Wikitongues Archival Policy

Authors

Daniel Udell
Erik Pagan
Madhdvie Bhagwandeen
Melanie Chin
Theresa Condon
Marybeth Coscia-Weiss
Riley Ellison
Jennifer Grunenberg
Tomasz Gubernat
Jennifer Hayes
Alice Lambert
Sophia Lopez
Shayla Nastasi
Jakub Pieczarski
Ryann Riggs
Elizabeth Rowland
Veronica Smith
Katherine Stoutenburgh
Andrew Tripp
Samuel Weinberg
David Yang

Table of contents

Introduction **3**
- Mission 3
- Scope 3
- Content types 3
- Infrastructure 3

Metadata **3**
- Schema 4
  - Contributors 4
  - Oral histories 5
  - Oral history captions 10
  - Lexicons 10
  - Languages 13
  - Macrolanguages 15
  - Language Families (Top-level genealogy) 17
  - Glottocodes 17
  - Continents 18
  - Nations 19
  - Territories 20
  - Rights 21
  - Institutions 22
  - Writing Systems 23
  - Language Status 23
  - Publishers 23

Inventory and Storage **24**
  - Dropbox 24
  - External Harddrives 24

External Partners Storage 24

Maintenance 25

Intake 25

Technology Setup 25

Processing Oral History Donations 25

Processing Lexicon Donations 26

How to Classify Content by Languages 26

Backup schedule 26

Database 26

Hard Drives 27

External partners 27

Short-term maintenance 27

Intake 27

Pruning metadata 27

Changing and deprecating fields 27

Metadata and storage reconciliation 27

Long-term maintenance 27

Updating metadata 27

Stakeholders 28

Policy review 28

Introduction

Mission

All initiatives to sustain marginalized languages have one thing in common: media, so their language can be shared and taught, so Wikitongues is building a seed bank in every language in the world. Our long-term goal for every language is to collect up to eight hours of oral history videos and a 3,000-word dictionary, enough to support grassroots revitalization efforts.

Scope

Wikitongues selects, curates, and retains languages from around the world. In doing so, the project encompasses numerous types of files which may include, the files which have been stored in a repository or cloud-based service, any digital resources, and any institutional records relating to the functions of Wikitongues itself. The scope may expand over time, and Wikitongues has recognized this fact in the implementation of policy.

Content types

The content file type of each submission will vary depending on the creator, but upon submission, the item will be converted for use by Wikitongues to a standardized file format. An example may be converting video files (any type) to .mp4 files.

Infrastructure

Wikitongues manages metadata using Airtable and uses cloud hosting donated by Dropbox as primary servers. In accordance with the 3-2-1 rule, we back up all content on external hard drives and select content at external archives.

Metadata

Wikitongues stores and manages metadata related to language documentation and languages through Airtable. Schemas for each table in our database are listed below. Field names are not yet fully normalized.

Schema

Creators

This table contains the names and information of individuals who have added language documentation in our archive, either by producing or appearing in it. Personal data, such as contact information or precise location, are withheld from public views and may only be accessed by Wikitongues staff or core volunteers.

Field	**Description**
Identifier	Unique identifier assigned to each contributor. Formatted accordingly: **firstName_lastName_dateAdded**
Title	Creator's full name, formatted: **Last Name, First Name**
Date [Available]	The date of the creator's first contribution to our archive. Formatted: **YYYY-MM-DD**
Subject [Territories]	Sub-national territories where the creator lived as of our last contact with them. Links to Territories.
Subject [Country]	De jure or de facto nation-state where the contributor lived as of our last contact with them. Links to Nations.
Subject [Continent]	Continent where the contributor lived as of our last contact with them. Links to Continents.
Subject [Created: Oral Histories]	Language videos submitted by the creator; or videos of which the creator is the primary author. Links Oral Histories.
Subject [Speaker]	Language videos in which the creator appears. Links to Oral Histories.
Subject [Videographer]	Language videos that the creator helped record, without featuring in them or being the primary author. Links to Oral Histories.
Subject [Facilitator]	Language videos that the creator facilitated without featuring in them, recording them, or being the primary author. Links to Oral Histories.
Subject [Captioned]	Language videos that the creator helped caption or translate. Links to

	Oral History Captions.
Subject [Videos created count]	The number of videos submitted by the creator; or of which the creator is the primary author
Subject [Speaker count]	The number of videos in which the creator appears
Subject [Captioned count]	The number of videos captioned by the creator

Oral histories

Index of every language video in our archive. Metadata for content management can only be accessed by Wikitongues staff and core volunteers. Descriptive and technical metadata can be viewed separately or comprehensively.

Field	**Description**
Identifier	Unique ID assigned to each oral history. Formatted accordingly: **SpeakerFirstName_DateAdded_languageISOcode**. If the speaker name is unavailable, it is formatted: **Anonymous_DateAdded_languageISOcode**. Multiple ISO codes and speaker names are joined by a dash.
Title	Semantic title for the video. With exception, formatted: **[speaker(s)] speaking [language(s)]**. If the speaker name is not available, the title is formatted: **speaking [language(s)]**.
Creator	Primary video author. Links to the Contributors table
Description	Description of the video. Some are written by Wikitongues contributors; others are pulled from Wikipedia
Subject [Language: Genealogy]	The highest-level language families to which the featured languages pertain. Links to the Language Families table.
Subject [Language: Continent of Origin]	Continents of origin for the languages featured in the video. Links to the Continents table
Subject [Language: Nation of Origin]	De jure or de facto nation-states where the languages featured originate. This field links to the Nations table

Subject [Speaker Gender]	Identified gender of the creators featured in the video.
Creator [Speakers]	Creators featured in the video. Links to Creators.
Creator [Caption Authors]	Creators who helped caption or translate the video. Links to Creators.
Creator [Videographer]	Creators who helped produce the video without being its primary author or featuring in it; i.e., holding the camera, editing, etc. Links to Creators.
Creator [Facilitator]	Creators who facilitated the video without recording it, appearing in it, or being its primary author; i.e., arranging the interview. Links to Creators.
Date Created	Date the final, edited video file was created
Type	Media type: moving image or sound
Format	File type and resolution
Language names	Primary English-language name for the languages featured in the video
Languages: Speaker preferred names	The contributors' preferred names of the languages featured in the oral history, if different from the Language names above.
Languages: ISO code (639-6)	ISO 639-6 codes for the languages featured in the oral history
Languages: Glottocode	Glottocode for the languages featured in the oral history
Languages: Dialect Glottocode	Glottocode for the unique dialectal variety featured in the video, if applicable. Links to the Glottocodes table.
Languages: Macrolanguage ISO code	ISO code for the macrolanguages to which the oral history's featured languages pertain, if applicable. Links to the Macrolanguages table
Caption Languages	Primary English-language names of the languages of the oral history's captions
Caption Languages: ISO Code (639-3)	ISO 639-3 Codes for the languages of

	the oral history’s captions
Caption Languages: Glottocode	Glottocodes for the languages of the oral history’s captions. Links to the Glottocodes table.
Caption File Identifier	Unique ID for each oral history caption file. Links to Oral History Captions.
Caption File Links	Links to the oral history’s caption files
Coverage: Video Nation	De jure or de facto nation-state where the video was recorded. This field links to the Nations table
Coverage: Video Territory	Sub-national territories (i.e., U.S. states, Canadian provinces, etc) where the video was recorded. This field links to the Territories table
Coverage: Distribution	The extent to which the oral history is publicly available: on Wikitongues platforms only, or on Wikitongues external partner platforms
Coverage: Dropbox Link	Link to the finished video file on Dropbox. This only applies to videos that have been reviewed and edited.
Rights	The oral history license, chosen by the primary contributor: usually CC-by-NC 4.0, CC-by-SA 4.0, or Protected Copyright, but other licenses are accepted. Links to the Rights table.
Publisher	The oral history’s original publishing institution: either Wikitongues, or one of our partner organizations.
Date Received	The date this record was created.
Encoded Data	
Tagged Data	
Duration	Duration length of the final, edited oral history file.
Format T	File type of the final, edited video file.
Format Profile	
Codec ID	
File Size	Size of the final, edited oral history file

Format Info	
Format Settings	
Format Settings CABAC	
Format Settings ReFrames	
Codec ID/Info	
Bit rate	Bit rate of the final, edited oral history file
Width	If the oral history file is video, resolution width of the final, edited video file
Height	If the oral history file is video, resolution height of the final, edited video file
Display Aspect Ratio	If the oral history file is video, aspect ratio of the final, edited video file
Frame Rate	If the oral history file is video, frame rate of the final, edited video file
Standard	
Color Space	
Chroma Subsampling	
Bit Depth	
Scan Type	
Bits (Pixel*Frame)	
Stream size	
Color range	
Color primaries	
Transfer characteristics	
Matrix coefficients	
Codec configuration box	
Format audio	
Format/Info audio	
Bit Rate Audio	

Bit rate mode audio	
Codec ID Audio	
Channel(s)	
Channel layout	
Compression mode	
Sampling rate	
Stream size audio	
Subjects Reference ID: Ethnologue	External reference confirmation that the oral history's featured languages have been attested. Links to the Languages table
Quality [Thumbnail]	
Quality [Aspect ratio]	Frame dimension category, either *portrait* or *landscape*.
Quality [Stability]	The stability of the frame, either *1 - Jumpy, unable to see*, *2 - moves a lot*, *3 - some movement*, *5 - fairly static*, *8 - no camera movement, tripod*
Quality [Lighting]	The quality of lighting, either *1 - Unable to see face*, *2 - Difficult to see*, *3 - Okay*, *5 - Well-lit*, *8 - Professional lighting*
Quality [Distractions]	Interruptions in the video or audio that distract from the contents. Either *1 - Full of awkward jumps, many distractions*, *2 - A lot*, *5 - A little, but not bothersome*, *8 - No distractions*
Quality [Signing space]	The amount of visible physical range for videos of sign languages. Either *Yes*, *No*, or *Not relevant*
Quality [Audio]	The quality of audio. Either *1 - Distorted, unable to hear*, *2 - Scratchy, difficulties*, *3 - Quiet at times*, *5 - Overall good*, and *8 - Professional audio*
Quality [Background noise]	The intensity of background noise. *1 - Unable to hear/focus*, *2 - Extremely distracting*, *3 - Annoying*, *5 - A little at times, but not distracting*, *8 - No background noise*.
Quality [Duration]	Category of duration. Either *Less than 1 minute*, *1-3 minutes, 3-5 minutes*, *5-7 minutes*, or *7+ minutes*.

Quality [Captions count]	The number of languages in which the video is captioned.
Captions flagged	Automatic flag for high-quality videos that are uncaptioned or under-captioned.
Quality [Overall]	Aggregate documentation quality score of the video or audio file.

Oral history captions

Indexes of every caption set for Wikitongues oral histories. Field names are not yet normalized.

Field	**Description**
Identifier	Unique ID assigned to each oral history. Formatted accordingly: **SpeakerFirstName_DateAdded_featuredLanguageISOcode_captionLanguageISOcode**. Multiple ISO codes and speaker names are joined by a dash.
Description	Raw, unformatted text transcript of the captions, if available
Language [ISO 639-3]	ISO 639-3 code for the captions language. Links to Languages.
Language [Glottocode]	Glottocode for the captions language. Links to Glottocodes.
Language [English Name]	Primary English-language name for the captions language
Source [Video]	The oral history that the caption set transliterates. Links to Oral Histories.
Source [Text File]	Link to the formatted caption file on Dropbox
Source [Unformatted Text File]	Unformatted text document (pdf, docx, etc) of the captions, if available
Creator	Author of the captions
Format [Medium]	
Format [Extent]	
Misc. / Notes	Loose notes about the caption set

Lexicons

Index of every lexicon in our archive.

Field	Description
Identifier	Unique ID assigned to each lexicon. Formatted accordingly: **Contributor_DateAdded_SourceLanguage--TargetLanguage**
Title	Title of the lexicon. If none specified, defaults to the source language name.
Creator	Contributor who submitted the lexicon or is the primary author.
Description	Text description of the lexicon. Information may include an abstract, nation or territory of origin, etc.
Date [Created]	The date the lexicon was created
Subject [General]	Library of Congress data category. Automatically formatted accordingly: **sourceLanguages------TargetLanguages**. Multiple values separated by commas
Subject [Source Language: Genealogy]	The highest-level language families to which the featured languages pertain. Links to the Language Families table.
Subject [Source Language: Continent]	Continents of origin for the lexicon's source language. Links to the Continents table
Subject [Source Language: Nation]	De jure or de facto nation-state for the lexicon's source language. This field links to the Nations table
Language [Source]	Primary English-language name of the lexicon's source language.
Language [ISO 639-3]	ISO 639-3 Code for the lexicon's source language. Links to the Languages table
Language [Source: Glottocode]	Glottocode for the lexicon's source languages. Links to the Glottocodes table.
Language [Source Dialect: Glottocode]	Glottocode for the unique dialectal variety of the lexicon's source

	language, if applicable. Links to the Glottocodes table.
Language [Source Macrolanguage: ISO 639-3]	ISO code for the macrolanguages to which the lexicon's source language pertains, if applicable. Links to the Macrolanguages table
Language [Target]	Primary English-language names of the lexicon's target languages.
Language [Target: ISO 639-3]	ISO 639-3 Code for the lexicon's target languages. Links to the Languages table
Language [Target: Glottocode]	Glottocode for the lexicon's target languages. Links to the Languages table
Language [Target Dialect: Glottocode]	Glottocode for the unique dialectal variety of the lexicon's target languages, if applicable. Links to the Glottocodes table.
Language [Target Macrolanguage: ISO 639-3]	ISO code for the macrolanguages to which the lexicon's target languages pertain, if applicable. Links to the Macrolanguages table
Publisher	The entity responsible for making the resource available. In the case of an object that existed in another form before being digitized, the publisher of the earlier form may be entered.
Format [Medium]	The material or physical carrier of the resource.
Format [Extent]	The size or duration of the resource.
Type	Media type; e.g., *text*
Type [Document: LC]	Library of Congress data category. Either *Dictionaries* or *Polyglot glossaries, phrase books, etc*
Type [Document Category: LC]	Library of Congress data category. Currently, all records are defined as *idioms*; this could change
Coverage [Nation]	De jure or de facto nation-state where the lexicon was created. This field links to the Nations table
Coverage [Territory]	Sub-national territories (i.e., U.S. states, Canadian provinces, etc) where the video was recorded. This field links

	to the Territories table
Coverage [Dropbox]	Link to the lexicon file on Dropbox, if hosted by Wikitongues.
Coverage [Web]	Link to the lexicon file on an external platform, if also hosted elsewhere.
Source	DCMI defines Source as: Where the content originally delivered from or a resource that is related intellectually to the described content. The Source element may consist of a combination of elements such as free text combined with a formal identification system (such as an ISBN to describe a book or journal). Whenever possible, include a unique standard identifier such as an ISBN, ISSN, or LC call number.
Relation [Is Version Of]	DCMI defines Is Version Of as: the described resource is a version, edition, or adaptation of the referenced resource. Changes in version imply substantive changes in content rather than different format. Here most likely used to reference a physical original of a digitized resource.
Relation [Is Part Of]	DCMI defines Is Part Of as: The described resource is a physical or logical part of the referenced resource (such as a chapter of a book or a part of a webpage). a URI, for linking directly to the other resource.
Rights	The lexicon's license, chosen by the primary contributor: usually CC-by-NC 4.0, CC-by-SA 4.0, or Protected Copyright, but other licenses are accepted. Links to the Rights table.

Languages

Index of every attested language, as listed by the ISO 639-3:2007 code set, as well as attested languages that are presently excluded from ISO.

Field	Description
Identifier	Unique identifier for each language. If

	a three-letter ISO 639-3 language code is available, we use it. If the language is unclassified per ISO, we create a custom four-letter code, prefixed with the letter 'w'.
Language [Ethnologue]	Standardized, English-language name for the language, as featured in Ethnologue and substantiated by academic literature.
Language [Dialectal/Alternative]	Array of other names for the language, including names specific to dialectal and regional variants
Language [Glottocode]	Official Glottocode for language, if available. Links to Glottocodes.
Language [Macrolanguage]	ISO code for the macrolanguage to which the language pertain's, if applicable. Links to Macrolanguages.
Description	Generated prose description of the language. Contains location origin.
Subject [Territories]	Sub-national territories (i.e., U.S. states, Canadian provinces, etc) where the language originated. Links to Territories.
Subject [Nation of Origin]	De jure or de facto nation-state where the language originated. Links to Nations.
Subject [Continent of Origin]	Continental region of the language's origin. Links to Continents.
Subject [Writing System]	Writing system predominantly used by speakers of the language. Links to Writing Systems.
Subject [Genealogy]	Top-level genealogy of the language. Links to Language Families.
Subject [Typology]	Notes on the language's typology.
Subject [Status: EGIDS]	Language vitality according to the Expanded Graded Intergenerational Disruption Scale (EGIDS). Links to Language Status.
Subject [Institutions]	International institutions and organizations that use the language as a working or official language. Links to Institutions.

Subject [Official Recognition: National]	De jure or de facto nation-states where the language is official. Links to Nations.
Subject [Official Recognition: Regional]	Sub-national territories (i.e., U.S. states, Canadian provinces, etc) where the language is officially recognized. Links to Territories.
Relation [Is Required By: Oral Histories]	Unique IDs of oral history videos featuring this language. Links to Oral Histories.
Relation [Is Required By: Video Captions]	Unique IDs of caption files featuring this language. Links to Oral History Captions.
Relation [Is Required By: Lexicon source]	Lexicons for which this language is the source language. Links to Lexicons.
Relation [Is Required By: Lexicon target]	Lexicons for which this language is the target language. Links to Lexicons.
Relation [Is Required By: External Resources]	Externally-hosted resources for this language.
Reference ID [Language Archives]	Link to the OLAC page about this language, if available.
Reference ID [Ethnologue URL]	Link to the Ethnologue page about this language, if available.
Reference ID [ISO]	Link to the SIL-ISO page about this language, if available.
Reference ID [Wikipedia]	Link to the Wikipedia page about this language, if available.

Macrolanguages

Index of every classified macrolanguage as listed by the ISO 639-3:2007 code set. Broadly speaking, a macrolanguage is a group of closely related languages bound together by an overarching cultural identity, such as Arabic.

Field	**Description**

Identifier	Unique identifier for each macrolanguage. If a three-letter ISO 639-3 language code is available, we use it. If the language is unclassified per ISO, we create a custom four-letter code, prefixed with the letter 'w'.
Language [Macro: Ethnologue]	Standardized, English-language name for the macrolanguage.
Language [Individual]	Unique IDs of individual languages that pertain to this macrolanguage. Links to the Languages table
Description	Generated prose description of the language. Contains location origin.
Subject [Territories]	Sub-national territories (i.e., U.S. states, Canadian provinces, etc) where the macrolanguage originated. Links to Territories.
Subject [Nation of Origin]	De jure or de facto nation-state where the macrolanguage originated. Links to Nations.
Subject [Continent of Origin]	Continental region of the language's origin. Links to Continents.
Subject [Genealogy]	Top-level language family to which the macrolanguage pertains. Links to Language Families.
Subject [Institutions]	International institutions and organizations that use the macrolanguage as a working or official language. Links to Institutions.
Subject [Official Recognition: National]	De jure or de facto nation-states where the macrolanguage is official. Links to Nations.
Subject [Official Recognition Regional]	Sub-national territories (i.e., U.S. states, Canadian provinces, etc) where the macrolanguage is officially recognized. Links to Territories.
Relation [Is Required By: Oral Histories]	Unique IDs of videos recorded in languages within the macrolanguage. Links to Oral Histories.
Relation [Is Required By: Lexicon Source]	Lexicons in which a language within the macrolanguage is the source language. Links to Lexicons.

Relation [Is Required By: Lexicon Target]	Lexicons in which a language within the macrolanguage is the target language. Links to Lexicons.
Relation [Is Required By: External Resources]	Externally-hosted resources for languages within the macrolanguage.
Reference ID [Language Archives]	Link to the OLAC page about this macrolanguage, if available.
Reference ID [Ethnologue]	Link to the Ethnologue page about this macrolanguage, if available.
Reference ID [ISO]	Link to the SIL-ISO page about this macrolanguage, if available.

Language Families (Top-level genealogy)

Index of every top-level language family.

Field	Description
Identifier	The standardized, English-language name of the language family.
Description	Generated prose description of the macrolanguage. Contains individual language and macrolanguage counts.
Subject [Languages]	Unique IDs for languages within the language family. Links to Languages.
Subject [Macrolanguages]	Unique IDs for macrolanguages within the language family. Links to Macrolanguages.
Subject [Language count]	The number of languages with the language family.
Subject [Macrolanguage count]	The number of macrolanguages that pertain to this language family.

Glottocodes

Index of every classified language, as listed by the Max Planck Society's Glottocode code set. While ISO 639-3:2007 lists languages only, Glottocode categorizes speech varieties up and down the classification tree, including dialectal varieties.

Field	Description

Identifier	Unique Glottocode as assigned by the the Glottolog project.
Language [English]	The standardized, English-language name for the language.
Language [ISO 636-3]	The corresponding ISO code for the language. Links to Languages.
Language [Langoid]	The standardized, English-language name for the language, macrolanguage, or dialectal variant.
Subject [Classification]	The classification level—dialect, language, etc—of the Langoid.
Subject [Macroarea]	Cultural-geographic area of the Langoid
Relation [Is Required By: Oral Histories]	Oral history videos in which the Langoid dialect variant is featured. Links to Oral Histories.
Relation [Is Required By: Oral Histories: Dialect]	Oral history videos in which the Langoid language variant is featured. Links to Oral Histories.
Lexicon Source Language: Dialect	Lexicon document in which the Langoid dialect variant is the source language. Links to Lexicons.
Lexicon Target Language: Dialect	Lexicon document in which the Langoid dialect variant is the target language. Links to Lexicons.
Reference ID [Glottolog]	Link to the Glottolog page about this macrolanguage, if available.

Continents

Index of every top-level continental region.

Field	**Description**
Identifier	Name of the continental region
Subject [Nations]	De facto or de jure nation-states in the continental region. Links to Nations.
Subject [Territories]	Sub-national territories (U.S. states, Canadian provinces, etc.) in the continental region. Links to Territories
Subject [Territory Count]	Number of sub-national territories in

	the cont
Subject [Nation Count]	Number of de jure or de facto nation-states in the continental region
Subject [Language Count]	Number of languages that originated in the continental region
Reference ID [Wikipedia]	The Wikipedia article about the continental region, if available.
Reference ID [M49]	The M49 entry about the continental region, if available.

Nations

Index of every de jure nation-state, including all full and partial United Nations members, and de facto independent states, such as Taiwan.

Field	**Description**
Identifier	Standardized, English-language name of the jure or de facto nation-state.
Title [Common]	Common English-language name of the nation-state.
Title [Official]	Official English-language name of the nation-state as listed by the nation-state's government.
Creators	Creators who are based in the nation-state. Links to Creators.
Description [Continent]	Continental region of the nation-state. Links to the Continents table.
Description [Language Count]	Number of languages that originated in the nation-state's modern territory.
Description [Language Count Details]	Description of language count, such as which languages could be considered at-risk, as displayed on a pre-paywall version of Ethnologue.
Description [Immigrant Languages]	Notes on prominents diasporic language communities in the nation-state, as displayed on a pre-paywall version of Ethnologue.
Languages [ISO 639-3]	ISO 639-3 codes for the languages that originated in the nation-state. Links to the Languages table.

Subject [Official Languages ISO 639-3: National]	ISO 639-3 codes of the nation-state's official languages. Links Languages.
Subject [Official Languages English: National]	Standardized, English-language names of the nation-state's official languages.
Subject [Official Languages ISO 639-3: Regional]	ISO 639-3 codes of the nation-state's regionally official languages. Links to Languages.
Subject [Official Languages English: Regional]	Standardized, English-language names of the nation-state's regionally official languages.
Subject [Territories]	The nation-state's sub-national territories. Territories from federal countries and geographically expansive countries are included. Links to Territories.
Relation [Is Referenced By: Oral Histories]	Video oral histories recorded in the nation-state. Links to Oral Histories.
Relation [Is Referenced By: Lexicons]	Lexicon documents recorded in the nation-state. Links to Lexicons.
Subject [ISO 3166]	ISO 3166 code for nation-state.
Subject [Macrolanguages: Indigenous]	Macrolanguages that originated in the nation-state's modern territory. Links to Macrolanguages.
Subject [Macrolanguages: Official]	Macrolanguages that are nationally official in the nation-state. Links to Macrolanguages.
Subject [Macrolanguages: Official: Region]	Macrolanguages that are regionally official in the nation-state. Links to Macrolanguages.
Reference ID [Ethnologue]	References about the nation on Ethnologue, if available.
Reference ID [LOC Naming Authority	References about the nation from the Library of Congress, if available.

Territories

An index of sub-national territories from federations, such as the United States, and geographically expansive, but still centralized, countries, such as China.

Field	**Description**
Identifier	Unique ID assigned to the territory. Formatted: **Territory-Name_CountryCode**
Title	The territory's common English-language name.
Description [Type]	Type of sub-national territory: Special Administrative Region, Federated Entity, Autonomous Country, Autonomous Region, Capital Territory, Province, or County.
Description [Sovereignty]	Sovereign (nation-)state to which the territory pertains. Links to Nations.
Description [Sovereignty: ISO 3166]	Two-letter country code of the nation-state to which the territory pertains.
Description [Continent]	Continental region to which the territory pertains. Links to Continents.
Description [Language Count]	Number count of the languages predominantly spoken in the territory.
Creators	Creators who lived in the territory at the time of our last contact with them. Links to Creators.
Subject [Official Languages: English]	List of languages with official recognition in the territory.
Subject [Official Languages: ISO 639-3]	ISO 639-3 codes for the languages predominantly spoken in the territory. Links to Languages.
Subject [Macrolanguages]	Macrolanguages with official status or recognition in the territory. Links to Macrolanguages.
Subject [Oral Histories]	Oral histories that were recorded in the territory. Links to Oral Histories.
Subject [Lexicons]	Lexicon documents created in the territory. Links to Lexicons.
Reference ID	U.S. Library of Congress reference

	page for the territory, if available.

Rights

Index of licenses of video oral histories and lexicon documents. Field names, order, and public archival view are not yet normalized.

Field	Description
Identifier	The shorthand name of the license.
Title	The full legal name of the license
Description	Full text of the license, its applications and attribution information
Description [Abbreviated]	Summary of the license
Description [Wikitongues]	Official text for Wikitongues platforms to summarize the language in question
Subject [Oral Histories]	Video oral histories licensed under the license. Links to the Oral Histories table
Subject [Lexicons]	Lexicons using the license
Type [Copyright]	Creative Commons or Copyright
Additional Resources	Relevant external links for the license
Reference ID [License]	External link to the license

Institutions

Index of international and geopolitical institutions. Field names, order, and public archival view are not yet normalized.

Field	Description
Identifier	Name of the institution
Description	Generated prose description of the institution. Contains a working languages list and whether the institution is regionally or internationally focused.
Description [Type]	International or regional

Subject [Languages]	ISO 639-3 codes for the Institution's official or working languages. Links to Languages.
Subject [Macrolanguages]	ISO code for the macrolanguages used as official or working languages by the Institution. Links to the Macrolanguages table
Reference ID [Institution]	External link to the Institution's website, if available.
Reference ID [Wikipedia]	External link to the English-language Wikipedia article about the Institution, if available.

Writing Systems

Index of Writing Systems. Field names, order, and public archival view are not yet normalized.

Field	**Description**
Identifier	Predominant English-language name of the Writing System.
Description	Generated prose description of the writing system. Contains an approximate count of how many languages use the writing system and lists these languages ISO 639-3 code.
Languages [ISO 639-3]	ISO 639-3 codes of the languages that are predominantly written with this writing system. Links to Languages.
Subject [Language Count]	Number of languages that are predominantly written with the language.
Reference ID [Wikipedia]	Link to the English-language Wikipedia article about the writing system, if available.

Language Status

Index of language vitality statuses per the EGID scale. Field names, order, and public archival view are not yet normalized.

Field	Description
Identifier	Number value and official name of the vitality status, per EGIDS.
Description	Generated prose description of the EGIDS status. Contains an approximate language count and a list of these languages by ISO 639-3 code.
Subject [Languages: ISO 639-3]	ISO 639-3 Code for languages associated with the vitality status
Subject [Languages: Count]	Number of languages associated with the vitality status
Reference ID [EGIDS]	Official EGIDS entry about the language status, if available.

Publishers

Index of Publishers whose content we archive. Field names, order, and public archival view are not yet normalized.

Field	Description
Identifier	The publisher name
Description	Brief description of the publisher
Contributor [Oral Histories]	Oral history videos by the publisher
Reference ID [Publisher]	Publisher's website, if available

Inventory and Storage

Wikitongues stores all content on Dropbox, with two external hard drive backups in New York City and Pittsburgh. Depending on storage agreements with content donors, selected content is also backed up at the U.S. Library of Congress, the Internet Archive, and the Wikimedia Commons.

Dropbox

In the root directory of our Dropbox server, **Teamwide** > **1_Oral_Histories** contains every Wikitongues language video, organized in individual directories labeled by video identifier. In each video directory, a metadata file, the final edited video, and a video thumbnail are located in the directory root. Video components, such as raw media and caption files, are located in the **Raws** directory.

In the root directory of our Dropbox server, **Teamwide** > **2_Lexicons** contains copies of some of the lexicons listed in the Lexicons table of our database, organized in individual directories labeled by video identifier. Lexicon directory structures are not yet normalized.

In the root directory of our Dropbox server, **Teamwide** > **3_Metadata_Backup** contains time-stamped, .csv backups of our database.

External Harddrives

All content is backed up on two external hard drives in the United States in New York City, New York and Pittsburgh, Pennsylvania.

External Partners Storage

Unless otherwise specified by the content donor, all content is preserved by the U.S. Library of Congress and uploaded to the Internet Archive.

Unless otherwise specified by the content donor, all content under a CC-by-SA or more open license is uploaded to the Wikimedia Commons.

Maintenance

Intake

Technology Setup

To manage intake processes at Wikitongues, you'll need to familiarize yourself with Airtable, especially navigating *bases* and managing *records*.

You will also need to install the Dropbox desktop app and our Oral History Instantiator. Please contact scott@wikitongues.org and daniel@wikitongues.org for help with installing these tools.

Processing Oral History Donations

For content submitted through our open form at Wikitongues.org, access the form's backend on Airtable. Each content donation is stored as a unique record.

For content submitted manually (e.g. over email, WhatsApp, etc.), consult the donor for appropriate metadata about the content.

In Airtable, navigate to the Oral Histories table and create a new record for the content donation, populating it with all available metadata.

> A unique identifier will be automatically generated for the record after the **Date Added** and **Languages: ISO Code 639-3** fields are populated.
>
> If the content donation came with captions, after creating your record in Oral Histories, navigate to the Oral history captions table and create unique records for each set of captions. These records will automatically populate the corresponding record in the Oral Histories table.

Using the Oral History Instantiator tool, create a new directory for the record, named for the unique identifier. In the newly created directory, store video and audio files in **Raws** > **clips** and **Raws** > **audio**. Store caption files in **Raws** > **captions**. Rename raw files with their corresponding record names in Airtable.

When you don't have time in a single sitting to create records for large content donations, download all content before you've created any records and store it in the root directory of our Dropbox server in **Teamwide** > **4_Video_Drop**. This ensures the donation is safely stored on the cloud until there is time to organize it.

Occasionally, we receive content donations with metadata but no files, or a broken link to download the files. When this happens, create the record anyway, but note that the files are missing in the **Intake Notes** field.

Processing Lexicon Donations

We do not yet maintain an open form at Wikitongues.org for submitting lexicon documents, so all content donations are received manually. Consult the content donor to make sure you have appropriate metadata to create records.

In Airtable, navigate to the Lexicons table and enter the record's available metadata. For lexicons, a unique identifier will be automatically generated after the **Date Created**, **Source Language: ISO Code (639-3)** and **Target Language: ISO Code (639-3) fields** are populated.

In the root directory of our Dropbox server, store the lexicon in **Teamwide** > **2_Lexicons**. Name the document for its corresponding Airtable record.

How to Classify Content by Languages

Language classification is messy and content donations are usually submitted using a language's common colloquial name, so you'll need to manually pair a content donation with the appropriate ISO code when creating a record.

In the **Worksheet** view of the Languages table on Airtable, run a filter on the Language Names column: **Where Language Names contains [content donation name]**. The Identifier of the filtered record will be the correct ISO code.

If filtering the Languages table doesn't work, look up the content donation name on the English-language Wikipedia. This should direct you to the appropriate article. There, the infobox on the article's right-hand side will list the corresponding ISO 639-3 and Glottocodes, if they exist.

Since the ISO code set is narrower in scope than Glottocode, many language varieties have Glottocodes but don't have ISO codes. In this case, populate the **Languages: Dialect Glottocode** field with the correct Glottocode, and populate the **Languages: ISO Code 639-3** field with a custom identifier (see below).

From time to time, we receive content in a language that can't be paired with an existing ISO code. When this happens, create a new record in the Languages table, with a unique, four-letter identifier that begins with the letter *w*. The remaining three are up to you, as long as they don't conflict with another four-letter ID. Since this is a new record in the Languages table, populate it with as much metadata as you have. Make sure that at least one Subject Reference ID link is populated.

Backup schedule

Database

On the first business day of each month, download each table as a .csv. In the root directory of our Dropbox server, store it in **Teamwide** > **3_Metadata_Backup** in a unique directory named for the month: YYYY-MM.

Hard Drives

Our external hard drive backup schedule is not yet standardized.

External partners

Our external partners backup schedule is not yet standardized.

Short-term maintenance

Intake

On a daily basis, process content donations submitted through our open form at Wikitongues.org or by email and social media. See Intake above for process notes.

Pruning metadata

On a quarterly basis, remove records with missing files that are more than three-months old.

Changing and deprecating fields

From time to time, we need to make changes to the structure of our Airtable bases. Changes are documented in our archival journal and reflected in a new version of this document with updated schema.

When a field is marked for deprecation, prefix its name with an ✖ emoji. In the appropriate **Archival View** on Airtable, move the field to the end of the table until it can be safely deleted.

Metadata and storage reconciliation

On a monthly basis, check that our Airtable records match our directories on Dropbox to ensure that we're not missing any files. If we are missing files, you should be able to recover them from one of our external backups or through Dropbox's file recovery feature. To streamline the Airtable-Dropbox reconciliation process, you'll need to install our Oral History Directory Comparer tool from Github. Please contact scott@wikitongues.org and daniel@wikitongues.org for help with installing this tool.

Long-term maintenance

Updating metadata

On an annual basis, you'll need to update the following metadata:

Check ISO 639-3 and Glottocode code sets for any languages that may have been added, and incorporate them into Airtable. If new languages have been added, update record names and metadata, as well as file directories, accordingly.

Update the Wikipedia Intros field using Airtable's Wikipedia automation, based on the Subjects Reference ID: Wikipedia Intro field.

Check that our list of sovereign states accurately reflects current geopolitics. Are there new de facto independent states? Has the United

Nations admitted new members? Has any sovereign state changed its official, English-language name?

Stakeholders

Stakeholders in the digital preservation of languages and all-associated content include Wikitongues, users of Wikitongues, volunteers and researchers, libraries, students, educators, and others who contribute content. The roles of these stakeholders may vary, but in any action there is a contribution to the overall goals of the project.

The roles of the stakeholders may encompass digital management which may include: the migration of files, the creation and renaming of files, the prevention of corruption or bit-rot by assessing the collection and the implementation of solutions as necessary.

Stakeholders may need to constantly reassess the goals and the mission in order to ensure the digital preservation plan is being properly implemented.

Policy review

This policy will be appraised regularly to ensure that strategies continue to support Wikitongues mission and policies, that resources are being used effectively, and that adaptations are being made to address developing technologies, while using ISO programs, files and recommendations. Departmental review will occur annually to assist an organization-wide appraisal which will be conducted at least once every five years.

This document will be published in versions using the MAJOR.MINOR.PATCH (x.x.x) system, starting with Version 1.0.0.

> Small revisions, such as format revisions, a single edit to one section, or a series of minor edits to multiple sections, are PATCHes.
>
> Significant revisions to a section or sections, or the addition of new sections, are MINOR.
>
> Structural changes based on annual policy reviews are MAJOR.