SIL Ethnologue® Languages of the World

Ethnologue Global Dataset

Twenty-seventh edition data

David M. Eberhard, Gary F. Simons, and Charles D. Fennig, Editors

Based on information from *Ethnologue*, 27th edition:

Eberhard, David M., Gary F. Simons, and Charles D. Fennig (eds.). 2024. *Ethnologue: Languages of the World.* Twenty-seventh edition. Dallas, Texas: SIL International. Online: http://www.ethnologue.com.

SIL International, 7500 West Camp Wisdom Road, Dallas, Texas 75236-5699 USA
Web: www.sil.org, Phone: +1 972 708 7404, Email: publications_intl@sil.org

Contents

0. Introduction

This document describes the contents and structure of the *Ethnologue Global Dataset* for the 27th edition of *Ethnologue*. If you have used a previous edition of the dataset, see "6. Change History" (page 20) for a description of changes to the data tables since the previous edition. Whereas the www.ethnologue.com website provides information about the world's languages in a presentation view, this product makes much of that same information available in an actionable format. Before describing the details of the data, this introduction first describes the rationale that lies behind the development of the product and the terms under which it may be used.

Underlying rationale

The purpose of the *Ethnologue Global Dataset* is to make it possible for researchers to replicate the statistical summaries that are published in *Ethnologue* (see http://www.ethnologue.com/statistics) and to use data from *Ethnologue* in their own analyses. Most of the information published in *Ethnologue* is in the form of textual comments that are not amenable to statistical analysis; these fields of information are specifically not included in the dataset. This dataset contains only data fields with simple values (like Booleans, numbers, categories) that can be submitted to statistical analysis.

Another criterion we have followed in selecting data fields for inclusion in the dataset is that they be adequately complete and reliable in terms of global coverage. There are many fields of information that *Ethnologue* reports when the information is available to us, but which are missing for a significant number of languages or which we know to be inadequately comparable across languages. Such fields have not been included in this dataset since any conclusions drawn from global analysis using such fields would be suspect.

A final principle we have followed in designing the product is convenience for the user. That is why the tables are not in fully normal relational form; rather, we felt it would be more convenient for users to deal with a lower number of denormalized tables. Some manifestations of this denormalization are the redundancy of including both codes and their associated names or labels and the inclusion of some columns that are computed from others.

With each new edition of *Ethnologue*, we seek to transform more fields of information from being arbitrary text to becoming actionable data fields that can be added to this product. If the analysis you are seeking to do needs information that is in *Ethnologue* in some form, but is not found in this product, you are invited to send a description of your use case to ethnologue_editor@sil.org. This will give input to the editors as they make plans for the next edition. If you publish or post visualizations of these data or findings from analyzing them, we would appreciate hearing about your work. Please use the same editorial email address to tell us about how you have used this dataset.

Terms of use

The *Ethnologue Global Dataset* is a licensed product with restricted terms of use. If you have licensed the dataset under the Personal Research License, or are an Authorized User of an organization that has licensed it under the Institutional Research License, you agreed to the following terms before receiving the data files:

- You may not share your copy with others, except to do so temporarily with someone who is assisting you in the analysis you are performing.
- You may freely publish and distribute visualizations you create from your analysis of these data and tables presenting results that aggregate over the data.
- You should cite this product as the source in any work you produce that is based on analysis of this dataset or that includes visualizations of any of this information.
- You may *not* redistribute the raw data in any form, including displaying it on a public website, posting the files on an intranet site, or incorporating any of these data into a dataset that you are distributing. To inquire about uses such as these, contact the publisher at customer_service@ethnologue.com.

If you represent an organization that has licensed the dataset under the Institutional Research License, you agreed to the following terms:

- You may only distribute copies of the dataset for the use of faculty, students, and staff currently affiliated with your institution or organization.
- Your authorized users will be required to read and agree to the above terms for personal research use before getting access to the data.
- The dataset will not be placed on a public server, but made available by a means such that only your authorized users can get access to it.

The above are only the highlights of the licensing agreements. To get more information about licensing, contact us at https://www.ethnologue.com/contact-us.

1. Overview of Product

The data are distributed as three tables in the ubiquitous tab-delimited file format that can be loaded into virtually any spreadsheet, database, or other data analysis tool. The string data are encoded in UTF-8, so be sure to specify that encoding scheme when importing the tables into application software. (For a test that the character encoding has been correctly interpreted, check the name of Côte d'Ivoire in the CI row of the Table_of_Countries.) Note, too, that the first line of each data file contains the names of the data columns.

This product includes the following tab-delimited data files:

- *Table_of_Countries.tab* has a row for each country that *Ethnologue* reports on. It is documented below in "2. Table of Countries" (page 6).
- *Table_of_Languages.tab* has a row for each language that *Ethnologue* reports on. It is documented below in "3. Table of Languages" (page 9). The data in this table pertain to the language in general or provide aggregated results over all the countries in which it is used.
- *Table_of_LICs.tab* has a row for each language-in-country (LIC) that *Ethnologue* reports on. It is documented below in "4. Table of LICs" (page 14). The data in this table pertain to the language in a particular country where it is used.

In Table_of_Languages, there is exactly one row for each language. This table provides the source data that were used for producing tables 1 through 6 of the Statistics section on the *Ethnologue* website; see http://www.ethnologue.com/statistics (accessible when logged in with a subscription). In Table_of_LICs there may be multiple rows for a language, one for each country in which it is used. This table provides the source data from which the information in tables 7 and 8 on the *Ethnologue* site were calculated. Table_of_Countries captures those same calculations.

In the next three sections, the columns contained within these three tables are documented. The columns are listed in the order in which they appear in the rows of the tab-delimited files. The data types of the columns are expressed in terms of the following ANSI SQL data types:

char(*n*)	The value is fixed-width string that is *n* characters in length.
varchar(*n*)	The value is a variable-width string that is up to a maximum of *n* characters in length.
integer	The value is an integer.
decimal(*p,d*)	The value is a decimal number with *p* digits of precision and *d* digits after the decimal point.

2. Table of Countries

The table of countries has 242 rows, one for each country that *Ethnologue* reports on. Each row contains the following columns:

Column	Type	Description
Country_Code	char(2)	The unique two-letter identifier for the country from ISO 3166; see http://www.iso.org/iso/country_codes.
Country_Name	varchar(40)	The name used for the country in *Ethnologue*.
Languages	integer	The number of living languages for this country. It is the sum of *Established* plus *Unestablished*.
Indigenous	integer	The number of LIC records in this country for languages that are living and are indigenous to the country (that is, where *EGIDS* < "x10" and *Is_Indigenous* = "T").
Established	integer	The number of LIC records in this country for languages that are living and are well-established within the country (that is, where *EGIDS* < "x10" and *Is_Established* = "T"). See the documentation of *Is_Established* below for the full definition of what constitutes an "established" language.
Unestablished	integer	The number of LIC records in this country for languages that are not established within the country (that is, where *Is_Established* = "F"). For instance, it is a major language that many have chosen to learn or is the language of temporary workers or recent refugees.
Diversity	decimal(4,3)	The Greenberg diversity index. This is the probability that any two people of the country selected at random would have different mother tongues. The highest possible value, 1, indicates total diversity (that is, no two people have the same mother tongue) while the lowest possible value, 0, indicates no diversity at all (that is, everyone has the same mother tongue). See: Greenberg, J. H. (1956) The measurement of linguistic diversity, *Language* 32(1):109–115; Lieberson, S. (1981) *Language diversity and language contact,* Palo Alto: Stanford University Press.

Column	Type	Description
Included	integer	The number of languages in the country for which population estimates are available and which are therefore included in the calculation of the diversity index. Comparing this number to the value in the *Languages* column gives an indication of the relative coverage of the diversity calculation.
Sum_Of_Populations	integer	The sum of *L1_Users* for all the languages in the country, based on the most recent population estimate that is available to *Ethnologue*. If the L1 estimates for all languages were up-to-date and if every person had only one L1, this sum would match *Population*, which is the latest estimate of the national population. However, this number may differ significantly from *Population*. When *Sum_Of_Populations* is significantly less than *Population*, it is an indication that the population within the country has grown significantly since the most recent population estimates for the individual languages. When *Sum_Of_Populations* is significantly greater than *Population*, it is an indication that a significant portion of the country's population has grown up with bilingualism and thus has L1 proficiency in more than one language.
Mean	integer	The mean first-language speaker population of a language in the country. It is *Sum_of_Populations* divided by *Included.*
Median	integer	The median first-language speaker population of a language in the country. That is, half of the languages are larger than this size, and half are smaller. (When there is an odd number of *Included* languages, it is the middle value when the populations are sorted into numerical order. When there is an even number, it is calculated as the average of the two middle values.)
Population	integer	The population of the country. This figure is taken from the most recent national census data where available or the current estimated population from the United Nations or another reliable source. See the *Ethnologue* page of the relevant country for the specific source information.

Column	Type	Description
Literacy_Rate	decimal(3,2)	This is an estimate of the proportion of the population in the country that is literate in any language of the country. Data are primarily from UNESCO but may also come from various other sources if more recent estimates are available. See the *Ethnologue* page of the relevant country for the specific source information.
Conventions	integer	*Ethnologue* has identified 9 conventions within the body of international law that affirm the language and culture rights of indigenous and minority peoples. This column gives a count of the number of these conventions that each country has subscribed to. See the *Ethnologue* page of the relevant country for the list of international conventions to which the country is a party.

3. Table of Languages

The table of languages has 7,613 rows, one for each language that *Ethnologue* reports on. Each row contains the following columns:

Column	Type	Description
ISO_639	char(3)	The unique three-letter code for identifying the language. These codes are taken from the ISO 639-3 standard; see https://iso639-3.sil.org/.
Language_Name	varchar(50)	The primary name used in *Ethnologue* for the language associated with this ISO 639-3 code. The name may include a comma, in which case a generic ethnic name precedes the comma and a specific name for the particular language follows it.
Uninverted_Name	varchar(50)	The uninverted form of the *Language_Name*. That is, if the latter contains a comma, the two parts of the name are reversed and the comma is removed.
Country_Code	char(2)	The unique two-letter identifier from ISO 3166 for the primary country of the language. This is typically the country of origin.
Country_Name	varchar(40)	The name used in *Ethnologue* for the primary country.
Region_Code	char(3)	A three-letter mnemonic code representing the geographical region in which the primary country is located. *Ethnologue* follows the scheme of regions used by the UN in its statistical reporting; see https://en.wikipedia.org/wiki/United_Nations_geoscheme. The UN standard uses three-digit codes; these mnemonic three-letter codes were devised for use in the *Ethnologue* database.
Region_Name	varchar(30)	The name of the UN geographical region.
Area	varchar(8)	The name of the major world area in which the primary country is located. The field has one of five values: Africa, Americas, Asia, Europe, Pacific.

Column	Type	Description
L1_Users	integer	The total population of native (or first-language) users of the language in all countries. The value is the sum of *L1_Users* in all the LICs for this language. When there are no known first-language speakers remaining, the value is 0. When the population is unknown, the value is null.
Digits	integer	The order of magnitude of *L1_Users* expressed as the number of digits in the population number, computed as *floor(log10(Population))+1.* When there are no known first-language speakers remaining, the value is 0. When the population is unknown, the value is null.
All_Users	integer	The total number of all users of the language in all countries, including both first-language users and those who have learned it as a second language. The value is the sum of the non-null *All_Users* in all the LICs for this language. When no populations are known, the value is null.
Countries	integer	The number of countries in which use of the language is established. That is, this is the count of LICs for the language in whihch *Is_Established* = "T".
Family	varchar(30)	The name of the highest level grouping in the genealogical classification of the language. It is the same as the beginning of the value for *Classification* up to the first comma.
Classification	varchar(250)	The complete genealogical classification path from highest level grouping to lowest level. It is a comma-delimited list of subgroup names. The longest path contains 14 levels.
Latitude	decimal(7,4)	The latitude of a centroid point for the language expressed in decimal degrees. Positive values represent North latitude; negative values represent South latitude.
Longitude	decimal(7,4)	The longitude of a centroid point for the language expressed in decimal degrees. Positive values represent East longitude; negative values represent West longitude. The goal in constructing this dataset has been to ensure that every known language can be represented by a dot when the results of queries are mapped. In cases where the maps in *Ethnologue* show a polygon for the language, the centroid is the middle point within that polygon. In the

Column	Type	Description
		remaining cases, a point within the primary country has been selected (sometimes arbitrarily). For instance, the points for national languages are often placed near the national capital. Consult the region information reported in the *Ethnologue* description of the language to determine whether the point is arbitrary or represents a precisely known location. To the best of our knowledge, all centroid points are located on land with respect to the Digital Chart of the World (VMap Level 0, Edition 5). When superimposing these points on other base maps, some centroids (especially those for languages on tiny islands) may end up in the water. Note that there are approximately 30 languages (all extinct) for which we can find no information on where it was located and these remain with no centroid point.
EGIDS	varchar(3)	The estimated level on the Expanded Graded Intergenerational Disruption Scale. It is the strongest of the EGIDS levels for all the countries in which the language is used, except for the six official languages of the United Nationas in which case the value is EGIDS 0 (International). See https://www.ethnologue.com/methodology/#languageStatus for the definitions of the levels and literature references. Note that the values of the scale are strings rather than numbers because of the “a” and “b” distinctions. Also level 10 (Extinct) is coded as “x10”. This ensures that when the levels are sorted, the extinct languages appear at the bottom of the list (rather than falling between “1” and “2”).

Column	Type	Description
YLSD	Integer	Year last speaker died. For all EGIDS 9 and 10 languages we endeavor to report the year when the last fluent speaker died. In the publication, this is often reported as a decade or a part of a century. In this dataset, such values are transformed to a year in the middle of that range. For instance, if the last speaker died in the 1980s, this column reports "1985". If the last speaker died in the early 20th century, it reports "1925". In 2% of cases we have not yet discovered the YLSD; in these cases the value is left blank. This column provides an up-to-date dataset for replicating our 2019 study of "Two centuries of spreading language loss"; see https://dx.doi.org/10.3765/plsa.v4i1.4532.
Is_Written	char(1)	The value is "T" (true) if the language has a writing system such that literature in a standardized form is known to have been used by some members of the language community (even if it is no longer in use); otherwise, it is "F" (false). A value of "T" indicates either that the language achieved at lesat incipient literacy (EGIDS 5) at some point in its history, or that a heritage community of L2 users is writing the language as part of a revitalization effort. The fact that outsiders are known to have written the language is not enough to make the language qualify as written for the current purpose. For instance, specifically excluded are publications for an academic audience or materials printed in an orthography that was never accepted by the language community.
DLs_Level	varchar(10)	The level of digital language support (DLS) as measured by the method developed by *Ethnologue*, which is described in the paper "Assessing digital language support on a global scale"; see https://aclanthology.org/2022.coling-1.379/. The scale for measuring the digital vitality of a language has five levels: Still, Emerging, Ascending, Vital, and Thriving. See http://www.ethnologue.com/methodology/#DLS for the definitions of the levels.

Column	Type	Description
DLS_Score	decimal(5,4)	The DLS score is a more refined measure that is produced by the same methodology. It is a number from 0 to 1 that reports how far the language has climbed on the growth curve of digital language support. The measurements are based on web harvessting done in the fourth quarter of the previous year.
Institutional	integer	The value is 1 if the language has an EGIDS level from 0 to 4; otherwise, this value is 0. Computing the sum of this column over selected rows counts the number of languages that fall within the "Institutional" summary category; see http://www.ethnologue.com/language-development/#egids.
Developing	integer	The value is 1 if the language has an EGIDS level of 5; otherwise, this value is 0. Computing the sum of this column over selected rows counts the number of languages that fall within the "Developing" summary category; see http://www.ethnologue.com/language-development/#egids.
Vigorous	integer	The value is 1 if the language has an EGIDS level of 6a; otherwise, this value is 0. Computing the sum of this column over selected rows counts the number of languages that fall within the "Vigorous" summary category; see http://www.ethnologue.com/language-development/#egids.
In_Trouble	integer	The value is 1 if the language has an EGIDS level of 6b or 7; otherwise, this value is 0. Computing the sum of this column over selected rows counts the number of languages that fall within the "In trouble" summary category; see http://www.ethnologue.com/endangered-languages/#egids.
Dying	integer	The value is 1 if the language has an EGIDS level from 8a to 9; otherwise, this value is 0. Computing the sum of this column over selected rows counts the number of languages that fall within the "Dying" summary category; see http://www.ethnologue.com/endangered-languages/#egids.
Extinct	integer	The value is 1 if the language has an EGIDS level of 10; otherwise, this value is 0. Computing the sum of this column over selected rows counts the number of languages that fall within the "Extinct" summary category; see http://www.ethnologue.com/endangered-languages/#egids.

4. Table of LICs

The table of LICs has 12,585 rows, one for each language-in-country (LIC) that *Ethnologue* reports on. Only individual languages are included in the dataset. The pages for countries on the *Ethnologue* website also list "macrolanguages", both in the main language listings and in the immigrant language lists; however, these are omitted from the dataset since they are ambiguous with respect to individual language and including them in statistical analyses would introduce double counting of individual languages. Each row of the table contains the following columns:

Column	Type	Description
ISO_639	char(3)	The unique three-letter code for identifying the language. These codes are taken from the ISO 639-3 standard; see https://iso639-3.sil.org/. The identity of the same language across multiple countries is captured in this three-letter code, since the same language may have different names in different countries.
Language_Name	varchar(50)	The primary name used in *Ethnologue* for this language in this country. The name may include a comma, in which case a generic ethnic name precedes the comma and a specific name for the particular language follows it.
Uninverted_Name	varchar(50)	The uninverted form of the *Language_Name*. That is, if the latter contains a comma, the two parts of the name are reversed and the comma is removed.
Country_Code	char(2)	The unique two-letter identifier from ISO 3166 for a country in which the language for this row is used.
Country_Name	varchar(40)	The name used in *Ethnologue* for the country.
Region_Code	char(3)	A three-letter mnemonic code representing the geographical region in which the country is located. *Ethnologue* follows the scheme of regions used by the UN in its statistical reporting; see https://unstats.un.org/unsd/methodology/m49/#geo-regions. The three-letter codes are not from a standard, but were devised for use in *Ethnologue*.
Region_Name	varchar(30)	The name of the UN geographical region.
Area	varchar(8)	The name of the major world area in which the country is located. The field has one of five values: Africa, Americas, Asia, Europe, Pacific.

Column	Type	Description
Is_Primary	char(1)	The value is "T" (true) if this is the primary country for the language in this row; otherwise, it is "F" (false). The primary country is typically the country of origin of the language. If the area of origin spans multiple countries, it is the country that has the most speakers. For each unique value of *ISO_639*, there is one and only one country identified as primary.
Is_Indigenous	char(1)	The value is "T" (true) if the language is indigenous to this country; that is, it is known to have originated within the country. Where country borders run through the original homeland, the language will be recorded as being indigenous to multiple countries. In the case of a pidgin or creole language, it is considered indigenous to the country (or countries) where the language contact that gave birth to the language took place. On the other hand, if the value is "F" (false), the language is known to have immigrated from elsewhere in relatively recent history (at least since the time of European colonial expansion).
Is_Established	char(1)	The value is "T" (true) if the language is well-established in this country. It is considered establishsed if: (1) it is indigenous to the country, (2) it is a language of immigrants who established a multi-gnerational language community that has successfully transmitted the language to rising generations in this country, or (3) it is a language from elsewhere that is being successfully transmitted as an L2 through formal education even though it may not have an L1 community. Otherwise, if the value is "F" (false), the language is used as a first language only by recent immigrants.
All_Users	integer	The estimated population of all users of the language in this country. It is the sum of *L1_Users* and *L2_Users*. When both *L1_Users* and *L2_Users* are unknown, the value is null. If *L1_Users* is greater than zero and *L2_Users* is unknown, the value of *L1_Users* is given. Note that in the latter case, *All_users* may be under-reported since it would be higher if *L2_Users* were known.

Column	Type	Description
L1_Users	integer	The estimated population of first-language users of the language in this country. When there are no known first-language users remaining, the value is 0. When the population is unknown, the value is null.
L2_Users	integer	The estimated population of second-language users of the language in this country. This includes all users for whom it is not a native language, regardless of whether it was learned second, third, or so on. When no such population is known, the value is null.
Family	varchar(30)	The name of the highest level grouping in the genealogical classification of the language. The complete genealogical classification path from highest level grouping to lowest level may be found in the *Classification* column of the Table_of_Languages by doing a relational JOIN on the *ISO_639* column.
EGIDS	varchar(3)	The estimated level on the Expanded Graded Intergenerational Disruption Scale for this language within this country; it is either equal to or lower than the value for the language as a whole. The EGIDS column is null for immigrant languages (that is, *Is_Established* = "F"). See http://www.ethnologue.com/methodology/#languageStatus for the definitions of the levels and literature references. The values of the scale are actually strings rather than numbers (because of the "a" and "b" distinctions). Therefore level 10 (Extinct) is coded as "x10". This ensures that when the levels are sorted, the extinct languages appear at the bottom of the list (rather than falling between "1" and "2").

Column	Type	Description
Function_Code	char(3)	A three-letter code representing the function for which this language is officially recognized in this country; if there is no official recognition for the language in this country, the value is null. The recognized functions are defined in Table 3 at http://www.ethnologue.com/methodology/#FICLabels. The value is RCL for "Recognized language" and LRN for "Language of recognized nationality." The other 12 possible values represent the first 12 categories in Table 3. These code values may be decomposed as follows in order to do analysis of the various components of recognition: • First character: S (statutory) versus D (de facto) • Second character: N (national) versus P (provincial) • Third character: L (both working language and identity language) versus W (working language only) versus I (language of identity only)
Function_Label	varchar(42)	A human-readable label of the function represented by the *Function_Code*.
Institutional	integer	The value is 1 if the LIC has an EGIDS level from 0 to 4; otherwise, this value is 0. Computing the sum of this column over selected rows counts the number of languages that fall within the "Institutional" summary category; see http://www.ethnologue.com/language-development/#egids.
Developing	integer	The value is 1 if the LIC has an EGIDS level of 5; otherwise, this value is 0. Computing the sum of this column over selected rows counts the number of languages that fall within the "Developing" summary category; see http://www.ethnologue.com/language-development/#egids.
Vigorous	integer	The value is 1 if the LIC has an EGIDS level of 6a; otherwise, this value is 0. Computing the sum of this column over selected rows counts the number of languages that fall within the "Vigorous" summary category; see http://www.ethnologue.com/language-development/#egids.

Column	Type	Description
In_Trouble	integer	The value is 1 if the LIC has an EGIDS level of 6b or 7; otherwise, this value is 0. Computing the sum of this column over selected rows counts the number of languages that fall within the “In trouble” summary category; see http://www.ethnologue.com/endangered-languages/#egids.
Dying	integer	The value is 1 if the LIC has an EGIDS level from 8a to 9; otherwise, this value is 0. Computing the sum of this column over selected rows counts the number of languages that fall within the “Dying” summary category; see http://www.ethnologue.com/endangered-languages/#egids.
Extinct	integer	The value is 1 if the LIC has an EGIDS level of 10; otherwise, this value is 0. Computing the sum of this column over selected rows counts the number of languages that fall within the “Extinct” summary category; see http://www.ethnologue.com/endangered-languages/#egids.

5. Other Data Sources

Other freely available data sources may be downloaded and used in conjunction with this dataset. The following use ISO 639-3 codes and may be used to get additional information about languages:

- The *Ethnologue Global Dataset* includes only the primary names for languages. The full set of alternate language names published in *Ethnologue*, including the primary and alternate names of all dialects associated with a language, may be downloaded from http://www.ethnologue.com/codes. *CountryCodes.tab* and *LanguageCodes.tab* are a selection of columns from the tables of countries and languages in this product. However, *LanguageIndex.tab* includes all of the alternate names and is not duplicated in this product. Use the *Country_Code* and *ISO_639* columns in this product to JOIN to that table.
- The languages listed in the *Ethnologue Global Dataset* do not include macrolanguages, since this introduces double counting of the same speaker communities as part of both an individual language and a macrolanguage. In order to include macrolanguages in your analysis, you may download the ISO 639-3 code table including the macrolanguages plus the table of macrolanguage mappings from https://iso639-3.sil.org/code_tables/download_tables.

The following licensed product is keyed to ISO 639-3 codes and can be used in conjunction with this dataset:

- The *Ethnologue Language GIS Dataset* contains the polygons that were used in creating the maps for the current edition of *Ethnologue*. The polygons represent the current home areas of each indigenous language; see https://www.ethnologue.com/data-consulting.

The following sources provide open data on a wide variety of indicators that could be analyzed in relation to language:

- The official United Nations site for Sustainable Development Goals indicators provides downloadable data: https://unstats.un.org/sdgs/indicators/database/.
- The United Nations Development Program (UNDP) offers downloadable data for the Human Development Index and hundreds of human development indicators: http://hdr.undp.org/en/statistics/.
- The United Nations Population Division provides downloadable population-related data: https://population.un.org/wpp/
- The World Bank Open Data site gives download access to over 8,000 indicators from World Bank data sets: http://data.worldbank.org/.

6. Change History

The following subsections document all of the changes that have been made to the structure of the *Ethnologue Global Dataset* since it first appeared in 2014 as part of the 17th edition release.

Changes for the 27th edition

There are no changes to the columns of the data tables in the 27th edition release of the *Ethnologue Global Dataset*.

Changes for the 26th edition

The 26th edition of *Ethnologue* reports on digital language support (DLS) for the first time. This release of the *Ethnologue Global Dataset* also adds data on the year the last fluent speaker died for dormant and extinct languages. The changes occur in “3. Table of Languages” (page 9):

Column	Action	Description of change
YLSD	Added	This column reports on the “year last speaker died”.
DLS_Level	Added	This column reports the digital language support (DLS) for each language in terms of a five-level scale.
DLS_Score	Added	This column gives a more refined measue of DLS as a number in the range of 0 to 1.

Changes for the 25th edition

There are no changes to the columns of the data tables in the 25th edition release of the *Ethnologue Global Dataset*.

Changes for the 24th edition

There are no changes to the columns of the data tables in the 24th edition release of the *Ethnologue Global Dataset*.

Changes for the 23rd edition

The 23rd edition of *Ethnologue* has replaced the reporting of a list of"immigrant” languages in each country description with the reporting of a full language-in-country entry for each. A part of that change is to identify these entries more broadly as having a status of “Unestablished” since they are not all explained by immigration. This release of the *Ethnologue Global Dataset* reflects this change by updating the relevant column name in “2. Table of Countries” (page 6):

Column	Action	Description of change
Immigrant	Dropped	The table no longer contains a column by this name. It has been renamed to *Unestablished*.
Unestablished	Added	This column is equivalent to the column that was named *Immigrant* in the previous edition.

Changes for the 22nd edition

There are no changes to the columns of the data tables in the 22nd edition release of the *Ethnologue Global Dataset*.

Changes for the 21st edition

There are no changes to the columns of the data tables in the 21st edition release of the *Ethnologue Global Dataset*.

Changes for the 20th edition

The 20th edition of *Ethnologue* is reporting a new data element, namely, a list of the international conventions pertaining to language rights that each country has adopted. Therefore, this release of the *Ethnologue Global Dataset* introduces the following new column in “2. Table of Countries” (page 6):

Column	Action	Description of change
Conventions	Added	This new column gives the count (out of 9 possible) of the number of intenational conventions related to language rights that the country has adopted.

Changes for the 19th edition

The 19th edition release of the *Ethnologue Global Dataset* introduces three new data elements:

- The categorization of each LIC as being indigenous as opposed to non-indigenous in that country.
- The categorization of each language as being written or unwrittten.
- The inclusion of populations of L2 users where they are known. We originally refrained from including these data due to the criterion of “adequate completeness” discussed in “0. Introduction” (page 3). But we include them now, since customer demand has been so high, even given the known gaps. The EGIDS assessments allow us to know exactly where the gaps are. A language at EGIDS 4 or lower is, by definition, a local language and L2 users are

not expected. However, languages at EGIDS 3 and higher are vehicular and, by definition, they should have a significant number of L2 users. The current coverage of L2 estimates for these cases is as follows:

- 50% (154 out of 308) for EGIDS 1 LICs
- 18% (19 out of 106) for EGIDS 2 LICs
- 30% (78 out of 257) for EGIDS 3 LICs

The addition of the above data elements has caused the following changes in "2. Table of Countries" (page 6):

Column	Action	Description of change
Indigenous	Redefined	Because of the redefinition of *Is_Indigenous* in the Table_of_LICs, this column now produces a different result than it did in the previous edition. It counts the subset of established languages in the country that are truly indigenous. If you still want a column that is equivalent to *Indigenous* as it was in the previous edition of the dataset, use the *Established* column from this edition.
Established	Added	This new column gives the count of established languages in the country. It is equivalent to the *Indigenous* column of the previous edition.

The addition of the above data elements has caused the following changes in "3. Table of Languages" (page 9):

Column	Action	Description of change
Population	Dropped	The table no longer contains a column by this name. It has been renamed to *L1_Users*.
L1_Users	Added	This column is equivalent to the column that was named *Population* in the previous edition.
All_Users	Added	This is a new column. It reports the total of L1 and L2 users and is always equal to or greater than *L1_Users*.
Is_Written	Added	This is a new column which reports whether or not literature in a standardized form is known to have been used by some members of the language community.

The addition of the above data elements has caused the following changes in "4. Table of LICs" (page 14):

Column	Action	Description of change
Is_Indigenous	Redefined	In the previous edition, this column indicated that a language was not recently immigrated. It is now redefined to indicate more precisely that the language is indigenous to the country and not a language from elsewhere that has become established in the country.
Is_Established	Added	This new column indicates that the language has become established in the country, as opposed to being only the language of recent immigrants. It is equivalent to the *Is_Indigenous* column of the previous edition.
Population	Dropped	The table no longer contains a column by this name. It has been renamed to *L1_Users*.
Digits	Dropped	This column is dropped without an equivalent. It can be easily computed from a full population number if such a value is needed.
All_Users	Added	This is a new column. It reports the total of *L1_Users* and *L2_Users*.
L1_Users	Added	This column is equivalent to the column that was named *Population* in the previous edition.
L2_Users	Added	This is a new column that reports the number of L2 users in the country, if it is known.

Changes for the 18th edition

There are no changes to the columns of the data tables in the 18th edition release of the *Ethnologue Global Dataset*. The structure of the tables remains the same as in the original release of the dataset (17th edition).