municipio*, which in the case of Tataltepec includes both the town of Tataltepec de Valdés (where most every Chatino speaker speaks Cháknyá) and the equally-populous Santa Cruz Tepenixtlahuaca (where a variety of a different Chatino language is widely spoken).
Date:	2016
Genre:	Dataset
ailla:243783	The phonology and inflectional morphology of Chá?knyá, Tataltepec de Valdés Chatino, a Zapotecan language
Description:	This dissertation is a description of the phonology and inflectional morphology of an endangered indigenous language of Mexico stemming from a collaborative research project that places an emphasis on natural language and on describing a language on its own terms. The language described is Tataltepec Chatino (ISO 639-3: cta), a Zapotecan language spoken by fewer than 500 people only in the community of Tataltepec de Valdés in Mexico's Oaxaca state. The language has a complex system of tone in which tone sequences are the crucial morphological element rather than the constituent tones of the tone sequences. The tone system has a slightly peculiar inventory, with the level tones Low, High, and Superhigh rather than Low, Mid, and High in addition to a High-Low contour tone. The tonal system is also notable given the unlinked tone in two tone sequences which only surfaces in particular phonological contexts, but is never displaced from the word it is associated with, unlike

Continued from previous page

	canonical floating tones. The segmental phonology shows a language that permits a large number of often very complex onset clusters many of which violate the Sonority Sequencing Principle, but maintains tight restrictions on codas, allowing only a simple coda which can only be filled by one of two consonants in the language. Tataltepec Chatino also has interesting morphological features in its complex systems of verb aspect and person inflection which are instantiated by a system of prefixes and a system of complex paradigmatic alternations which only partially intersect. The language also has an unusual word I analyze as a "pseudoclassifier" which appears to serve some pragmatic functions of numeral classifiers while failing to do any lexical classification.
Date:	2015-05
Genre:	Thesis

Philological Research on Chatino, Zapotec, and Mixtec Languages

ailla:257329	A brief report on early 19th century Ixtapan Eastern Chatino
Description:	Abstract of the paper: A travel diary published in 1839 includes transcriptions of sixteen words in Ixtapan Eastern Chatino recorded in 1829. Though small, this is the earliest known sample of written Chatino, and shows both that a language very similar to contemporary Ixtapan Eastern Chatino was spoken around that site in the early 19th century, and that unstressed vowel loss–a process known to have affected nearly all Chatino topolects–was not complete, and may have been incipient in 1829.
Language(s):	Chatino, Eastern, Ixtapan
Date:	2015-09-03
Genre:	Document
ailla:244868	Reintroducing Teojomulco Chatino
Description:	A journal article and supplementary appendix on the Chatino language of Santo Domingo Teojomulco as transcribed by Francisco Belmar in 1902. The article is also available at the publisher's website (http://www.journals.uchicago.edu/toc/ijal/2016/82/4). This resource also includes the data of the appendix in a less formatted text file. Abstract: Belmar (1902) contains a word list of Teojomulco Chatino forms which is the only extant data from this language. I show that Teojomulco Chatino is the most divergent Chatino language, a sister to the most recent reconstructions of Proto-Chatino (Campbell 2013),

Continued from previous page

	and that it has traits not found elsewhere in Chatino, including Zapotec-like stem-initial prominence, nasalization of penultimate vowels, and a unique analogue of the Proto-Zapotecan *kwe= proclitic. Lexical evidence suggests that Teojomulco Chatino has undergone significant contact with Zapotec languages. Based on this data, I propose a revised picture of the historical development of the Chatino branch of the Zapotecan languages.
Language:	Chatino, Teojomulco
Date:	2016-10
Genre:	Article
ailla:257325	Report on Coatecas Altas Zapotec in 1911
Description:	Abstract of the paper: Boas (1912) contains a number of transcriptions of Coatecas Altas Zapotec. In spite of the difficulties presented by the handwriting, the symbols used, and the fading pencil marks, the data clearly show a great similarity to the Coatecas Altas Zapotec of Benton (2016a,b). This report offers my own interpretations and analysis of Boas (1912) and I note some of the differences I see between these forms and those reported by Benton (2016a,b).
Language:	Zapotec, Coatecas Altas
Date:	2016-08-09
Genre:	Document
ailla:257319	Tataltepec Chatino in Investigaciones sobre el idioma chatino
Description:	Abstract of the paper: Belmar (1902) is the earliest extant description of the Chatino languages, and contains transcriptions of words and phrases in a number of Chatino topolects. While the provenance of three relatively short lists is given, the bulk of the data comes from an unidentified topolect distinct from the three identified topolects. I will show that this unidentified topolect is Tataltepec Chatino (ISO 639-3 cta, henceforth TAT) based on the criteria used by Campbell (2013) to subgroup the modern Chatino languages and by other features identified from descriptions of Chatino topolects (Campbell 2014; Villard 2015; Sullivant 2015b) and my own field notes.
Language:	Chatino, Tataltepec de Valdés
Date:	2015-07-28
Genre:	Document
ailla:257331	The original text and English translation of Ausflug an die Ufer der Südsee in Frühjahre 1829
Description:	This document presents Mühlenpfordt's "Ausflug an die Ufer der Südsee im Frühjahre 1829" and a rough preliminary English translation. This article is a travel diary that was published in 1839 in *Das Ausland* across twenty serialized articles. "Ausflug" is notable since it contains the earliest

Continued from previous page

	known transcriptions of Ixtapan Eastern Chatino.
Language:	Chatino, Eastern, Ixtapan
Date:	2015-08-24
Genre:	Document
ailla:257321	Tututepec Mixtec data in Belmar (1902)
Description:	Abstract of the paper: Belmar (1902) gives word lists in Chatino, Zapotec, and Mixtec to demonstrate the Chatino languages' Zapotecan affiliation. The Mixtec words are from San Pedro Tututepec Mixtec and represent a small vocabulary of this language as it was spoken at the beginning of the 20th Century. This paper presents Belmar's transcriptions together with my own interpretations of them in order to make future researchers aware of this source of data and assist in diachronic and comparative studies of the Mixtec languages.
Language:	Mixtec, Tututepec
Date:	2015-07-15
Genre:	Document

Administrative Documents and Metadata

ailla:252369	AILLA Administrative Documents – Sullivant
Description:	Includes: metadata spreadsheets and depositor agreement.
Language(s):	Chatino, Tataltepec de Valdés, Chatino, Santo Domingo Teojomulco
Date:	2017-08-23
Genre:	Document