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

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

Contributors

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
ID	Unique identifier assigned to each contributor. Formatted accordingly: **firstName_lastName_dateAdded**
Date Added	The date of the individual's first contribution to our archive
First Name	Given name or names
Last Name	Family name or names
Country	De jure or de facto nation-state where the contributor lives. This field links to the Nations table
Continent	Continent where the contributor lives. Links to the Continents table
Videos featured in	Language videos in which the contributor appears. Links to the Oral Histories table
Videos created	Language videos submitted by the contributor; or videos of which the contributor is the primary author. Links to the Oral Histories table
Videos recorded	Language videos that the contributor helped record, without featuring in them or being the primary author. Links to the Oral Histories table
Videos captioned	Language videos that the contributor helped caption or translate. Links to the Video Captions table
Video Descriptions	Video descriptions that the contributor helped write. Links to the Oral Histories table
Videos featured in count	The number of videos in which the contributor appears

Videos contributed count	The number of videos submitted by the contributor; or of which the contributor is the primary author
Captions count	The number of videos captioned by the contributor

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: Top level 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	Biological sex of the contributors featured in the video.
Contributor: Speakers	Contributors featured in the video. Links to the Contributors table

Contributor: Caption Authors	Contributors who helped caption or translate the video. Links to the Contributors table
Contributor: Videographer	Contributors who helped produce the video without being its primary author or featuring in it; i.e., holding the camera, editing, etc. Links to the Contributors table
Contributor: Description	Contributors who help write the description for this video. Links to the Contributors table
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 the Video Captions table
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
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.
Aspect ratio	Frame dimension category, either *portrait* or *landscape*.
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*
Lighting	The quality of lighting, either *1 - Unable to see face*, *2 - Difficult to see*, *3 - Okay*, *5 - Well-lit*, *8 - Professional lighting*
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*
Full signing space	The amount of visible physical range for videos of sign languages. Either *Yes*, *No*, or *Not relevant*
Quality of sound	The quality of audio. Either *1 - Distorted, unable to hear*, *2 - Scratchy, difficulties*, *3 - Quiet at times*, *5 - Overall good*, and *8 - Professional audio*

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*.
Duration	Duration length of the final, edited oral history file.
Length	Category of duration. Either *Less than 1 minute*, *1-3 minutes, 3-5 minutes*, *5-7 minutes*, or *7+ minutes*.
Captions language count	The number of languages in which the video is captioned.
Captions flagged	Automatic flag for high-quality videos that are uncaptioned or undercaptioned.
Documentation status	Aggregate documentation quality score of the video or audio file.
Date Received	The date this record was created.
Encoded Data	
Tagged Data	
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

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.
Oral History	The oral history that the caption set transliterates. Links to the Oral Histories table
Caption Language ISO 639-3	ISO 639-3 code for the captions language
Caption Language Glottocode	Glottocode for the captions language
Caption Language Name	Primary English-language name for the captions language
Author(s)	Author of the captions
.srt File	Link to the formatted caption file on Dropbox
Unformatted Text (raw)	Raw text transcript if the captions were submitted unformatted
Unformatted File	Text document (pdf, docx, etc) of the captions if they were formatted this way
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.
Source Language: Name	Primary English-language name of the lexicon's source language.
Source Language: ISO Code (639-3)	ISO 639-3 Code for the lexicon's source language. Links to the Languages table
Source Language: Glottocode	Glottocode for the lexicon's source languages. Links to the Glottocodes table.
Source Language: Dialect Glottocode	Glottocode for the unique dialectal variety of the lexicon's source language, if applicable. Links to the Glottocodes table.
Source Language: Macrolanguage ISO Code	ISO code for the macrolanguages to which the lexicon's source language pertains, if applicable. Links to the Macrolanguages table
Source Language: Top level genealogy	The highest-level language families to which the featured languages pertain. Links to the Language Families table.
Source Language: Continent of Origin	Continents of origin for the lexicon's source language. Links to the Continents table
Source Language: Nation of Origin	De jure or de facto nation-state for the lexicon's source language. This field links to the Nations table
Target Languages: Name	Primary English-language names of the lexicon's target languages.
Target Languages: ISO Code (639-3)	ISO 639-3 Code for the lexicon's target languages. Links to the Languages table
Target Languages: Glottocode	Glottocode for the lexicon's target languages. Links to the Languages table
Target Languages: Dialect Glottocode	Glottocode for the unique dialectal variety of the lexicon's target languages, if applicable. Links to the Glottocodes table.
Target Languages: Macrolanguage ISO Code	ISO code for the macrolanguages to which the lexicon's target languages

	pertain, if applicable. Links to the Macrolanguages table
Document Category	Library of Congress data category. Currently, all records are defined as *idioms*; this could change
Document Type	Library of Congress data category. Either *Dictionaries* or *Polyglot glossaries, phrase books, etc*
Subject: General	Library of Congress data category. Automatically formatted accordingly: **sourceLanguages------TargetLanguages**. Multiple values separated by commas
Description	Text description of the lexicon.
Date Created	The date the lexicon was created
Type	Media type; e.g., *text*
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: Cloud	Link to the lexicon file on Dropbox, if hosted by Wikitongues.
Coverage: External Link	Link to the lexicon file on an external platform, if also hosted elsewhere.
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.
Relation Label	Clarifies if the lexicon is a stand-alone document or part of a larger work; i.e., the chapter of a bigger dictionary.
Relation Source	The larger work of which the lexicon is a part, if applicable
Relation	Statement of the lexicon's relation to a larger work, if applicable
Format: Medium	Technical metadata to be defined

Format: Extent	Technical metadata to be defined

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'.
Standardized Names	Standardized, English-language name for the language, as featured in Ethnologue and substantiated by academic literature.
Language Names	Array of other names for the language, including names specific to dialectal and regional variants
Glottocode	Official Glottocode for language, if available. Links to the Glottocodes table
Continent of Origin	Continental region of the language's origin. Links to the Continents table
Nations of Origin	De jure or de facto nation-state where the language originated. This field links to the Nations table
Territories	Sub-national territories (i.e., U.S. states, Canadian provinces, etc) where the language originated. This field links to the Territories table
Macrolanguage	ISO code for the macrolanguage to which the language pertain's, if applicable. Links to the Macrolanguages table
Dialects	List of dialects, as listed by Ethnologue
Genealogy	Top-level genealogy of the language. Links to the Language Families (Top-level genealogy) table.
Genealogy Keywords	Full-tree genealogy as listed by Ethnologue. Scheduled for

	deprecation.
Writing System	Writing system predominantly used by speakers of the language. Links to the Writing Systems table.
Language Description	Prose description of the language written by someone at Wikitongues.
Typology	Notes on the language's typology.
Demographics	Notes on the language's demographics, especially as it pertains to population, as listed by a deprecated (pre-paywalled) version of Ethnologue.
Geographic Notes	Notes on the language's geographic distribution, especially as it pertains to population, as listed by a deprecated (pre-paywalled) version of Ethnologue.
Social Context Notes	Notes on the vitality and political recognition of the language,as well as the social (especially religious) context of the community, as featured on a deprecated (pre-paywalled) version of Ethnologue.
Language Use	Notes on the vitality of the language in its community.
Language Development	Notes on documentation, materials availability, and literacy. Under review for deprecation.
Language Notes	Additional miscellaneous notes on the language.
Language Status	Language vitality per the Expanded Graded Intergenerational Disruption Scale (EGIDS). Links to the Language Status table.
Language Status Raw	Language vitality per the Expanded Graded Intergenerational Disruption Scale (EGIDS). Unlinked. Under review for deprecation.
Institutions	International institutions and organizations that use the language as a working or official language. Links to the Institutions table

Nationally Official In	De jure or de facto nation-states where the language is official. This field links to the Nations table
Regionally Recognized In	Sub-national territories (i.e., U.S. states, Canadian provinces, etc) where the language is officially recognized. This field links to the Territories table
Speakers recorded	Unique IDs of oral history videos featuring this language. Links to the Oral Histories table
Lexicon source	Lexicons for which this language is the source language. This field links to the Territories table
Lexicon target	Lexicons for which this language is the target language. This field links to the Territories table
Subjects Reference ID: Language resources URL	Link to the OLAC page about this language, if available.
Subjects Reference ID: Ethnologue URL	Link to the Ethnologue page about this language, if available.
Subjects Reference ID: ISO URL	Link to the SIL-ISO page about this language, if available.
Subjects Reference ID: Wikipedia Intro	Text introduction of the English-language Wikipedia article about the language.
Identifier Source	Either ISO 639-3 or Wikitongues

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
ISO 639-3	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’.
Macrolanguage name	Standardized, English-language name for the macrolanguage.

Individual languages	Unique IDs of individual languages that pertain to this macrolanguage. Links to the Languages table
Nations of Origin	De jure or de facto nation-state where the macrolanguage originated. This field links to the Nations table
Territories	Sub-national territories (i.e., U.S. states, Canadian provinces, etc) where the macrolanguage originated. This field links to the Territories table
Speakers Lookup	Unique IDs of videos recorded in languages that pertain to the macrolanguage. Links to the Contributors table
Genealogy	Top-level language family to which the macrolanguage pertains. Links to the Language Families (Top-level genealogy) table
Demographics	Notes on the demographics of the macrolanguage, especially as it pertains to population numbers.
Language Use	Notes on the vitality of the macrolanguage in its community. Under review for deprecation.
Language Development	Notes on documentation, materials availability, and literacy. Under review for deprecation.
Language Description	Under review for deprecation.
Continent of Origin	Continental region of the language's origin. Links to the Continents table
Language resources URL	Link to the OLAC page about this macrolanguage, if available.
Ethnologue URL	Link to the Ethnologue page about this macrolanguage, if available.
ISO URL	Link to the SIL-ISO page about this macrolanguage, if available.
Institutions	International institutions and organizations that use the macrolanguage as a working or official language. Links to the Institutions table
Nationally Official In	De jure or de facto nation-states where the macrolanguage is official. This

	field links to the Nations table
Regionally Recognized In	Sub-national territories (i.e., U.S. states, Canadian provinces, etc) where the macrolanguage is officially recognized. This field links to the Territories table
Geographic Notes	Notes on the macrolanguage's geographic distribution, especially as it pertains to population, as listed by a deprecated (pre-paywalled) version of Ethnologue.

Language Families (Top-level genealogy)

Index of every top-level language family.

Field	Description
Name	The standardized, English-language name of the language family.
Languages	Unique IDs for the languages that pertain to this language family. Links to the Languages table
Language count	The number of languages that pertain to this language family
Macrolanguages	Unique IDs for the macrolanguages that pertain to this language family. Links to the Macrolanguages table
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
glottocode	Unique Glottocode as assigned by the the Glottolog project
name	The standardized, English-language name for the language

isocodes	The corresponding ISO code for the language. Links to the Languages table
Langoid Names	The standardized, English-language name for the language, macrolanguage, or dialectal variant.
level	The classification level—dialect, language, etc—of the Langoid.
macroarea	Cultural-geographic area of the Langoid
latitude	Latitude of the Langoid macroarea
longitude	Longitude of the Langoid macroarea
Individual Languages	Unlinked corresponding ISO code. Scheduled for review and possible deprecation
Oral Histories: Dialect	Oral history videos in which the Langoid dialect variant is listed
Lexicon Source Language: Dialect	Lexicon document in which the Langoid dialect variant is the source language
Lexicon Target Language: Dialect	Lexicon document in which the Langoid dialect variant is the target language

Continents

Index of every top-level continental region.

Field	Description
Name	Name of the continental region
LOC Naming Authority Name	Standardized, English-language name of the continental region as listed by the Library of Congress
Reference Link ID	External source material about the continental region
Nations	De facto or de jure nation-states in the continental region. Links to the Nations table
Territories	Sub-national territories (U.S. states, Canadian provinces, etc.) in the continental region. Links to the

	Territories table
Nation Count	Number of de jure or de facto nation-states in the continental region
Territory Count	Number of sub-national territories in the cont
Language Count	Number of languages that originated in the continental region

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
Polities	Standardized, English-language name of the jure or de facto nation-state
LOC Naming Authority	Standardized, English-language name of the nation-state as listed by the Library of Congress
Official Name	Official, English-language name of the nation-state as listed by the nation-state's government
Reference Link ID	External source material about the continental region
Contributors	Contributors who are based in the nation-state. Links to the Contributors table
Languages	ISO 639-3 codes for the languages that originated in the nation-state. Links to the Languages table
Language Count	Number of languages that originated in the nation-state
Language Count Details	Description of language count, such as which languages could be considered at-risk, as displayed on an early (pre-paywalled) version of Ethnologue
Immigrant Languages	Notes on prominents diasporic language communities in the nation-state
Nationally Official Languages	ISO 639-3 codes of the nation-state's official languages. Links to the

	Languages table
Nationally Official Names	Standardized, English-language names of the nation-state's official languages
Regionally Official Languages	ISO 639-3 codes of the nation-state's regionally official languages. Links to the Languages table
Regionally Official Names	Standardized, English-language names of the nation-state's regionally official languages
Territories	The nation-state's sub-national territories. Territories from federal countries and geographically expansive countries are included. Links to the Territories table
Continent	Continental region of the nation-state. Links to the Continents table
Oral Histories Recorded in Nation	Video oral histories recorded in the nation-state. Links to the Oral Histories table
Lexicon Documents Created in Nation	Lexicon documents recorded in the nation-state. Links to the Lexicons table
ISO 3166	ISO 3166 code for nation-state
Indigenous Macrolanguages	Macrolanguages that originated in the nation-state. Links to the Languages table
Nationally Official Macrolanguages	Macrolanguages that are nationally official in the nation-state. Links to the Languages table
Regionally Official Macrolanguages	Macrolanguages that are regionally official in the nation-state. Links to the Languages table

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
Polities	Unique ID assigned to the territory. Formatted:

	Territory-Name_CountryCode
Name	The territory name
LOC naming authority name	Standardized, English-language name as assigned by the U.S. Library of Congress
Reference Link ID	U.S. Library of Congress reference page for the territory
Type	Type of sub-national territory: Special Administrative Region, Federated Entity, Autonomous Country, Autonomous Region, Capital Territory, Province, or County
Sovereignty	Sovereign (nation-)state to which the territory pertains. Links to the Nations table
Continent	Continental region to which the territory pertains. Links to the Continents table
ID	Two-letter country code of the nation-state to which the territory pertains
Regionally official	List of languages with official recognition in the territory
Languages (names)	Deprecated field with local language data, to be consolidated with the Languages field below.
Languages	ISO 639-3 codes for the languages predominantly spoken in the territory. Links to the Languages table
Language Count	Number count of the languages predominantly spoken in the territory
Immigrant Languages	Notes about established diaspora communities in the territory
Contributors	Contributors who are based in the territory. Links to the Contributors table
Oral Histories	Video oral histories that were recorded in the territory. Links to the Oral Histories table
Language Names	Standardized, English-language names of the languages spoken in the territory

Macrolanguages	Macrolanguages with official status or recognition in the territory. Links to the Macrolanguages table
Lexicons	Lexicon documents created in the territory. Links to the Lexicons table

Rights

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

Field	Description
Name	The shorthand name of the license.
Full Name	The full legal name of the license
Short Text	Summary of the license
Full Text	Full text of the license, its applications and attribution information
Video License Statement	Official text for Wikitongues platforms to summarize the language in question
URI	External link to the license
Type	Creative Commons or Copyright
Oral Histories	Video oral histories licensed under the license. Links to the Oral Histories table
Oral History Distribution	Number of Oral History videos licensed under the license
Lexicons	Lexicons using the license
Lexicon Distribution	Number of lexicon documents using the license. Links to the Lexicon table
Additional Links	Relevant external links for the license

Institutions

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

Field	Description
Name	Name of the institution
ISO	ISO 639-3 codes for the Institution's official or working languages. Links to the Languages table
Type	International or regional
URL	External link to the Institution's website
Wikipedia	External link to the English-language Wikipedia article about the Institution
Macrolanguages	ISO code for the macrolanguages used as official or working languages by the Institution. Links to the Macrolanguages table

Writing Systems

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

Field	Description
Name	Predominant English-language name of the Writing System
Languages	ISO 639-3 codes of the languages that are predominantly written with this writing system. Links to the Languages table
Count	Number of languages that are predominantly written with the language

Language Status

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

Field	Description
ID	Number value assigned to the vitality status

Classification	Official ID of the vitality status
Name	Official name of the vitality status
Languages	ISO 639-3 Code for languages associated with the vitality status
Count	Number of languages associated with the vitality status

Publishers

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

Field	Description
Name	The publisher name
Oral Histories	Oral history videos by the publisher

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.