Following the Legacy of Pioneer .CAT: Identifying World Languages Best Suited for Linguistic gTLDs

Daniel Pimienta, OBDILCI, for puntCAT - 7/1/2025

Note: This study has been contracted by puntCAT to Daniel Pimienta, acting as a consultant. The intellectual property of the study belongs to puntCAT which makes the decision to publish it in cc-by-sa 4.0 license. That license allows others to use, adapt, and share this work, even commercially, as long as they credit both puntCAT and Daniel Pimienta and license their new creations under the same or compatible terms. The term "Extensio" which appears in different place refers to a would-be project of puntCAT using this study.

MANAGER SUMMARY

The aim of the study is to list and hierarchize potential candidates for the definition of a generic TLD based on linguistic and/or cultural consideration, gaining from the pioneer experience of the .cat, which has been followed by some 20 additional cases, mainly in Europe. The methodology used stands on the data produced by OBDILCI model and starts by observing, by region, the languages with more than one million L1 speakers which are not official national languages. The constraint on the number of speakers is, in a second stage, released in one way or the other (checking families of languages, artificial languages, languages with less speakers but yet potential candidates). Some particular situations are observed separately (Africa, Italy, Russia, China, indigenous languages of America, Oceanic region, Creoles). The finally selected 75 languages are spread into 4 categories: top potential, high potential, medium potential and low potential. A complete form gathering key data is filled for the 3 first categories and a matrix with less details gathers the last 2 categories. The existence of a competent counterpart is a key criterion for hierarchy, beyond Internet connectivity and linguistics considerations. The final selection is the following:

TYPE	NUMBER	LIST
TOP	5	Esperanto, French Creole, Friulian, Romani, Tamazight
HIGH	20	Afrikaans, Aymara, Fulfulde, Hakka, Hausa, Hokkien (Nan), Kurdish, Mayan, Nahuatl, Napolitano, Papiamentu, Rohingya, Sami, Sardinian, Sicilian, Swahili, Tagalog, Uyghur, Venetian, Yiddish
MEDIUM	24	Asturian, Bavarian, Emilian/Romagnolo, Eskimo, Gagauz, Gondi, Guarani, Kurux, Iban, Kashubian, Ligurian, Limburgish, Lingala, Lombard, Mandingo, Mapuche, Occitan, Otomanguean, Piemontese, Quechua, Saxon low, Tulu, Yoruba
LOW	27	Algic, Cantonese (Yue), English Pidgin, Extremadurian, Hassaniyya, Hunsrik, Ladin, Muong, Oceania, Okinawan Central, Plautdietsch, Sango, Scottish Gaelic, South African regional (Ndebele, Xhosa, Zulu, Sepedi, Sesotho, Setswana, Siswati, Tshivenda, Xitsonga), Tupian, Upper Saxon, West Flemish, Wolof

+Table of contents

1.TERMS OF REFERENCE (in Spanish)

Fundació PuntCAT considera la posibilidad de desarrollar un programa de cooperación internacional destinado a fomentar y acompañar la creación de una serie de dominios lingüísticos comparables al .cat, en dirección de lenguas que se pueden considerar apropiadas para tales fines. Ese proyecto requiere una primera aproximación a las que podrían ser los criterios para integrar la lista de lenguas apropiadas. A partir de esos criterios, una primera selección de las lenguas candidatas, y, para cada lengua, una ficha descriptiva incluyendo la identificación de potenciales interlocutores a convencer del interés del proyecto. Finalmente, se contempla una estrategia de selección y desarrollo, así como una lista de fuentes potenciales de financiamiento para cubrir el costo de las gestiones correspondientes frente al ICANN, el cual no está necesariamente al alcance de los interlocutores.

La consultoría consiste en proponer un primer nivel de acercamiento a los parámetros de ese proyecto de manera a agilizar y facilitar las decisiones finales de determinación de las lenguas susceptibles de beneficiar de este esfuerzo y la definición de una estrategia para el proyecto.

Por *primer nivel de acercamiento* se entiende la colecta y presentación de una serie amplia de elementos organizados para ayudar a una toma de decisión acertada, tanto en cuento a la selección de lenguas candidatas como la estrategia para superar los obstáculos propios a ese objetivo identificados como, de manera no exhaustiva:

- Posible resistencia de los registradores de ccTLD de los países concernidos que pueden considerar los registradores de dominios lingüísticos como competencia.
- Dificultad de elegir interlocutores reconocidos y representativos en un contexto local donde el consenso entre lingüistas no es la norma y la competencia digital de los lingüistas no se puede considerar como sistemática.
- Tamaño critico de la base de hablantes de las lenguas concernidas (se considera que 96% de las aproximativas 7500 lenguas existentes son habladas por menos de 4% de la población mundial).

La consultoría debe proponer, basado en criterios explícitos, una serie de lenguas candidatas, en tres niveles:

- Lenguas muy indicadas para el proyecto (alrededor de 15)
- Lenguas candidatas potenciales (alrededor de 25)
- Lenguas candidatas posibles (alrededor de 35)

La cantidad y precisión de los datos asociados a cada lengua de ese cuadro en tres niveles, será determinada acorde al nivel (nivel uno, lo más completo y detallado; nivel dos, una ficha descriptiva sintética; nivel tres, algunas líneas descriptivas).
Las fuentes usadas para la consultoría serán los últimos datos demo-lingüísticos de Ethnologue, los datos de conectividad de la UIT, y los datos asociados con el modelo de creación de indicadores para las lenguas en la Internet de OBDILCI, así como otros documentos idóneos que podrán ser identificados y referenciados.
La consultoría tratara de ofrecer cuadros que permiten una gestión organizada de los resultados: por regiones lingüísticas, por tipo de agrupación (lingüística o geográfica), por número de locutores. La tendencia natural en esos tipos de ejercicios siendo de favorecer las lenguas con más locutores, un parte del esfuerzo debe ser orientado a pensar en soluciones de reagrupamiento para lenguas minoritarias.

Los datos reunidos serán colectados vía información accesible en la Web, salvo excepción, la consultoría no contempla contactos directos con interlocutores potenciales.
El producto de la consultoría tendrá la forma de dos documentos en inglés:

- Un documento Word en su versión inicial (para facilitar los comentarios) y pdf en su versión final, articulando una estrategia para el proyecto y reuniendo datos y fuentes de datos pertinentes de la toma de decisión para ese proyecto (en particular datos asociados a las lenguas preconizadas).
- Un documento PowerPoint soporte para una presentación ejecutiva planteando la síntesis de los resultados.

2. INTRODUCTION

Following Ethnologue, 7615 languages exist, more than half with less than ten thousand speakers (L1+L2) and only 447 languages have more than one million speakers. Below some statistics directly computed from Ethnologue Global Dataset #27, 2024.

Table 1: Count of languages per number of speaker's categories

LANGUAGES	COUNT	PORCENTAGE
With more than 10 million speakers	125	1.6 %
With more than 1 million speakers	447	6 %
With less than 1 million speakers	7158	94 %
With less than 100 000 speakers	6157	81 %
With less than 10 000 speakers	4292	56 %
With less than 1 000 speakers	2310	30 %
TOTAL	**7615**	**100 %**

Source: Ethnologue Global Dataset #27, 2024, before grouping macro-languages

Some languages can be regrouped into macro-languages allowing to increase the number of speakers associated (see https://iso639-3.sil.org/code_tables/macrolanguage_mappings/data). For instance, the macro language Arabic (ara) groups 31 variants of Arabic language, which together represents 438 million speakers while only one component (Egyptian Arabic) reaches 100 million. This process transforms, by grouping, 454 individual languages into 61 macro-languages.

The number of languages with minimal digital existence has grown from few tenth in the early years of the Internet to some 750 today, approximative figures computed from the number of languages with codification in Unicode Consortium (https://Unicode.org). Obviously, the level of digital existence varies from just a codification scheme for digital representation, to full presence in a variety of applications or spaces (global application interfaces, translator, syntax analyzer, web contents, IA).

The model created by OBDILCI, which will be used as a basis for that study, works on the basis of macro-languages. It stands on a probable hypothesis which states that some economic law determines the size of the web contents per language. This law would link the demand (the number of connected speakers in a given language) and the offer (the number of webpages in that language). In an ideal world, the link would be linear, proportionality of number of contents with number of connected speakers. In the reality many factors are involved in pushing up or down that proportionality that the model tries to measure.

Following that hypothesis, it becomes obvious that a critical mass is required for such economic law to function. Languages with less than 10 000 speakers are way below that threshold. Where is exactly the threshold to see that law fully performing is not an easy question, is it 100 000 or is it one million?

In the referenced model, only languages with more than one million speakers has been processed, considering that the biases would be too high for figures lower.

In this study, we start with the same approach, however some flexibility will be allowed, considering two factors:

- Many language families exist which have not been defined by a macro-language, preventing them to reach the critical mass (one example is the Maya language of Mexico which does not reach one million speakers, although its language family includes 20 languages whose total speakers is above 2.5 million speakers). We will explore such situations, especially targeting **indigenous languages of America's**.
- Some **existing language oriented TLDs** concern languages with less than one million speakers (as for instance. corsica or .cymru). We will then release the condition and explore, especially in Europe, minority languages with less than one million speakers.

3. METHODOLOGY

What are the criteria required to belong to the list of candidate's languages?

BASIC REQUIRED CRITERIA

- Not being unique official language of a country.
- If it is one of the official languages of a country, not being the majority one and having a notable proportion of speakers in another countries (example: Guarani)
- Having a critical mass of speakers (the threshold is set at one million L1 speakers). Exception: the two exceptions expressed in the introduction and also the case of artificial languages having large L2 speaker's basis (example: Esperanto)
- Having digital existence and presence (being localized and existing strong basis of websites or a strong potential for websites)
- Existence of an institutional direct or indirect representation of the language.

POSSIBLE EXTENSION CRITERIA

- Consistent and rational regrouping of several minority languages (example indigenous languages from a specific region like Canada or Nordic countries or creole languages).
- Languages not rooted in any particular region/country having a solid institutional representation (examples: Esperanto, Yiddish, Romani)

BONUS CRITERIA

- Presence in the upper part of the cyber-globalization table.
- Strong presence in Wikimedia
- Presence in Google Translate
- Existing consensual representation already sensitized to digital challenges
- Other criteria may appear along the process.

What are the chronological different steps the study will cover?

1. Inventory of existing linguistic domains beyond .cat and collect of relevant information for each of them.
2. General understanding of the ICANN process of TLD registration
3. Identify potential language candidates in Europe[1].
4. Identify potential language candidates in Latin America.
5. Identify potential language candidates in Asia.[2]
6. Identify potential language candidates in Africa[3].
7. Identify potential language candidates in Oceania
8. Identify potential language candidates not rooted in a specific region/country
9. Identify potential regrouping of minority or indigenous languages and explore European minority languages (the mentioned exceptions in introduction)
10. PREPARE INTERIM REPORT
11. **For each identified segment, proceed to a selection in 3 levels and gather corresponding information, according to each level, including potential representations and funding**
12. **Check special cases, exceptions and particular situations releasing the fixed rules[4]**
13. Present the findings in a structured fashion
14. Write final report and presentation

The format of the **language form** is the following:

LANGUAGE FORM ISO639-3: Iso639-3 unique 3 characters language code and English name[5]
NAME (English, local): English name of the language and local name(s)
Classification: The selected languages are classified between **** (top candidate), *** (high potential candidate), medium potential candidate (**) and low potential candidate (*). The focus could be on language (*) or culture (x) or mixed (*x).
If macro language: Yes/no. If yes, the different components are listed with code and English name.
L1+L2: The number of speakers first language (mother tongue) + second language
L1+L2/L1: The ratio L1+L2 speakers on L1 speakers. If almost no L1 or no L2 it is specified instead.
Connected L1+L2: The percentage of persons connected to the Internet as computed by OBDILCI model, when available. If not, an approximative value is set.
Countries with speakers: The number of countries with speakers of that language. If more than one, the complete list with associated speakers.

[1] For steps 3 to 7, the results of the OBDILCI model limited to languages with more than one million L1 speakers have served as starter. In further steps, this condition has been released under certain circumstances.

[2] Languages from China, India and Russia will not be considered for selection, with some possible exceptions. The rationale is both pragmatic and acknowledging that those countries have strong local language policies which could easily create conflict of interests with the project.

[3] For Africa, the current situation of deep digital divide for some countries have obliged to introduce a complementary filter, the rate of connectivity to the Internet, considering that rate below 25% is too low for considering the creation of a TLD.

[4] As a matter of fact, steps 11 and 12 has been looping for a while until convergence for final results.

[5] Within the more than 7 000 languages the probability that some name variants are shared by different languages is far from being null, this is why the unique identifier is a must. Note that there are very few cases where another scheme has been used to uniquely specify the language.

Virtual Presence Indicator: The ratio between % of contents and % of speakers, as computed by OBDILCI model. Values much higher than 1 indicates high virtual presence and reciprocally.
Cyber-Globalization rank: Position in the cyber-globalization ranking as computed by OBDILCI.
Wikimedia: Presence in Wikimedia, no, yes (low), fair or high
GoogleTranslate: Presence in the list of GoogleTranslated languages.
Comments:
Pros: Arguments in favor of selection.
Cons: Arguments against selection.
Potential representation: Possible counterparts for partnership with contact.
Potential funding: If specific funding possibility has been identified.
References: Recent documents about the presence of the language on the Internet.
ccTLD: List of involved ccTLDs[6]

Note: In the last control reading of the report it has been decided to add a last parameter to help final decision on selection. A flag has been added to the language's description:
@: means clear representation (if repeated means extremely clear)
#: means difficulty or lack of clear representation.

What are the main sources used by this study?

- https://obdilci.org, a model online under cc-by-sa 4.0 license.
- Ethnologue Data Set #27 May 2024, a proprietary resource protected by copyright for demo-linguistic details.
- https://www.ethnologue.com/language/, https://www.ethnologue.com/country/, https://www.ethnologue.com/subgroup/, Ethnologue open sources online, with member access for more details
- ITU, percentage of persons connected by country.
- Other specific sources will be mentioned in the report when applicable.

Note on Ethnologue: It is considered the most reliable existing source for data on the more than 7500 existing languages. The theme is complex and require a worldwide precise attention that no organization is in capacity to maintain at the level of details required for an exhaustive and high level of confidence. Those data are easily susceptible of errors and sometimes competent local sources could offer more precise or trustable data. Religious motivations[7] at the root of Ethnologue endeavor are often criticized and could encompass a systematic bias. However, no other single source can cover the whole panorama of languages the way Ethnologue is doing and, for the sake of methodology coherence, it is preferable to rely on the same unique source, whenever possible, for all demo-linguistic data. The fact that some

[6] Unless it is otherwise specified, it is highly recommended to negotiate a deal with each mentioned ccTLD registries under the terms that new domains can be registered simultaneously and linking to the same website in both TLD at half the fixed price. In case of a domain already registered in a ccTLD it can apply to the new TLD at half price. The rationale is that any TLD does not create harm to the other one and to offer a win-win-win situation to the user, the ccTLD registry and the new TLD registry, so to avoid harmful competition and trigger synergy.

[7] SIL International (formerly known as the Summer Institute of Linguistics International), the institution responsible for Ethnologue, is an evangelical Christian nonprofit organization whose main purpose is to study, develop and document languages, especially those that are lesser-known, in order to expand linguistic knowledge, promote literacy, translate the Christian Bible into local languages, and aid minority language development. Its headquarters are in Dallas, USA and SIL has sometimes been accused by Indigenous groups of cultural imperialism because the majority of its members are from the USA.

languages are grouped into "macro-languages" allow them to reach the threshold of one million L1 speakers that we have set; other families or group of languages, not defined as macro-languages at this moment, could be considered penalized for not beneficiating of such regrouping. We have tried to extend the analysis, especially for indigenous languages, mainly in America, into those possible grouping.

4. OVERVIEW OF gTLD REGISTRATION ICANN PROCESS

4.1 gTLD registration

Source: https://newgtldprogram.icann.org/en

The process of gTLD registration is both complex and costly, the cost being direct (the fees required by ICANN) and indirect (the cost of the required processes to support the application). Both direct and indirect costs are in the order of some hundreds of thousands of US$. Furthermore, invested costs and fees are not recovered in case the application does not conclude positively.

The application fee has been set recently at 227 000 US$ but it is not the only fee which apply, additional fees has to be expected, such as:

Conditional Fees:

- **Community Priority Evaluation (CPE):** For applicants seeking community-based TLD status.
- **Community Registration Policies Review (Specification 12):** Assessment of specific registration policies for community TLDs.
- **Geographic Name Review:** For applications involving geographic names, requiring additional scrutiny.
- **Brand Exemptions (Specification 13):** For applications seeking Brand TLD status.
- **Code of Conduct Exemption:** Requests to be exempted from certain registry operator codes of conduct.
- **Reserved Names Review:** Evaluation related to names reserved under ICANN policies.
- **Re-evaluations Due to Change Requests:** If changes are made to the application that necessitate re-evaluation (e.g., background screening).
- **Limited Challenges/Appeals:** Fees associated with challenging or appealing evaluation results.
- **Registry Voluntary Commitments (RVC) Review:** Assessment of voluntary commitments made by the applicant.
- **Name Collision High-Risk Mitigation Plan Review:** For strings identified with potential name collision risks, requiring mitigation strategies.
- **"Occupancy" Fee for Lingering Applications:** Applicable to applications that remain in the process for extended periods.
- String variants for **internationalized domain names** (IDNs)
- **Trademark-related review**

The exact fees for these conditional evaluations are determined based on the effort required and are communicated before the application window opens.

Post-Delegation Fees:

- **Annual Registry Fees:** Ongoing fees for operating the gTLD registry. For reference, the 2024 base registry agreement outlines these fees in section 6.1.
- **Trademark Clearinghouse (TMCH) Fees:** Fees associated with trademark protections and services.

4.2 Applicant Support Program (ASP)

Source: https://www.icann.org/en/announcements/details/icann-opens-application-period-for-new-gtld-applicant-support-program-19-11-2024-en

The good news is the existence of an **Applicant Support Program (ASP)** which is exactly tailored for the candidates for language TLD. Applicants are supposed to be non-for-profit organizations. Success in applying to this program could allow a reduction of the fee of the order of 85% and access to facilities which could reduce considerably the indirect cost of the actions required to conduce the process.

ICANN offers financial and non-financial assistance to eligible applicants through the ASP, which provides:

- **Access to Pro Bono Services:** Volunteer professional services to assist with the application process.
- **Training and Resources:** Materials to help applicants understand the gTLD application and evaluation processes.

4.3 ICANN grant program

Source: https://www.icann.org/grant-program-en

Would this program be reconducted in 2025, it will provide an opportunity for funding within the realm of ICANN and it should be considered as a priority option for funding, the maximum amount (500 K$) being in the order of magnitude of the total process cost.

Note: The focus of the consultancy is on selecting language's candidate and is therefore not oriented towards the details of that process, for which Punto.cat has all the possible experience and skills; this section is only mentioned as an introduction. The same for the important question of the threshold, in terms of number of domain registered, for the return on investment of the creation of a TLD, which is not treated in the study.

5. EXISTING TLDs ANALOGOUS TO .CAT

Some TLD domains dedicated to specific languages exist, others are dedicated to a region with an associated language and another group have a looser connection with language. They are listed below, associated with some parameters. Note that IDN ccTLD are not listed below (such as .ελ for .gr of Greece) as they are not linguistic or cultural gTLD.

Sources:
https://en.wikipedia.org/wiki/List_of_Internet_top-level_domains for identification
https://icannwiki.org for general information

https://domainnamestat.com or https://zonefiles.io/detailed-domain-lists/ for number of registered domains.
https://en.wikipedia.org/wiki/Geographic_top-level_domain

The gTLDs of the following table are more or less explicitly linked to a specific language. In parenthesis the number of active domains when the information has been found. The first data represents all the registered domains, following https://domainnamestat.com, the second data, in parenthesis, is the number of active ones, only mentioned when the data has been obtained from the NIC website. The third number in red color is the L1+L2 population of speakers of the concerned language which is used to create the second table to compare the performance by speaker of each TLD.

Table 2: Existing linguistic TLDs

gTLD	Language	IDN	Since	Registered domains	Comments
.alsace	Alsatian		2014	4219 (1 450) **0.9**	Cover more the region of France than the language
.bzh	Breton		2014	19 277 (12 000) **0.2**	Dedicated to both the language and Brittany a region of France
.cat	Catalan		2004	222 237 (112 000) **9.3**	Dedicated to the language
.corsica	Corsican		2015	3 800 (1 300) **0.12**	Cover both the region of France and the language. Managed by local government.
.cymru	Welsh		2014	15 770 **0.55**	More oriented towards Welsh language than Wales
.eus	Basque		2014	24 348 (9 000) **1.1**	Cover both Basque language and culture
.frl	Frisian		2014	10 307 **1.2**	For the Friedland region of Netherlands
.gal	Galician		2014	10 891 (6 500) **3.4**	Cover both Galician language and culture
.irish	Irish		2015	10 626 **1.2**	More oriented towards culture than language
.scot	Scot and Gaelic		2014	12 000 **1.7**	Cover both Scotland, the region of United Kingdom, and the languages.
.shiksha	Hindi		2013	2 525 (1 000)	For education purpose
.wales	Wales and Welsh		2014	27 722 (12 771) **.6**	Cover both Wales, the region of United Kingdom, and the languages. Managed by Welsh government
.みんな	Japanese	xn--q9jyb4c	2013	3 910	
.krd	Kurdish		2014	1 831	Geographic to Kurdistan region of Iraq

.arab	Arabic		2015	1 884	More oriented towards culture than language. Managed by League of Arab States.
عرب	Arabic	xn—ngbrx	2015	147 102	Same management than .arab
.شبكة	Arabic	xn--ngbc5azd	2015	1 248	.shabaka
.موقع	Arabic	xn--4gbrim	2015	464	
网址	Chinese	xn--ses554g	2014	296 398	
政府	Chinese	xn--mxtq1m	2015	184 411	.gov
公司	Chinese	xn--55qx5d	2014	71 945	
我爱你	Chinese	xn--6qq986b3xl	2014	65 555	
商标	Chinese	xn--czr694b	2014	62 219	
网络	Chinese	xn--io0a7i	2014	49 418	
商城	Chinese	xn--czru2d	2014	37 654	
手机	Chinese	xn--kput3i	2014	36 803	Oriented toward mobile users
中文网	Chinese	xn--fiq228c5hs	2014	18 349	
移动	Chinese	.xn--6frz82g	2014	14 047	Oriented toward mobile users
集团	Chinese	xn--3bst00m	2014	10 083	
购物	Chinese	xn--g2xx48c	2016	9 192	e.commerce
网店	Chinese	xn--hxt814e	2014	8 648	e.commerce
	Chinese				They are some 10 additional Chinese IDN with less than 5 000 registered domains
онлайн	Cyrillic alphabets	xn--80asehdb	2013	7 203	
сайт	Cyrillic alphabets	xn--80aswg	2013	4 551	
.рус	Russian	xn--p1acf)	2014	110 385	Dedicated to ethnic community of Russian-speaking people who originated in Kiev Rús in the 13th century. This includes but is not limited to residents of Belarus, Kazakhstan, Norway, Russia, Ukraine and the United States

Other gTLD exist which may have some connection with specific languages, although it is not clear enough from the documentation that there is a linguistic purpose. They have not been included in this table: .okinawa, .ryukyu, .quebec, .kiwi, .africa, .capetown**,** .joburg**,** .durban, .vlaanderen.

Apart of that, IDN transliterations of .com & .net has been defined in the following languages: Thai, Deva, Korean (Hang), Chinese (Hans/Traditional & Hans/Simplified), Hebrew, Russian, Arabic, Japanese. Note that a request for .thai was rejected.

To conclude this part, some regional domain exist which can be indirectly connected to various languages: .asia, .africa and .lat (for Latin America).

What this compilation shows is that the number of registrations per population of speakers vary in a large factor of 1 to 30 with .cat being in the middle of the table. In absolute terms, few linguistic domains have managed to congregate large number of subscribers, apart .cat and .pyc, and this situation has to be considered in the Extensio project. It also shows that .cat was the very early promotor of the concept. Maybe, the Extensio project should add a component directed towards already existing linguistic domains in terms of providing learnt lessons experience and advices to boost the results which seems quite modest so far for the lower part of the following table.

Table 3: TLDs number of registrations per 1000 speakers

TLD	Registration/ 1000 speakers
.bzh	96,39
.wales	46,20
.corsica	31,67
.cymru	28,67
.cat	23,90
.eus	22,13
.fls	16.09
.irish	8,86
.scot	7,06
.alsace	4,69
.gal	3,20

6. CANDIDATES LANGUAGES

6.1 First stage: pre-selection

The first stage of the study creates a first selection, following the established rules, and using data from Ethnologue and the OBDILCI model, expanded in a working Excel file. The first selection classifies the selected languages into 4 categories:

- Top potential languages (****)
- High potential languages (***)
- Medium potential languages (**)
- Checking list to determine low potential or discard (*)

The orientation could be towards linguistic matters and then marked * or towards cultural (or regional) matters and then marked x, if both it is marked with combination.

According to each category, the second stage of the study process will be to establish a form whose level of details vary from extremely detailed to brief, depending on the category. For the last category (*), the second stage will make an informed decision to put in one of the previous categories or to discard.

Note that the split into categories decided in the first stage is only transitory and changes of category can be decided during the second stage of the study, especially if obstacles arise, especially on the critical issue of searching for representative counterparts.

During the second stage, most languages classified in categories ****, *** or ** will be described in the form defined below:

Table 4: Language form

LANGUAGE FORM ISO639-3:
NAME (English, local):
Classification:
If macro language:
L1+L2:
L1+L2/L1:
Connected L1+L2:
Countries with speakers:
Virtual Presence Indicator:
Cyber-Globalization rank:
Wikimedia:
GoogleTranslate:
Comments:
Pros:
Cons:
Potential representation:
Potential funding:
References:
ccTLD:

Languages classified *, will be added to the cumulative table below with essential parameters:

Table 5: Matrix form

ISO	LANGUAGE	COUNTRY	L1+L2	%C	M/F/G	Cw. Sp	W	GT	COMMENTS

Where:

- ISO is the 3 characters iso code 369-3
- LANGUAGE is the English name of the language
- COUNTRY is, if applicable, the main country of speakers
- L1+L2 is the total number of speakers
- %C is the percentage of speakers connected to the Internet
- M/F/G will specify if it is a Macro-language, a Family of languages or a Group of languages
- CwSp: Number of countries with speakers

- W: Wikimedia presence: No, Yes, Fair, Good
- GT: Google Translate presence: Yes or no.

Note that the Excel file which served to conduct that selection is an internal working resource for the consultant may contain proprietary data from Ethnologue. It can be provided, on request by Fundació Punto.cat, but cannot be shared to third persons and must not be published.

The process of pre-selection started from the last results from OBIDILCI model with 362 languages (L1 > 1M) all associated with a set of useful parameters. The first stage was to eliminate official national languages, with few exceptions. From the remaining languages, the results are split into regions (Europe, America, Africa, Asia, Oceania) and some additional parameters are added from other sources (presence in Wikimedia and in GoogleTranslate). The pre-selection, with hypothetical notations to be confirmed, is done in the working Excel file.

In a second step, the list has been extended with a look to categories not included in the OBDILCI model:

- European languages with less than 1M speakers, but anyway possible candidates.
- Families or groups of languages not defined as macro-languages and then possibly outside of OBDILCI model. A systematic analysis of America's indigenous languages is part of that step (Brazil, Mexico, Colombia, North America) as well of creole languages.
- Languages not anchored in any specific country (artificial languages, languages such as rom).

A look at the cyber-geography table produced by OBDILCI, which split and gather the indicators by region is useful at this stage (remembering it concerns only languages with L1 speakers higher than one million):

Table 6: Cyber-geography of language families

LANG. FROM >	Africa	Americas	Arab world	Asia	Europe	Pacific	Not Incl.	TOTAL
Internauts %	36,3%	69,0%	68,0%	60,8%	89,2%	62,2%	49,53%	64,19%
Contents	4,18%	0,29%	3,63%	44,28%	45,87%	0,03%	1,73%	100%
Virt. Pres.	0,36	0,83	0,89	0,92	1,43	0,52	0,51	1
Cont. Prod.	0,63	0,81	0,84	0,93	1,19	0,73	0,66	1
POP.L1+L2	11,68%	0,35%	4,08%	48,37%	32,06%	0,05%	3,42%	100%
%POP. CONN.	6,66%	0,35%	4,33%	47,46%	38,53%	0,04%	2,64%	100%
NUM.LANG	152	8	1	147	51	2		361

Source: OBDILCI V5.2 - https://obdilci.org/wp-content/uploads/2024/11/RESULTS-5.2.xlsx

6.1.1 African languages pre-selection

The figures of 36% for the percentage of connected speaker for African languages as well as the figure of 0.36 for the Virtual Presence indicator are a call for caution in the selection, in spite the fact that the list of African languages with more than one million speakers is exceptionally large (152 languages).

A first filtering has been to select for scrutiny the list of languages with high number of countries with speakers, filtering out those which virtual presence is lower than 0.3 and national languages. Those represents potential high candidates for their role as lingua franca at regional level, inside a country or between countries.

Table 7: African languages spoken in many countries

ISO	Language	VIRT.PRES.	L1+L2	L2	L1+L2/L1	NB. Co.	Main country
afr	Afrikaans	0.83	18 093 000	7 778 400	2.33	**15**	South Africa
wol	Wolof	0.65	22 646 100	7 139 820	3.17	**13**	Senegal
swa	Swahili Macro	0.37	97 658 480	5 265 080	18.55	**24**	Tanzania
snk	Soninke	0.36	2 280 700	2 280 700	1.00	**8**	Mali
man	Mandingo Macro	0.36	9 134 300	9 134 300	1.00	**7**	Guinea
yor	Yoruba	0.35	47 195 900	45 171 300	1.04	**18**	Nigeria
sna	Shona	0.35	10 877 780	7 375 510	1.47	**9**	Zimbabwe
mey	Hassaniyya	0.35	5 206 980	5 206 980	1.00	**10**	Mauritania
ibo	Igbo	0.34	30 913 160	30 903 800	1.00	**7**	Nigeria
ful	Fulfulde Macro	0.33	39 955 240	37 275 240	1.07	**19**	Senegal
bam	Bamanankan	0.30	14 188 850	4 186 390	3.39	**7**	Mali
lin	Lingala	0.30	40 540 300	20 522 160	1.98	**11**	Democratic Rep. of the Congo
hau	Hausa	0.29	88 238 080	53 829 060	1.64	**18**	Nigeria

More attention will be then given to the languages of African countries with the higher connectivity rates, see table below, and a special attention will be given to Morocco, for its very high rate (91%) and South Africa, for its demographic and economic importance apart of its high connectivity rate (75%).

The below table is established from the list of African countries with a connectivity rate higher than 50%, and list African languages with more than one million speakers in those countries, except those which are official national language in some other country. Those languages, together with those of Morocco and South Africa, plus the languages with serve as lingua franca will receive further analysis for Africa.

Table 8: African languages selection based on countries well connected

COUNTRY	% Conn.	Iso, Language, speakers, main country if different
Algeria	71	kab Kabyle 7.9
Djibouti	65	
Egypt	72	
Equatorial Guinea	67	fan Fang 1.1
Eswatini	58	
Gabon	73	
Gambia	54	mnk Mandinka 2 fuk Pulaar 6.3
Ghana	70	abr Abron 1.4 Côte d'Ivoire aka Akan 9.9 ada Dangme 1 ewe Éwé 5.5 yor Yoruba 47.2 Nigeria

Libya	78	
Mauritius	76	
Namibia	62	
Saudi Arabia[8]	100	swa Swahili macro 5.3 Tanzania
Senegal	60	mnk Mandinka 2M fuk Pulaar 6.3 bam Bambara 14 ful Fulah macro 37M (incl. Pulaar) mey Hassaniya 3.8 Mauritania mlq Maninkakan, Western 2 fuf Pular 4.8 Guinea srr Serer-Sine 1.9 Wolof wol 22M

The final list obtained after stage 1 is consigned below.

Top and High potential languages (** or ***)**

Table 9: First selection for top and high candidates

EUROPE	AMERICAS	AFRICA	ASIA	
Bavarian	Quechua Macro	Swahili Macro	Tagalog	Yiddish
Saxon, Low	Guaraní Macro	Hausa		Esperanto
Occitan	Hunsrik	Berber family		Romani
	Aymara Macro			Kreyol (French creole)

Medium and low potential languages (**) or (*)

Some of the following languages could be discarded after scrutiny or pushed up in category.

Table 10: First selection for medium and low candidates

EUROPE	AMERICAS	AFRICA	ASIA	OCEANIC
Italian local	Mayan	Yoruba	Tatar	Tok Pisin
Aragonese	Otomanguean	Afrikaans	Chechen	Papua New Guinea[9]
Asturian	Uto-Aztecan	Fulfulde Macro	Okinawan, Central	Hawaian
West Flemish	Tupian	Amharic	Kurdish	Maori
Frisian	Chocoan	Wolof	Uyghur	
Sorbian		South-African local	Cebuano	**NOT LOCATED**
Saxon, upper	Jivaroan	Mandingo	Ilocano	Pidgin (English creoles)
Limburgish		Kanuri Macro	Armenian, W	Portuguese creoles

[8] Saudi Arabia, as a matter of fact, gathers very large population of speakers of Asian languages (Balochi, Bengali, Rohingya, Hindi, Pashto, Punjabi, Nepali, Saraiki, Tagalog, Tamil, Telugu, Urdu…)

[9] Papua New Guinea, with a population of over 9 million, split between 600 islands, holds 852 languages, the large majority of them with less than 1 000 speakers. The ones with most speakers (over 50 000) and playing a role of lingua franca have been retained in first preselection. The Internet connection rate is only 27% so probability to remain in the list are low.

Cornish	Maipurean	Tigrigna	Rohingya	Other artificial languages
Gagauz	Algic		Muong	
Rusyn	Eskimo	Hassaniyya	Tibetan, Central	
Kashubian	Eyak-Athabaskan	Soninke	Yue	
Uralic		Lingala	Hakka	
Plautdietsch		Sango	Nan	
Sami				

At the end of stage 1 we have gathered:
5 **** languages
10 *** languages
70 languages to be determined if categorized or discarded

6.2 Second stage: individual analysis, classification and form writing

Each pre-selected language is individually analyzed and the parameters are completed. For most of the languages finally classified as Top potential (****), High potential (***) or Medium potential (**) a complete form is filled in annex 1. The rest of the pre-selected languages is either discarded or kept as Low potential language (*), documented in Annex 2 with less details.

Some situations have been processed separately, dur to their particular characteristics; they are exposed in sequence.

6.2.1 Italy

Italy which has so far not assigned a TLD to any of its local languages. It is uneasy to distinguish between the Italian local languages and they have been processed separately below.

Table 11: Local languages from Italy

ISO	Language Name	Country	Users L1	TOTAL	STATUS	WIKI	GT	REPRES.	TYPE
nap	**Napolitano xx***	Italy	5 700 000	**5 700 000**	Developing	YES		LINK	L&C
scn	**Sicilian xx***	Italy	4 700 000	**4 700 000**	Developing	Yes	Y	LINK	L
vec	**Venetian xx***	Italy	3 800 000	**3 852 500**	Developing	Yes		LINK	L
vec	Venetian	Croatia	50 000		Developing				
vec	Venetian	Mexico	2 500		In Trouble				
vec	Talian	Brazil			Dying				
vec	Venetian	Slovenia			Dying				
lmo	**Lombard *x**	Italy	3 600 000	**3 903 000**	Vigorous	Yes	Y	LINK	L&C
lmo	Lombard	Switzerland	303 000		Vigorous				

pms	**Piedmontese x***	Italy	700 000	**700 000**	In Trouble	Yes		#	
pms	Piedmontese	Argentina			Dying				
fur	**Friulian ********	Italy	600 000	**600 000**	In Trouble	Yes		LINK @	**L&C**
src	**Sardinian**, xx* Logudorese	Italy	500 000	**1 200 000**	In Trouble	Yes		LINK	R
sro	Sardinian, Campidanese	Italy	500 000		In Trouble				
sdc	Sardinian, Sassarese	Italy	100 000		In Trouble				
sdn	Sardinian, Gallurese	Italy	100 000		In Trouble				
egl	**Emilian ****	Italy	440 000	**440 000**	In Trouble	Yes		#	
rgn	**Romagnol ****	Italy	160 000	**164 120**	In Trouble	Yes		#	
rgn	Romagnol	San Marino	4 120		Dying				
lij	**Ligurian ****	Italy	140 000	**148 420**	Developing	Yes	Y	#	
lij	Monégasque	Monaco	8 420		Institutional				
lij	Ligurian	France			Dying				
lld	**Ladin ***	Italy	38 000	**38 000**	In Trouble	Yes		LINK	L&C

The table lists all local languages of Italy with respective parameters. It is difficult to highlight one particular language, because the parameters are close, and even when a potential representation has been identified, none is really convincing, except for Friulian which have an institution really dedicated to this language: ARLEF - http://www.arlef.it/. Friulian has been added to the top candidates as a TLD could be an opportune strategy to help existing efforts to get it out of the threatened category, and because it is a statutory language of provincial identity in Friuli-Venezia Giulia autonomous region. Obviously, the decision to take or not that opportunity belongs to ARLEF.

Each of those languages is then classified as medium potential candidates for language and culture (*x). Acknowledging some additional features, the following 5 ones could be classified as high potential (**x) if a motivated counterpart could be identified:

- Neapolitan and Venetian, if the cities of Naples or Venice decides to create a geographic TLD, as have been done by some other famous cities in the world.
- Sardinian and Sicilian, because they are related to an island culture.

6.2.2 Oceanic Region

It has been difficult to select some languages from Oceanic region. Tok Pisin, from Papua New Guinea, is some sort of lingua franca inside the island but not spread abroad, except in Australia with very few speakers. Among the 850 languages of Papua New Guinea, a first selection has been made of the most spoken[10], but with an Internet connection rate of less than 30% it is not

[10] Dobu, Golin , Huli, Kuanua, Melpa, Motu, Hiri, Yuwei , Kâte, Enga, Kuman, Kamano, Kewapi, East.

possible to retain any with confidence. As a matter of fact, as a general rule, except Tok Pisin and Iban (iba), a language born in Borneo which migrated to Malaysia, no Oceanic language reaches one million speakers. Iban could be considered as a region-oriented candidate as it is mainly located in the specific region of Arawak, but Malaysian authorities will probably have a say; anyhow, a form has been filled out for Iban in medium category. Hawaiian has too few speakers to be considered.

The possibility for not leaving oceanic languages out of the study remains in paying tribute to its huge diversity of languages, although with low number of speakers, and dedicate a unique specific domain for a very large number of languages, something like .oceania.or, as an alternative, splitting it into .melanesia, .micronesia and .polinesia. Refer to the Spanish Wikipedia article https://es.wikipedia.org/wiki/Oceanía for more details about country and dependencies concerned. To retain the idea, .oceania has been added to the low category.

6.2.3 Russian Federation

Russian Federation encompasses a large set of 102 indigenous languages. Given the political context in Russia, but also given the fact that there is, historically, from the Soviet Union period and still today, strong state linguistic policies in support of many of those languages, they are not included in the study[11].

6.2.4 China

China holds 281 indigenous languages, of which the macro language Chinese (zho) groups 16 of them and 96% of speaker's population. China has its own Internet and language policies and except very few exceptions of languages spoken in China but with a large proportion of speakers outside, this study is not covering those languages. The only languages looked after are therefore:

- Chinese Yue (yue), 87M of speakers, mostly L1, in 35 countries, 15% of speakers out of China (finally classified Low)
- Chinese Hakka (hak), 44M speakers, mostly L1, in 20 countries, 17% of speakers out of China (finally classified High)
- Chinese Min Nan (nan), 51M speakers, mostly L1, in 15 countries with 44% of speakers out of China (de facto language of provincial identity in Taiwan). Also known as Hokkien it has finally been classified High.

Note: those 3 languages are included in zho macro-language.

6.2.5 India

India is second of population after China (1.38 vs. 1.41 billion) but is forecasted to become soon the most populated country on earth. India gathers 464 languages, of which 424 are living indigenous, of which some 50 have more than one million speakers. Twenty-two languages have received official status (such as Hindi, Bengali, Telugu, Marathi, Tamil, Urdu, Gujarati, Malayalam, Kannada, Odia, Punjabi, Assamese or Maithili). It is not easy to select languages as candidates for gTLD in such a wealth of languages and this could justify a specific study. If

[11] https://www.gencat.cat/llengua/noves/noves/hm02primavera/internacional/a_marc.pdf Extract: "From a sociolinguistic point of view, the outcomes of the Soviet nationality policies can be summed up as follows: "La politique linguistique est sans aucun doute le plus original de l'action menée par le pouvoir en matière nationale. C'est aussi, cela est certain, sa plus parfaite réussite".

the focus is set on regional language of more than one million speakers but threatened, a TLD seen as a mean to contribute for revitalization, some languages could be selected, this initial list would need to be completed if the interest arises to target specifically India.

Table 12: Local languages in India

LANGUAGE	SPEAKERS	COMMENTS	LINK
Kurux (kru) *x	2.1M (almost no L2)	Spoken by the Oraon tribe and has its own script called Tolong Siki.	Link
Tulu (tcy) *x	1.85M No L2	Localized in Karnāṭaka and Kerala States, with rich literary tradition and strong cultural presence	Link1 Link2
Gondi (gon) *x	2.4M No L2	Macro-language (wsg+esg+gno) Spoken by the Gond tribe, one of the largest indigenous communities in India.	Link1 Link2

6.2.6 Indigenous languages of Americas.

A selection has been made and forms filled for a diverse selection of those which could present the best chances of success: Aymara macro, Quechua, Mayan family, Nahuatl family and Guarani.

For Mayan and Nahuatl, a note invites to check with INALI (https://www.inali.gob.mx/) for the possibility to extend further the family selection: "*Would discussion with INALI be initiated about the possibility of TLD for Mexican indigenous languages, they should be extended to the cases of another possible candidate: Otomanguean, family of 176 languages of Mexico including Zapotec and Mixtec, with 1.8M speakers*". For Guarani, it is recommended to check the possibility to extend to the Tupian family (Brazil).

For the rest, the following table gather the prospects. The idea to select a family of languages instead of a single language is to allow more easily to reach a critical mass of speakers. However, when in spite of this effort, the number of speakers remain low, it is hard to recommend a TLD, except if special conditions which are beyond the capacity of this study to be identified could be found by further studies. Note that this table is not exhaustive and languages belonging to Ecuador, Venezuela, Colombia, Guyana, Argentina has been lest aside.

Table 13: Indigenous languages of Americas

LANGUAGE FAMILY	COUNTRY	#L	SP. M.	COMMENTS	
Otomanguean	Mexico	177	1.7	Incl. Mixtec and Zapotec.	*x
Uto-Aztecan	Mexico & USA	63		Nahuatl belongs here. The rest has no critical mass.	
Mixe-Zoquean	Mexico	14	0.17	No critical mass	
Many families[12]	Brazil Colombia			No critical mass	
Tupian	Brazil	76		Close to Guarani, see form.	*

[12] Arauan, Bororoan, Cariban, Jean, Maipurean, Panoan, Puinavean, Cariban, Yanomaman,..

Mapudungun (Mapuche)	Chile, Argentina (arn and huh)	1	0.26	Oral but written form has been created and is promoted especially in Internet. https://es.wikipedia.org/wiki/Consejo_de_Todas_las_Tierras https://uchile.cl/noticias/220046/u-de-chile-y-wikimedia-promueven-la-cultura-mapuche-en-internet	*x
Algic	Canada & USA	48	0.15	Include macro language Cree (cre) with 18.5K and macro language Ojibwa (oji) with 80K	*
Eyak-Athabaskan	Canada & USA		0.25	Include Navajo (nav) with 150K and Apache family with 14K	
Eskimo	Greenland, Canada, USA, Russia	11	0.12	Mainly Kalaallisut (Greenland) 51K, Inuktitut (Canada) 42K, Yupik (Alaska) 20K.	*x

Many endangered languages belong to those families and a TLD could be a strategy to preserve some and boost other languages, providing a centralized representation to handle such process or alternatively the collaboration between organizations from different countries.

As an example, for Eskimo family, the following institutions would need to cooperate **u**nder the umbrella of Inuit Circumpolar Council (ICC) (inuitcircumpolar.com) which represents Inuit communities in Canada, Greenland, Alaska, and Russia:

- Greenland Language Secretariat (oqaasileriffik.gl)
- Inuit Tapiriit Kanatami (itk.ca)
- Pirurvik Centre (pirurvik.ca)
- Alaska Native Language Center (ANLC) (uaf.edu/anlc)
- Alaska Native Language Preservation and Advisory Council (often cited but no website found)

This example shows the paradox for TLD for indigenous languages, an initiative which could contribute for decisive progress for those languages in their online presence but at the same time quite complex to setup due to the spread of languages with low number of speakers among many countries.

The Indigenous Decade activities promoted by UNESCO could be a place to extend and deepen another specific study for those languages.

The Wikipedia article https://en.wikipedia.org/wiki/Indigenous_languages_of_the_Americas is a sound basis to explore further this theme as it provides exhaustive listing of languages and concerned countries.

6.2.7 Creoles

A creole language is a stable natural language that develops from the process of different languages simplifying and mixing into a new form, expanding and elaborating into a full-fledged language with native speakers. Ethnologue treats creole as a single family and have identified 92 different creoles in function of the languages which has been creolized; and which allow to name the creole (French based creole or Malay based creole).

For the sake of that study, only those gathering high number of speakers or presenting a particular interest are analyzed in the following table.

Table 14: Preselected creoles

LANGUAGE BASE	VARIETIES	SP. M	COMMENTS
English	**33**	**162**	**Sum of pidgins from Nigeria (121), Cameroon (12), Sierra Leone (8), Ghana, Liberia, Papua New Guinea, Jamaica and Hawaii accounts for 158M.**
French	**11**	**17**	**Haitian creole accounts for 11.**
Arabic	2 (pga, kcn)	1.5	
Iberian	**Papiamentu (pap)**	**.32**	**Curacao, Aruba, Netherland Caribbean, Sint Maarten and Netherlands**
Kongo	Kituba (ktu+mkw)	14	Congo and Democratic Republic of the Congo
Malay	14	13	Indonesia (99%), Sri Lanka, Singapore, Netherlands, Malaysia and Coco island.
Ngbandi	**Sango (sag+snj)**	**5.2**	**Central African Republic**
Portuguese	13	3.3	68% between national languages of Guinea-Bissau and Cape Verde (Kabuverdianu -kea).
Tetun	Tetun Dili	1	East Timor

Only those marked in bold have been studied, the other being too much linked to a small number of countries (one or two). From those, the following resulted:

- French creoles have been set in the Top category (see form)
- English creoles are a complex piece with major speakers as national languages of a reduced number of countries. It has been put in category low potential (under the alternative name of English Pidgin) and it would deserve a deeper and particular study.
- Sango exists in Wikimedia and GoogleTranslate, however it has too weak the virtual presence indicator (0.06) at this stage. It could be a candidate in the future when digital divide is overcome in Central African Republic which holds today a tiny connection rate of 11%.
- Papiamentu has been set as a high potential candidate, given it is quite lively with strong presence in media and the Internet, including Wikimedia and GoogleTranslate. Curacao is without doubt a world model country for multilingualism, with kids being trained in 4 languages (Papiamentu, Dutch, Spanish and English), and such a project could be well received by National Language Institute of Curaçao (https://nticuracao.org/) the institution owning such decision.

7 FINAL COUNT

The final results are summarized in the tables below.

Table 15: Final list of candidates per type

TYPE	NUMBER	LIST
****	5	Esperanto, French Creole, Friulian, Romani, Tamazight
***	20	Afrikaans, Aymara, Fulfulde, Hakka, Hausa, Hokkien (Nan), Kurdish, Mayan, Nahuatl, Napolitano, Papiamentu, Rohingya, Sami, Sardinian, Sicilian, Swahili, Tagalog, Uyghur, Venetian, Yiddish
**	24	Asturian, Bavarian, Emilian/Romagnolo, Eskimo, Gagauz, Gondi, Guarani, Kurux, Iban, Kashubian, Ligurian, Limburgish, Lingala, Lombard, Mandingo, Mapuche, Occitan, Otomanguean, Piemontese, Quechua, Saxon low, Tulu, Yoruba
*	26	Algic, Cantonese (Yue), English Pidgin, Extremadurian, Hassaniyya, Hunsrik, Ladin, Muong, Oceania, Okinawan Central, Plautdietsch, Sango, Scottish Gaelic, South African regional (Ndebele, Xhosa, Zulu, Sepedi, Sesotho, Setswana, Siswati, Tshivenda, Xitsonga), Tupian, Upper Saxon, West Flemish, Wolof

Table 16: Repartition of candidates per region

	AFRICA	AMERICA	ASIA	EUROPE	OCEANIA	NO-GEO	TOTAL
****	1			1		3	5
***	4	4	6	5		1	20
**	3	5	3	12	1		24
*	12	3	3	6	1	1	26
TOTAL	**20**	**12**	**12**	**24**	**2**	**5**	**75**

Table 17: Countries concerned by candidates TLD[13]

TYPE	COUNTRIES
****	Netherlands, Haiti, Italy, Latvia, Morocco
***	South Africa, Bolivia, Nigeria (2), Taiwan (2), France, Mexico (2), Italy (4), Curacao, Malaysia, Norway, Zanzibar, Philippines, Germany, USA
**	Spain, Germany, Italy (4), Greenland, Moldova, India (3), Paraguay, Malaysia, Poland, Netherlands, Democratic Republic of the Congo, Guinea, Chili, France, Mexico, Peru, Germany, Nigeria
*	Canada (2), USA, Nigeria, Spain, Mauritania, Brazil, Italy, Vietnam, Oceania, Japan, Central African Republic, UK, South Africa (9), Paraguay, Germany, Belgium, Senegal

8 AI SIDE EXPERIMENT

In parallel with the study, an experiment has been conducted to check the capacity of ChatGPT (version with Internet search capacity) in this very specific and specialized subject, on the edge of Internet and linguistics. Apart from helping, in several opportunities, in identifying, faster than in a personal search, the potential institutions to represent a TLD candidate, ChatGPT cannot be credited of contributing significantly to the product. However, the fact is that its performance has been quite impressive, confirming in more than 80% the decisions or orientations taken[14], and showing thorough "artificial knowledge" of the wealth of languages as well as apparent expertise in the domain naming theme, with several cases of useful proactivity, anticipating potential future questions. It seems that ChatGPT have wisely taken profit from the excellent Wikipedia pages on languages as well as from the public part of Ethnologue knowledge. The capacity to combine and intersect both sources in giving the appearance of expertise in such a very specialized subject remains far above the expectations and absolutely impressive.

[13] The countries associated are the one of the potential representations

[14] Less than 10% of the case can be considered as gross mistakes, including very few hallucinations.

ANNEX 1: LANGUAGE FORMS FOR FIRST LEVELS LANGUAGES

LANGUAGE FORM ISO639-3: epo (Esperanto) **** @@

NAME (English, local): Esperanto, Esperanto, Lingvo Internacia
Classification: Most popular constructed language, top candidate for language (****)
If macro language: no
L1+L2: 101 000 (incl.1 000 L1) however other sources claim between 1 and 2 million users
L1+L2/L1: extremely high since mostly L2
Connected L1+L2: unknown but presumably very high given the countries most involved
Countries with speakers: widespread, mainly Japan, China, France, Germany, Italy, Poland, United States, Brazil, Belgium, and United Kingdom (in order of number of members in the World Esperanto Association).
Virtual Presence Indicator: unknown but presumably high
Cyber-Globalization rank: unknown but presumably top given the huge L1+L2/L1 and the number of countries
Wikimedia: Strong
GoogleTranslate: Yes
Comments: This is a top candidate, it is surprising they have not done this move already. Would not be a surprise that the subject had been under discussion. A closed petition existed for a .io domain with low success (https://www.change.org/p/icann-add-a-eo-top-level-domain-for-the-esperanto-community). The funding has probably been the blocking factor for a previous move and may remain the main issue.
Pros: An Esperanto domain would be the most natural and logical move for a constructed language with potential for reaching out in every country and for which the Internet is a second chance to counterbalance a very slow growth. No ccTLD to care about.
Cons: None
Potential representation: Universala Esperanto Asocio - https://uea.org
Funding opportunities:
References:
https://www.bbc.com/future/article/20180110-the-invented-language-that-found-a-second-life-online
https://en.wikipedia.org/wiki/Esperanto
ccTLD: None

LANGUAGE FORM ISO639-3: rom (Romani) **xx

NAME (English, local): Romani, Romani + Rromani ćhib + Romanes + Kaalengo tšibbaha + Romnimus
Classification: Top candidate for language and culture (**xx)
The .rom domain would contribute to strengthen the existing move from a mainly oral language to written practices and help the convergence towards standards.
If macro language: Yes

rmc	Carpathian Romani
rmf	Kalo Finnish Romani
rml	Baltic Romani
rmn	Balkan Romanirmo, Sinte Romani
rmw	Welsh Romani
rmy	Vlax Romani

L1+L2: 1 868 970

L1+L2/L1: 1,12
Connected L1+L2: 85,95%
Countries with speakers: 36
Bulgaria 482 830, Romania 276 000, Serbia 147 600, Slovakia 123 800, Russian Federation 108 000, Germany 88 500, Turkey 72 900, Argentina 59 000, France 48 900, Ukraine 48 700, North Macedonia 41 200, Greece 41 000, Hungary 39 800, Iran 36 800, Austria 25 000, Italy 23 000, Poland 21 080, Switzerland 21 000, Czechia 20 120, Brazil 20 000, Belgium, Croatia, Sweden, Moldova, Belarus, Bosnia and Herzegovina, Finland, Kazakhstan, Latvia, Montenegro, Colombia, Slovenia, Albania, Lithuania, Netherlands, Estonia (<20 000).
Virtual Presence Indicator: 1.17
Cyber-Globalization rank: 24
Wikimedia: yes
GoogleTranslate: yes
Comments:

- Romani is the only Indo-Aryan language spoken almost exclusively in Europe.
- No monolingual web sites (English mostly together with Romani)

Pros:

- No geographical anchorage
- Wide spread amongst countries
- Strong centralized representation

Cons:

- Large differences between language's varieties (but homogeneous core)
- Function is primarily oral, with no monolingual speakers, no written standard, no prescriptive norms

Potential representation:

- International Romani Union - https://iru2020.org/
- European Roma Institute for language and culture https://eriac.org/
- Possible contact: https://yaronmatras.org a UK academic who created a romani project https://scholar.google.com/citations?user=grsZylMAAAAJ&hl

Funding opportunities:
References:

- https://www.researchgate.net/publication/276963791_The_use_of_Romani_language_in_the_Internet_and_the_Roma_identity
- https://www.academia.edu/44154640/PRE_PUBLICATION_COPY_Romani_on_the_Internet

ccTLD: None

LANGUAGE FORM IETF Language code: cpf (French creole) ***x @

regrouping ISO-689-3: acf, cks, crs, gcf, gcr, hat, icr, lou, mfe, rcf, scf

NAME (English, local): French Creole, kreyòl
Classification: Top candidate for language and culture (***x)
The . kreyòl domain would contribute to unite, under the dominant Haitian creole many speakers of the creole family dispersed in many diasporic and non-diasporic places: Bahamas, **Brazil**, British Indian Ocean Territory, **Canada**, **Chile**, Dominica, **Dominican Republic, France, French Guiana**, Grenada, **Guadeloupe, Madagascar, Martinique**, **Mauritius**, New Caledonia, Panama, **Réunion**, Saint Barthélemy, **Saint Lucia**, Saint Martin, Seychelles, South Africa, Trinidad and Tobago, Turks and Caicos Islands, United Kingdom,

United States, and representing 16.7 million speakers**. In bold when > 0.1 M speakers**. **Underscored when > 0.5M**

If macro language: No, but coherent group of languages

acf	Lesser Antillean French Creole
cks	Tayo (New Caledonia)
crs	Seychelles French Creole
gcf	Guadeloupean French Creole
gcr	Guianese French Creole
hat	Haitian Creole
icr	Karipuna French Creole (Brazil)
lou	Louisiana Creole
mfe	Morisyen
rcf	Réunion French Creole
scf	San Miguel French Creole (Panama)

L1+L2: 16 694 303
L1+L2/L1: Mostly L1
Connected L1+L2: 53.75%
Countries with speakers: 25 Haiti holds 67% of speakers
Virtual Presence Indicator: n.a.
Cyber-Globalization rank: n.a.
Wikimedia: yes (Haitian creole)
GoogleTranslate: yes (Haitian, Morisyen and Seychelles creoles)
Comments:

- It is challenging but a champion has been identified and OBDILCI could help.
- It is a special opportunity to see a highly motivated Haitian leadership arise approved and supported by the rest of partners' country.
- Could be the first promising experience of a TLD dedicated to a family of languages with strong commonality, opening the door for many similar initiatives which would help the difficult situation of many language families having for each component a low number of speakers but together representing a large number (apply to some of the indigenous languages).

Pros:

- Could receive a strong speakers' support and drive potential content fostering. Strong centralized representation
- This is a move which would gain strong political support in the concerned countries.

Cons: Complex representation and institutionalization
Potential representation: A champion has been identified willing to take the lead with high skill in leadership and coordination together with good networking. Patrick Attié, is an Haitian successful entrepreneur with leadership capacity and strong technical team as well as large network both in Haiti and the Caribbean. OBDILCI is having a large experience and contacts in the referenced area, both geographic and institutional, and could play a facilitator role and help extend the contacts to the other concerned regions.
Funding opportunities: Funding opportunities could come from Francophonie and from Indigenous language decennia branded by UNESCO.
References: https://www.linkedin.com/in/patrickattie/
https://obdilici.org
ccTLD: .ht, .mq, .gp, .re, .fr, .mu; .sc

LANGUAGE FORM Berber Subgroup zgh (Tamazight Standard Moroccan) **xx @@

NAME (English, local): Berber - Amazigh/Tamazight/Tashelhit Tašlḥiyt + تشلحيت (tšlḥyt) + ⵜⴰⵛⵍⵃⵉⵢⵜ (taclhiyt)
Classification: Top potential for language and culture
If macro language: A sub-group which encompass 28 languages: Awjilah auj (Libya), Sawknah swn (Libya), Siwi siz (Egypt), Chenoua cnu (Algeria), Judeo-Berber jbe (Israel), **Tachelhit shi** (Morocco), Tamazight, **Central Atlas tzm** (Morocco), **Tamazight, Standard Moroccan zgh** (Morocco), **Kabyle kab** (Algeria), Ghadamès gha (Libya), **Nafusi jbn** (Libya), Sened sds (Tunisia), Ghomara gho (Morocco), Tagargrent oua (Algeria),Tamazight, Temacine tjo (Algeria), Taznatit grr (Algeria), **Tumzabt mzb** (Algeria), Senhaja Berber sjs (Morocco), **Tarifit rif** (Morocco), **Tachawit shy** (Algeria), Tamazight, Tidikelt tia (Algeria),**Tamahaq, Tahaggart thv** (Algeria), Tamajaq, **Tawallammat ttq** (Niger),Tamajeq, Tayart thz (Niger), **Tamasheq taq** (Mali), Tetserret tez (Niger), Zenaga zen (Mauritania). Note that Guanche (gnc) is an extinct language of Spain (Canary) belonging to this subgroup.
Among them the ones with significant demography are marked in bold: kab (7.9), shi (5.8M), tzm (3.1), shy (2.6), rif (1.9), ttq (1.3), taq (0.9), thz (0.4), jbn (0.3), mzb (0.2), thv (0.13)
Note that zgh is a Statutory National Language of Morroco developed by the Royal Institute of Amazigh Culture (IRCAM) by combining features of the major Berber languages in Morocco and used as L2 by all, while their speakers are accounted by Ethnologue in each component.

The intent to unite under the same TLD all Berber languages, mainly Tamazight, Kabyle and Tuareg, from Algeria, Morocco, Libya, Niger Mali and more, is certainly wise but the pragmatics tell that the current geopolitical complexity is a major obstacle. The highly feasible counter-alternative could be to target the effort started by IRCAM to unify Berber languages inside Morocco and focus on Moroccan Berbers, in coordination with IRCAM. The remaining of the form follows that alternative.

L1+L2: All Berber close to 25M, Moroccan Berbers 10.8 total incl. 9.6 M in Morocco
L1+L2/L1: almost no L2
Connected L1+L2: 90%
Countries with speakers: 6 Morocco (9.6), France (0.95) Netherlands (0.17), Western Sahara (0.12), Algeria, Canada
Virtual Presence Indicator:
Cyber-Globalization rank:
Wikimedia: Yes, recent by IRCAM
GoogleTranslate: Yes (Tamazight)
Comments: Tifinagh alphabet
Pros: Natural competent representative will decide opportunity and if decide to go on has the capacity to organize, fund and overcome obstacles. Morocco is an African leader in Internet.
Cons:
Potential representation: Royal Institute of Amazigh Culture https://www.ircam.ma/
Potential funding: Moroccan government
References: https://www.nationalia.info/new/11589/awal-the-popular-project-that-wants-to-make-the-internet-speak-amazigh
https://arbitrer.fib.unand.ac.id/index.php/arbitrer/article/view/390
ccTLD: .ma

LANGUAGE FORM ISO639-3: yid (Yiddish) **x

NAME (English, local): Yiddish, ייִדיש
Classification: High potential for language and culture (**x)
If macro language: Yes, Eastern Yiddish [ydd] and Western Yiddish [yih]

L1+L2: 421 797 (11 million in 1939)
L1+L2/L1: Almost no L2
Connected L1+L2: 93.8%
Countries with speakers: USA (0.2), Israel (0.17), Canada (0.02), Ukraine (0.01), less than 7 000: Belarus, United Kingdom, Germany, Turkmenistan, Sweden, Russian Federation, Latvia, Romania, Moldova, Poland
Virtual Presence Indicator:
Cyber-Globalization rank:
Wikimedia: Fair
GoogleTranslate: Yes
Comments: The data on speakers around the world seems to be unknown. High potential for growth. High potential for literature to digitalize. Online experience. Yiddish daily newspaper (https://forward.com/yiddish/).
Pros:

- Very dispersed community around a (re) developing language.
- Important digital libraries (https://www.yiddishbookcenter.org/)
- If motivated capacity to rise funding

Cons: Not a large critical mass at this stage
Potential representation: YIVO, institution for the study of Eastern European Jewry, https://www.yivo.org/ Hold a project to revive online huge data base of Yiddish lost documents.
Potential funding: High if motivated
References:
ccTLD: None

LANGUAGE FORM ISO639-3: aym (Aymara macro) **x

NAME (English, local): Aymara, aymar aru
Classification: High potential candidate for language and culture (**x)
If macro language: Yes Central Aymara [ayr] (Bolivia), Southern Aymara [ayc] (Peru).
L1+L2: 1 677 100
L1+L2/L1: L1 only
Connected L1+L2: 74%
Countries with speakers: 4 Bolivia (1M), Peru (0.65), Chile (0.02), Argentina (0.004)
Virtual Presence Indicator: 0.87
Cyber-Globalization rank: 3
Wikimedia: High
GoogleTranslate: Yes
Comments: Ivan Guzman de Rojas, a Bolivian scientist, has created an early translation program using Aymara as pivot language. He claimed that the grammatical matrix structure of Aymara makes it a special choice and allow to process in parallel translations to various languages. We have tried to convince him to open its program and algorithm but he died recently leaving unknown his secrets (https://www.obdilci.org/blog/ivan-guzman-de-rojas-and-the-aymara-language/)
Pros: A vigorous or developing language with growing digital existence and strong associated culture. Maybe the right time to boost its digital growth.
Cons: The counterpart is not obvious. It would be wise but a heavy task to create a multi-stakeholder structure for such project involving civil society and academic organizations together with various governmental agencies which could be cumbersome, except if ADSIB which is the NIC for .bo get motivated.

Potential representation: Agencia para el Desarrollo de la Sociedad de la Información https://adsib.gob.bo/ and/or Instituto Plurinacional de Estudio de Lenguas y Culturas https://www.ipelc.gob.bo/ Contact: ipelc@ipelc.gob.bo or https://www.minculturas.gob.bo/
Potential funding: UNESCO, indirectly thru the Decade of Indigenous languages (2022-2032)
References: https://www.goethe.de/prj/zei/en/art/24438469.html
https://rising.globalvoices.org/blog/2020/11/18/apthapi-digital-project-creates-digital-security-resources-in-the-aymara-language/
ccTLD: .bo, .pe (.bo managed by ADSIB)

LANGUAGE FORM ISO639-3: swa (Swahili macro) *** @

NAME (English, local): Swahili, Kiswahili
Classification: High potential candidate for language (***)
If macro language: Yes, Congo Swahili [swc] (Democratic Republic of the Congo) Swahili [swh] (Tanzania)
L1+L2: 97 658 480
L1+L2/L1: 18.6
Connected L1+L2: 32.9%
Countries with speakers: 24 *Tanzania* (59.4M ST), *Kenya* (21.6 ST), *RD Congo* (11.1 ST), *Uganda* (4.3), Saudi Arabia (0.4), Somalia (0.3) + United States, Canada, Oman, Zambia, Sudan, Réunion, Mozambique, Rwanda, Australia, Burundi, United Kingdom, Mayotte, Madagascar, United Arab Emirates, Malawi, Finland, Libya, Comoros, New Zealand. ST= Statutory National, *italic = mainly L2*- Vehicular language in a large portion of East Africa
Virtual Presence Indicator: 0.37
Cyber-Globalization rank: 3
Wikimedia: Strong
GoogleTranslate: Yes
Comments: Its large vehicular action and its solid digital existence make it a high potential candidate.
Pros: Large presence in Wikimedia. Latin alphabet. Top CGI indicate bright future. Key representation.
Cons: The inter-governmental nature of the key representation may turn the process slow to start.
Potential representation: East African Kiswahili Commission (EAKC) - https://kiswahili.eac.int/ Contact: eakc-hq@eachq.org
Potential funding: Thru UNESCO or ACALAN (https://acalan-au.org/) indirectly.
References: https://library.columbia.edu/libraries/global/virtual-libraries/african_studies/languages/swahili.html
https://rising.globalvoices.org/blog/2020/06/15/making-swahili-visible-identity-language-and-the-internet/
ccTLD: .tz, .ke, .cd

LANGUAGE FORM ISO639-3: hau (Hausa) ***

NAME (English, local): Hausa, Hausa
Classification: High potential candidate for language (***)
If macro language:
L1+L2: 88 238 080
L1+L2/L1: 1.64 Used as lingua franca by speakers of a huge number of West African languages
Connected L1+L2: 31.5%

Countries with speakers: 18 Nigeria (63.4M), Niger (19.6), Côte d'Ivoire (1.6), Benin (1.2), Sudan (0.8), Ghana (0.6), Chad, Cameroon, Central African Republic, Equatorial Guinea, Togo, Gabon, Gambia, Algeria, Congo, Canada, United Kingdom, Burkina Faso
Virtual Presence Indicator: 0.29
Cyber-Globalization rank: medium
Wikimedia: Fair
GoogleTranslate: Yes
Comments: Its large lingua franca action and its fair digital existence make it a high potential candidate.
Pros: Fair presence in Wikimedia. Latin alphabet.
Cons: Not sure the skills exist in the academic representation. Maybe it would be better to deal with NITDA which was past Registry of .ng
Potential representation: Centre for the Study of Nigerian Languages https://cnl.buk.edu.ng/, National Information Technology Development Agency (NITDA) https://nitda.gov.ng/
Potential funding: Thru UNESCO or ACALAN (https://acalan-au.org/) indirectly.
References:
https://www.researchgate.net/publication/376256658_Hausa_in_the_21st_Century_Internet_Environment_From_Easy_Access_to_Documentation
https://www.academia.edu/36403046/Promoting_the_use_of_the_Hausa_language_on_the_internet
ccTLD: .ng

LANGUAGE FORM ISO639-3: kur (Kurdish) **x

NAME (English, local): Kurdish Macro, (Kurdîyi başûrî) باشوور کوردیی + (Kurdî xwarg) خوارگ کوردی + zimanê soranî + Kurdî-Kurmancî + (Kurmancî) سۆران زمانێ
Classification: top potential candidate for language and culture downgraded due to complex context (**x)
If macro language: Yes Central Kurdish [ckb] (Iraq), Northern Kurdish [kmr] (Turkey), Southern Kurdish [sdh] (Iran) – Two alphabets, one Latin based, one Arabic based.
L1+L2: 26 088 540
L1+L2/L1: 1 Almost no L2
Connected L1+L2: 79%
Countries with speakers: 33 Turkey (8.9M), Iraq (8.8) **SNL**, Iran (5.5), Syria (1.9), Germany (0.2), France (0.08), Bahrain, Netherlands, United Kingdom, Turkmenistan, Kazakhstan, Russian Federation, Armenia, Canada, Lebanon, Greece, Belgium, Finland, Georgia, Kyrgyzstan, Italy, United States, Australia, Switzerland, Jordan, Denmark, Sweden, Azerbaijan, Norway, Tajikistan, Spain, Ukraine
Virtual Presence Indicator: 1
Cyber-Globalization rank: medium
Wikimedia: Fair
GoogleTranslate: Yes (both variants Kurmanji and Sorani)
Comments: A language of people united by a powerful nationalist movement and with very important and highly spread diaspora. Solid presence in the Internet (including Search Engine https://www.egerin.com/). Present all parameters for top candidate, however such an initiative is prone to provoke strong resistance from the 4 countries with more speakers and possible conflicts. Representation would need to be **diaspora centered** given the geopolitical context. Note that .krd exists as the geographic TLD for Kurdistan Region of Iraq used by Kurdistan Regional Government but apparently without vocation to extend further of that region (only 1881 domains registered so far).

Pros: All parameters for top candidate.
Cons: Potential conflicts. Negotiation for deal with concerned ccTLDs seems hardly feasible. Competition with .krd?
Potential representation: The Kurdish Institute of Paris https://www.institutkurde.org/ (non-political) and/or The Kurdish Academy of Language https://kurdishacademy.org/ (San Francisco, USA).
Potential funding: There is a Kurdistan Regional Government with a digital agenda: https://gov.krd/dxs/ which could either reject or support such a project.
References:
ccTLD: .tr, .ir, .iq,.sy

LANGUAGE FORM ISO639-2: smi Sami **x

NAME (English, local): Sami family Sami, Saami, Samic
Classification: High potential for language and culture (**x)
If macro language: No, but it is a family of Uralic languages, including some instinct or almost instinct ones, and only one with substantial speakers in Norway, North Sámi (sme): sma (Southern), sju (Ume), sje (Pite), smj (Lule), sme (Northern), sjk (Kemi), smn (Inari), sms (Skolt), sia (Akkala), sjd (Kildin), sjt (Ter).
L1+L2: 30 000
L1+L2/L1: No L2
Connected L1+L2:
Countries with speakers: Sami family is spread between Norway, Sweden, Finland and Russia, in their respective northern parts.
Virtual Presence Indicator:
Cyber-Globalization rank:
Wikimedia: Yes (sme and sms)
GoogleTranslate: No
Comments: Small speaker base but many strong arguments to try contribute to strengthen unification thru Internet.
Pros:
Cons: Requires a complex multi-stakeholder multi-countries approach
Potential representation: Norwegian Sami Parliament https://sametinget.no/. The same structure exists in Sweden and Finland, and such initiatives would require their cooperation together with the NGO https://www.saamicouncil.net/ with members in all the concerned countries.
Potential funding: Funding opportunities could come from the governments of the mentioned countries and from Indigenous language decennia branded by UNESCO.
References: https://finland.fi/life-society/sami-language-in-the-digital-age/,
https://journal.oraltradition.org/wp-content/uploads/files/articles/28i/07_28.1.pdf
https://www.iiisci.org/journal/pdv/sci/pdfs/PA003RU17.pdf
ccTLD: .no, .se, .fi, .ru

LANGUAGE FORM ISO639-3: Tagalog (tgl) ***

NAME (English, local): Tagalog Tagalog
Classification:
If macro language:
L1+L2: 83 357 610
L1+L2/L1: 2.85
Connected L1+L2: 74%

Countries with speakers: 46 Philippines (76.5M, 2/3 L2), USA (1.8), Saudi Arabia (0.9), Canada (0.7), Japan (0.5), Malaysia (0.5), United Arab Emirates (0.4, Italy, Qatar, Indonesia, Bahrain, United Kingdom, Oman, New Zealand, Spain, Guam, Germany, Brazil, South Korea, Norway, Netherlands, France, Nigeria, Israel, China–Macao, Switzerland, Northern Mariana Island, Lebanon, Libya, Cyprus, Brunei, Greece, Jordan, Denmark, Ireland, Austria, Finland, Sweden, Belgium, Egypt, Cayman Islands, Palau, American Samoa, Micronesia.
Virtual Presence Indicator: 0.87
Cyber-Globalization rank: Top
Wikimedia: Top
GoogleTranslate: Yes
Comments: De facto national language of Philippines, used as L2 by many and widespread internationally. The fact it is not official language and its huge spread may justify a TLD, providing agreement from both Philippines government and nic authorities. The fact that the official language Filipino (fil) is largely based on Tagalog and have less users than Tagalog, and that Philippines holds 186 languages, 175 of which are indigenous could orient the decision towards the direction of a TLD reflecting the large linguistic diversity of this island country. Those are matters of discussion with authorities. Note that among those languages some have large speaker's basis and could receive specific attention: Cebuano (ceb) 15.9M, Ilocano (ilo) 6.4, Hiligaynon (hil) 6.2, Bikol Macro (bik) 3.8, Waray- waray (war)2.6 , Kapampangan (pam) 2, Pangasinan pag 1.2, Maguindanaon (mdh) 1, Maranao (mrw) 0.9, Tausug(tsg) 0.8, Masbatenyo (msb) 0.7, Surigaonon (sgd) 0.5, Aklanon (akl) 0.5, Chavacano (cvk) 0.43, leaving aside various tenth of language between 20K and 500K.
Pros: Top language in Wikimedia, extended diaspora
Cons: The case of Philippines is very special and may require a specific more in-depth study and definitively any decision must be discussed with corresponding authorities. In any case, Cebuano, which is the second language in terms of articles in Wikipedia and Ilocano could also be candidate if the focus remains on Tagalog.
Potential representation: Official language commission of the Philippines https://kwf.gov.ph/; for diaspora, USA being in strong first demographic position, the National Federation of Filipino American Associations (NaFFAA)- https://naffaa.org/
Potential funding:
References: https://www.globalizationpartners.com/2022/06/15/website-translation-into-tagalog-5-things-to-consider/
ccTLD: .ph

LANGUAGE FORM: Mayan family (**x)

NAME (English, local): Mayan family
Classification: High potential for language and culture (**x)
If macro language: No, but sub-group of 31 languages, some principally from Mexico, other sfrom Guatemala with Belize also concerned.
L1+L2: More than 6M: 2.5 in Mexico, 3.7 in Guatemala + 0.03 in Belize
Most important are Q'eqchi' (kek) 1.1M, K'iche' (quc) 1, Mam (mam) 0.6, Kaqchikel (cak) 0.4, in Guatemala and Maya, Yucatec (yua) 0.8, Tzeltal (tzh) 0.6, Tzotzil (tzo) 0.6, in Mexico
L1+L2/L1: No L2
Connected L1+L2: 81% for Mexico, 54% for Guatemala
Countries with speakers: 3
Virtual Presence Indicator:
Cyber-Globalization rank:
Wikimedia: No
GoogleTranslate: Yes for Yua

Comments: The focus on family of languages instead of a single language is appropriate for this type of situation with strong cultural unity and strong linguistic diversity. However, the situation of non-inter-comprehension between the languages and the requirement for close cooperation between the 2 countries may represent a real challenge. If the cooperation is not obtained remains the possibility to separate into 2 TLD, one centered in Yucatan region of Mexico and the other in Guatemala. More specific studies on connectivity situation may be required, especially for Guatemala.

Pros: Maya culture is strong and span the languages and countries.

Cons: Bureaucracy of the representation could be a challenge.

Potential representation: in Mexico**:** Instituto Nacional de Lenguas Indígenas https://www.inali.gob.mx/; in Guatemala, Academia de Lenguas Mayas de Guatemala https://www.almg.org.gt/, for civil society, https://www.ukuxbe.org/

Potential funding: Funding opportunities could come from concerned governments and from Indigenous language decennia branded by UNESCO.

References: https://rising.globalvoices.org/blog/2022/08/14/we-promote-our-mayan-languages-online-so-they-will-not-be-forgotten/

ccTLD: .mx, .gt, .bz

NOTE: Would discussion with INALI be initiated about the possibility of TLD for Mexican indigenous languages, they should be extended to the cases of another possible candidate: Otomanguean, family of 176 languages of Mexico including Zapotec and Mixtec, with 1.8M speakers.

LANGUAGE FORM ISO-693-2: Nahuatl family (nah) **x

NAME (English, local): Nahuatl Nawatlahtolli, Mexikatlahtolli, Mexkatl, Mexikanoh, Masewaltlahtol

Classification: High potential for language and culture (**x)

If macro language: There is 28 varieties of languages under the Nahuatl name, all in Mexico. The fact Nahuatl have not been defined as a macro-language prevents it to appear in studies for languages with more than 1M L1 speakers.in spite the fact they gather 1.65 M speakers mostly concentrated in Tlaxcala and Puebla states of Mexico. Nahuatl, Eastern Huasteca (nhe) and Nahuatl, Western Huasteca (nhw) are the ones with most speakers (0.4 each in Mexico and 0.08 in USA for nhe). It is among the most studied and best-documented indigenous language of Americas.

L1+L2: 1 650 000

L1+L2/L1: No L2

Connected L1+L2: Mexico connection rate is 81%

Countries with speakers: 2

Virtual Presence Indicator:

Cyber-Globalization rank:

Wikimedia: Yes

GoogleTranslate: No

Comments: Discussions with INALI would determine if it is better to separate Nahuatl from its group, the Southern Uto-Aztecan which gather a total of a total of 63 languages (including the 28 for Nahuatl) or treat the whole group under the same TLD. Note that the Northern Uto-Aztecan group includes 14 languages from USA including Comanche and Hopi, none crossing 7 000 speakers and many almost instinct.

Pros:

Cons:

Potential representation: Instituto Nacional de Lenguas Indígenas https://www.inali.gob.mx/
Potential funding: Funding opportunities could come from concerned governments and from Indigenous language decennia branded by UNESCO.
References:
ccTLD: .mx

NOTE: Would discussion with INALI be initiated about the possibility of TLD for Mexican indigenous languages, they should be extended to the cases of another possible candidate: Otomanguean, family of 176 languages of Mexico including Zapotec and Mixtec, with 1.8M speakers.

LANGUAGE FORM ISO639-3: uig Uyghur **x

NAME (English, local): Uyghur (Uyghur tili) ئۇيغۇرچە + ئۇيغۇر تىلى (Uyghurche)
Classification: High potential candidate for language and culture (**x)
If macro language:
L1+L2: 10 548 782
L1+L2/L1: No L2
Connected L1+L2: 78%
Countries with speakers: 11 China (10M), India, Kazakhstan (.29), Uzbekistan (.05), Turkey (.04), Kyrgyzstan, Mongolia, Pakistan, Russian Federation, Saudi Arabia, Turkmenistan, United States
Virtual Presence Indicator: 1.12
Cyber-Globalization rank: medium
Wikimedia: Fair
GoogleTranslate: Yes
Comments: If it was not an extremely sensitive political situation with China, it could be a top potential candidate given its cultural strength, digital presence and strong diaspora. A cultural approach diaspora centered could still be a possibility and, in any case, a required one as censorship have been applied to Chinese websites in Uyghur.
Pros: significant speaker base, cultural identity, and important diaspora
Cons: China is prone to oppose officially or not.
Potential representation: https://www.uyghurcongress.org/
Potential funding:
References: https://www.wired.com/story/uyghur-internet-erased-china/
ccTLD: n.a.

LANGUAGE FORM ISO639-3: rhg Rohingya **x

NAME (English, local): Rohingya Ruwainggya
Classification: High potential candidate for language and culture (**x)
If macro language:
L1+L2: 2 529 270
L1+L2/L1: No L2
Connected L1+L2: 58%
Countries with speakers: 10 Bangladesh (.95), Saudi Arabia (.5), Myanmar (.48), Pakistan (.35), Malaysia (.15), United Arab Emirates, India, Thailand, Australia, Indonesia, Turkmenistan, United States
Virtual Presence Indicator: 0.59
Cyber-Globalization rank: medium

Wikimedia: No
GoogleTranslate: No
Comments: More speakers abroad than in the country of the language (Myanmar) due to huge refugees' situation. A TLD could serve as a digital hub for the Rohingya diaspora, strengthening their cultural identity, however will probably be opposed by Myanmar.
Pros: Could be part of (a diaspora centered) solution for a largely stateless population
Cons: Extremely sensitive political context
Potential representation: www.rohingyaproject.com
Potential funding: https://www.unhcr.org/
References: https://nethope.org/programs/connectivity-and-infrastructure/data-connectivity-in-the-rohingya-refugee-camps/
https://www.nature.com/articles/s41599-023-01553-w
https://rohingya-voice.com/internet/
ccTLD: n.a.

LANGUAGE FORM ISO639-3: nan (Chinese Min Nan) ***

NAME (English, local): Chinese Min Nan - 闽南语 (Minnanyu) (***)
Classification: High potential for language
If macro language: Part of macro-language zho.
L1+L2: 50.6M
L1+L2/L1: Almost no L2
Connected L1+L2: high
Countries with speakers: 10 China (28M), Taiwan (13.5), Malaysia (3.5), Thailand (1.5), Philippines (1.3), Indonesia (1), Singapore (.6), Hong Kong (.4), United States (.15), Cambodia (.09), Japan (.07), Myanmar (.07), Australia, Brunei, Canada, France, New Zealand
Virtual Presence Indicator: high
Cyber-Globalization rank:
Wikimedia: Yes
GoogleTranslate: No
Comments: Also known as Hokkien, one of the dialects. There is a current campaign in Taiwan to strengthen Hokkien: https://www.speakhokkien.org
Pros: It has all the characteristics for a top candidate for languages and cultures and could be embraced by Taiwan authorities.
Cons: Such a project would probably receive strong opposition from China.
Potential representation: Taiwanese authorities have linguistic policies both for Hakka and Hokkien and would be natural manager of such project.
Potential funding: Taiwan government
References: https://taiwaninsight.org/2022/08/24/the-many-faces-of-the-hokkien-language-internet/
ccTLD: .tw

LANGUAGE FORM ISO639-3: hak (Chinese Hakka) **x

NAME (English, local): Chinese Hakka 客家話 (Hakkafa) (**x)
Classification: High potential for language and culture
If macro language: Part of macro-language zho.
L1+L2: 44M
L1+L2/L1: Almost no L2

Connected L1+L2: High
Countries with speakers: 21 China (36.4M), China–Taiwan (4.24), Malaysia (1.8), Indonesia (.64), Hong Kong (.26), Singapore (.23), Thailand (.08), Brunei, Canada, French Guiana, Jamaica, Cambodia, Myanmar, Mauritius, New Zealand, Panama, French Polynesia, Réunion, Suriname, United States, Vietnam
Virtual Presence Indicator: High
Cyber-Globalization rank:
Wikimedia: Yes
GoogleTranslate: Yes
Pros: It has all the characteristics for a top candidate for languages and cultures and could be embraced by Taiwan authorities.
Cons: Such a project would probably receive strong opposition from China.
Potential representation: Taiwanese authorities have linguistic policies both for Hakka and Hokkien and would be natural manager of such project.
Potential funding: Taiwan government
References: https://jati.um.edu.my/index.php/jati/article/view/5913/3629
ccTLD: .tw

LANGUAGE FORM ISO639-3: afr (Afrikaans) **x

NAME (English, local): Afrikaans
Classification: High potential candidates for Language and Culture (**x)
If macro language:
L1+L2: 18 093 000
L1+L2/L1: 2.33
Connected L1+L2: 74%
Countries with speakers: 15 South Africa (17M), Namibia (.13), Zambia (.1), Zimbabwe (.09), United States (.05), Australia (.05), New Zealand, Canada, Netherlands, Eswatini, United Kingdom, Botswana, Malawi, Lesotho, Angola,
Virtual Presence Indicator: 0.83
Cyber-Globalization rank: Fair
Wikimedia: High
GoogleTranslate: Yes
Comments:
Pros: Strong diaspora, strong in Internet, universities create contents, strong in media (Afrikaanse Taal- en Kultuurvereniging (ATKV) - https://atkv.org.za/
Cons:
Potential representation: Pan South African Language Board https://www.pansalb.org/ Afrikaans Language Council https://www.afrikaansetaalraad.co.za/
Potential funding:
References: https://mybroadband.co.za/news/internet/73624-internet-in-sa-english-vs-afrikaans-vs-african-languages.html
ccTLD: .za

Note: Other potential *x or * candidates from South Africa could be discussed with PANSALB and .za.
Ndebele (nbl), a regional language with 1.4M L1 and 1M L2 all in South Africa, not in Wikipedia but in GT.
Xhosa (xho), a regional language with 19M, more than half L2, also in Lesotho, Botswana and Zimbabwe with few speakers, present in Wikimedia and GT.

Zulu (zul), a regional language with 27M, 60% L2, also in Lesotho (.3), Eswatini, Malawi, Mozambique, Botswana. present in Wikimedia and GT.
The following are the rest of South African languages with more than 1M speakers and, for most, regional with high L2 figures: **Sepedi (nso)**, **Sesotho (sot), Setswana (tsn), Siswati (ssw), Tshivenda (ven), Xitsonga (<tso).**

LANGUAGE FORM ISO639-3: ful (Fulfulde) ***

NAME (English, local): Fulfulde, Fulfulde + فُلْفُلْبِ (Fulfulde) + Maasinankoore + Pulaar + Pular
Classification: High potential for language (***)
If macro language: Yes Adamawa Fulfulde [fub] (Cameroon), Bagirmi Fulfulde [fui] (Chad), Borgu Fulfulde [fue] (Benin), Central-Eastern Niger Fulfulde [fuq] (Niger), Maasina Fulfulde [ffm] (Mali), Nigerian Fulfulde [fuv] (Nigeria), Pulaar [fuc], Pular [fuf] (Guinea), Western Niger Fulfulde [fuh] (Niger).
L1+L2: 39 955 240
L1+L2/L1: 1.07 (All L2 in Cameroon)
Connected L1+L2: 37%
Countries with speakers: 19 Nigeria (16.5), Cameroon (5.3), Senegal (4.7), Guinea (4.3), Mali (2.1), Burkina Faso (1.8), Benin (.7), Guinea-Bissau (.7), Côte d'Ivoire (.5), Gambia (.4), Mauritania, Chad, Central African Rep., Sierra Leone, Sudan, Togo, USA, Ghana, South Sudan
Virtual Presence Indicator: 0.33
Cyber-Globalization rank: medium
Wikimedia: Yes
GoogleTranslate: Yes
Comments: Large populations spanning in various African countries
Pros: High degree of mutual intelligibility, seems to be a top African candidate
Cons: Low virtual presence but enough to take it as an argument for boosting it
Coordination and representation of such a multi-country project is a challenge.
Potential representation: Will need to create a consortium. If this was not the weak point this language would have reached the top-level category… Maybe the right strategy is to motivate Nigerian National Information Technology Development Agency https://nitda.gov.ng/ to take the lead and create a coordination towards such goal.
Potential funding:
References: https://library.columbia.edu/libraries/global/virtual-libraries/african_studies/languages/fula.html
ccTLD: .ng, .cm, .sg, .gn, .ml, .mf

LANGUAGE FORM ISO639-3: pap (Papamientu) **x @@

NAME (English, local): Papamientu
Classification: High potential for language and culture (**x)
If macro language: No. It is a creole Iberian based
L1+L2: 318 100
L1+L2/L1: 20K L2 in Curacao
Connected L1+L2: High
Countries with speakers: 5 Curacao (.14), Netherlands (.08), Aruba (.08), Sint Maarten, Caribbean Netherlands
Virtual Presence Indicator: High
Cyber-Globalization rank:
Wikimedia: Yes

GoogleTranslate: Yes
Comments: Quite lively with strong presence in media and the Internet. Curacao is a world model for multilingualism and TLD could be well received by NTI, the institution owning such decision.
Pros: Decision belongs to NTI
Cons:
Potential representation: National Language Institute of Curaçao (NTI) https://nticuracao.org/
Potential funding: Curacao authorities
References:
https://www.researchgate.net/publication/332091969_Towards_a_language_database_of_Papiamentu
ccTLD: .cw, .aw

LANGUAGE FORM ISO639-3: bar (Bavarian) **

NAME (English, local): Bavarian, Boarisch
Classification: Medium potential candidate for language (**)
If macro language: No
L1+L2: 14 667 000
L1+L2/L1: 1
Connected L1+L2: 93,97%
Countries with speakers: 5 Austria (8.3M), Germany (6), Italy (0.3M), Netherland (0.02), Czechia (0.009) + Switzerland (0.05 not accounted)
Virtual Presence Indicator: 1.4
Cyber-Globalization rank: 110
Wikimedia: yes
GoogleTranslate: no
Comments:

- Upper German varieties spoken in the German state of Bavaria, most of Austria and the Italian region of South Tyrol.
- In recent developments, there has been a movement advocating for the recognition of Bavarian as a regional language, similar to Cornish, Welsh, or Catalan.

Pros:
Large speakers' population highly connected in Europe which could extend to Hungary, Brazil, Peru, USA and Canada (no data).
Cons:
The absence of a centralized institutional representation poses challenges to its formal recognition and standardized support.
Potential representation: There is a possible champion to be consulted to check further possibilities: Prof. Anthony Rowley
https://de-m-wikipedia-org.translate.goog/wiki/Anthony_Rowley?_x_tr_sl=auto&_x_tr_tl=en
Bayerische Akademie der Wissenschaften, Alfons-Goppel-Str. 11, 80539 München, Zimmer 228, Telefon: +49 (0)89 23031-1180, bwb@kmf.badw.de
Funding opportunities: Check Bavarian Academy of Science above
Reference:
https://www.thetimes.com/world/europe/article/is-bavarian-a-language-or-dialect-sounds-like-a-job-for-a-yorkshireman-djt9xdnbx
ccTLD: .de, .at

LANGUAGE FORM ISO639-3: nds (Saxon, Low) *x @

NAME (English, local): Saxon, Low, Nedderdüütsch, Plattdüütsch in Germany, Pomerano in Brazil – Also called Low German.
Classification: Medium otential candidate for language and culture (*x)
If macro language: no
L1+L2: 2.5M
L1+L2/L1: Essentially L2 in Germany and L1 in Brasil.
Connected L1+L2: 91.5%
Countries with speakers: Germany (2.2M) and Brasil (0.3). Also spoken in Netherlands and Denmark (no data)
Virtual Presence Indicator: High
Cyber-Globalization rank: Low
Wikimedia: Yes
GoogleTranslate: No
Comments: Officially recognized as a regional (separate) language in 8 states of Germany. Recognized as a regional (separate) language by the European Charter on Languages. Adults only. Shifting to Standard German [deu]. Used as L2 by Northern Frisia. Statutory language of provincial identity in Brazil. Printed fairly widely outside Europe, particularly in North and Latin America, Australia, Southern Africa, and Eastern Europe (Siberia, Kazakhstan). Regional broadcasters, particularly Norddeutscher Rundfunk (NDR), feature content in their programming, including radio shows, television segments, and online platforms, contributing to the language's visibility and accessibility.
Pros:

- Could help the decline on youngsters in spite strong cultural and education presence.
- Strong educative and cultural material for content creation

Cons:
Potential representation: Institute for Low German Language https://ins-bremen.de/
Potential funding: German Federal Government Commissioner for Culture and the Media (BKM). https://www.kulturstaatsministerin.de/DE/startseite/startseite_node.html
References:
ccTLD: .de, .br

SPECIAL NOTE: There is no macro-language nor language family for nds, however variants of this language exist as L1 languages and most are vigorous or developing. A TLD shared by all those variants could be a motivating, although more complex approach, joining a total of more than 4 million speakers. See below the table of variants.

ISO_639	Language_Name	Country_Name	L1_Users	L2_Users	STATUS
act	Achterhoeks	Netherlands	211 000		Developing
drt	Drents	Netherlands	255 000		Developing
frs	Saxon, East Frisian Low	Germany	200 000		In trouble
gos	Gronings	Netherlands	262 000		Developing
nds	*Pomeranian*	*Brazil*	*300 000*		*Institutional*
nds	*Saxon, Low*	*Germany*	*1 000*	*2 200 000*	*In trouble*
sdz	Sallands	Netherlands	347 000		Vigorous
stl	Stellingwerfs	Netherlands	5 000		Developing
twd	Twents	Netherlands	334 000		Developing
vel	Veluws	Netherlands	175 000		Vigorous
wep	Westphalien	Germany			Vigorous

LANGUAGE FORM ISO639-3: que (Quechua macro) *x

NAME (English, local): Quechua, Kechua + Runa Simi + Chanka runasimi + Kichwa + Quechua + Runa Shimi + … (several more not cited)
Classification: Medium Potential candidate for language and culture (*x)
If macro language: Yes, 42 languages grouped. Ambo-Pasco Quechua [qva], Arequipa-La Unión Quechua [qxu], Ayacucho Quechua [quy], Cajamarca Quechua [qvc],,Cajatambo North Lima Quechua [qvl], Calderón Highland Quichua [qud] (Ecuador), Cañar Highland Quichua [qxr] (Ecuador), Chachapoyas Quechua [quk], Chaupihuaranga Quechua [qur], Chimborazo Highland Quichua [qug] (Ecuador), Chincha Quechua [qxc], Chiquián Quechua [qxa], Corongo Ancash Quechua [qwa], Cusco Quechua [quz], Eastern Apurímac Quechua [qve], Huallaga Quechua [qub], Huamalíes-Dos de Mayo Quechua [qvh], Huaylas Ancash Quechua [qwh], Huaylla Wanca Quechua [qvw], Imbabura Highland Quichua [qvi] (Ecuador), Jauja Wanca Quechua [qxw], Lambayeque Quechua [quf], Loja Highland Quichua [qvj] (Ecuador), Margos-Yarowilca-Lauricocha Quechua [qvm], Napo Quichua [qvo], North Bolivian Quechua [qul] (Bolivia), North Junín Quechua [qvn], Northern Conchucos Quechua [qxn], Northern Pastaza Quichua [qvz] (Ecuador), Pacaraos Quechua [qvp], Panao Quechua [qxh]Puno Quechua [qxp] Salasaca Highland Quichua [qxl] (Ecuador), San Martín Quechua [qvs], Santa Ana de Tusi Pasco Quechua [qxt], Santiago del Estero Quichua [qus] (Argentina), Sihuas Ancash Quechua [qws] South Bolivian Quechua [quh] (Bolivia), Southern Conchucos Quechua [qxo], Southern Pastaza Quechua [qup], Tena Lowland Quichua [quw] (Ecuador), Yauyos Quechua [qux].
L1+L2: 7 252 540
L1+L2/L1: 1
Connected L1+L2: 74%
Countries with speakers: 5 Peru (3 980 920), Bolivia (1 726 000), Ecuador (1 479 500), Argentina (65 120), Chile (1 000)
Virtual Presence Indicator: 0.64
Cyber-Globalization rank: 119
Wikimedia: Fair
GoogleTranslate: Yes
Comments:
Pros: There is a basis for development as witnessed by Wikimedia presence and activities from linguists interested in digital apps.
Cons: The representation issue is not promising at first glance.
Potential representation: Academia Mayor de la Lengua Quechua - https://amlq.org.pe/. In their presentation they mention: "La academia trabaja en la incorporación del quechua en plataformas digitales y tecnológicas, adaptándose a los tiempos modernos y facilitando su acceso a públicos más amplios.". Contact: https://amlq.org.pe/contacto/ o admin@amlq.org.pe. The Academy website does not provoke the feeling of a busy organization nor that they have branches in Bolivia and Ecuador. The alternative would be to deal with Ministry of Culture of Peru who have a Department on Indigenous languages or to find "champions" in the Wikimedia space which is fairly active.
Potential funding:
References: https://elcomercio.pe/eldominical/quechua-conquista-internet-ecpm-noticia-673227-noticia/
https://globalvoices.org/2011/09/09/peru-the-state-of-quechua-on-the-internet/
https://www.academia.edu/42556724/Quechua_in_the_technology_spreading_the_voices_by_Internet
ccTLD: .pe, .bo, .ec

LANGUAGE FORM ISO639-3: oci (Occitan) *x

NAME (English, local): Occitan, occitan + lenga d'òc + provençal / provençau
Classification: Medium Potential for language and culture (*x)
If macro language: no
L1+L2: 1 111 560
L1+L2/L1: mostly L1 (except Aranés in Spain)
Connected L1+L2: 87%
Countries with speakers: France (1M), Italy (0.1M), Monaco (4 500), Spain (7 000)*
Virtual Presence Indicator: high
Cyber-Globalization rank: low
Wikimedia: Fair
GoogleTranslate: Yes
Comments: Occitan is fundamentally defined by its dialects, rather than being a unitary language, as it lacks an official written standard. The dialects include Gascon/Béarnese/Aranese, Languedocien, Limousin, Auvergnat, Provençal/Niçard, Vivaro-Alpine. In a 2014 study of local languages of France, realized by the author of the report, Occitan was classified, the same as Breton and Corso, with strong presence and dynamic in the Internet with strong citizen participation rather than local government. This may have evolved and today there is a large number of potential representations which could turn into a problem. It is surprising Occitan has not followed the path of other TLDs from France. Maybe the plurality of dialects and of institutional representations is the explanation. In previous studies
Pros: There is a vibrant associative life and dense Occitan documentation.
Cons: The potential representations do not seem particularly interested or competent specifically in Internet matters and searches of cross interests between Occitan and Internet do not get much results beyond old 2014 study referenced.
Potential representation: https://ieo-oc.org/ or https://www.locongres.org/
Potential funding: https://www.culture.gouv.fr/nous-connaitre/organisation-du-ministere/La-delegation-generale-a-la-langue-francaise-et-aux-langues-de-France
References: https://www.culture.gouv.fr/content/download/106582/file/lr_2014_11_lang-france-sur-internet.pdf?inLanguage=fre-FR&version=2
https://baseldf.fr/urls/index/languesDInterface:/langue_id%5B0%5D:15/note_id:/attribut_id:/initiative_id:/ressource_id:/reload_ok:1/langue_id[0]:15/page:1
https://shs.cairn.info/revue-hermes-la-revue-2016-2-page-101?lang=fr
ccTLD: .fr, .it, .mc, .es

LANGUAGE FORM ISO639-3: grn (Guarani) *x @

NAME (English, local): Guarani + Guaraní + Ava Guaraní + Nhandeayvu* + Avañe'ẽ
Classification: High potential for language and culture but complex situation (*x)
If macro language: Yes Ava Guaraní [nhd], Eastern Bolivian Guaraní [gui] (Bolivia), Mbyá Guaraní [gun] (Brazil), Paraguayan Guaraní [gug], Western Bolivian Guaraní [gnw] (Bolivia).
L1+L2: 6 652 790
L1+L2/L1: no L2
Connected L1+L2: 78%
Countries with speakers: 4 Paraguay (6.3M), Argentina (0.22), Bolivia (0.06), Brazil (0.006)
Virtual Presence Indicator: 0.96
Cyber-Globalization rank: low
Wikimedia: Yes
GoogleTranslate: Yes

Comments: It is a national language for Paraguay which could make it a serious issue. Furthermore, it belongs to the Tupian family which gathers 76 languages in Brazil for a total population of 76 000 speakers. While some of the Tupian family in Brazil are close enough for mutual inter-comprehension with Guarani, this is not the case for others of the family. Given the fact that Guarani is a national language of Paraguay, together with Spanish, and the importance given to language public policies for Guarani, this matter should totally rely on the Paraguayan government, with SPL as the focal point. They should decide of the opportunity of a TLD for Guarani and if it is wise to **extend it to all Tupian family** languages, in coordination with Brazilian counterpart if they decide so.
Pros: Guarani is a top candidate.
Cons: It is a very unique situation and decision shall belong to Paraguay authorities.
Potential representation: Secretaría de Políticas Lingüísticas. spl.gov.py
Potential funding: Paraguay government and https://oei.int
References: https://www.sapiens.org/language/guarani-digital/
https://www.uticvirtual.edu.py/revista.ojs/index.php/revistas/article/view/98/208
https://www.uticvirtual.edu.py/revista.ojs/index.php/revistas/article/view/98
ccTLD: .py Note: it is managed by two universities and this could add complexity to the situation.

LANGUAGE FORM ISO639-3: iba (Iban) *x @

NAME (English, local): Iban Jaku Iban
Classification: Medium Potential candidate for language and culture (*x)
If macro language:
L1+L2: 1 481 800
L1+L2/L1: 1.9
Connected L1+L2: 97%
Countries with speakers: 3 Malaysia (1.45), Brunei (15 000), Indonesia (14 000)
Virtual Presence Indicator: 1.08
Cyber-Globalization rank: medium
Wikimedia: No
GoogleTranslate: No
Comments: Could be .arawak if regional approach.
Pros: One of the few Oceanic languages reaching one million speakers.
Cons: No evidences have been found of digital presence. This need to be checked more deeply.
Potential representation: https://nativecustoms.sarawak.gov.my
Potential funding:
References:
https://www.researchgate.net/publication/375907928_IBAN_Language_in_National_Education_Issues_and_Challenges
ccTLD: .my

LANGUAGE FORM ISO639-3: yor (Yoruba) *x

NAME (English, local): Yoruba Èdè Yorùbá
Classification: Medium potential for language and culture (*x)
If macro language:
L1+L2: 47 195 900
L1+L2/L1: 1.04

Connected L1+L2: 36%
Countries with speakers: 18 Nigeria (46M), Ghana (.46), Benin (.24), United States (.21), Côte d'Ivoire (.13), Togo (.12), Burkina Faso, Niger, Italy, Canada, Liberia, Gambia, United Kingdom, Sierra Leone, Greece, Australia, Ireland, Finland
Virtual Presence Indicator: 0.35
Cyber-Globalization rank: medium
Wikimedia: Fair
GoogleTranslate: Yes
Comments: Regional spamming countries plus diaspora
Pros: Good internet presence
Cons: No appropriate representation fund, better deal with NITDA which also manage .ng
Potential representation: https://yorubaacademy.com/ - National Information Technology Development Agency https://nitda.gov.ng/
Potential funding:
References:
https://www.researchgate.net/publication/374269612_Language_Visibility_and_Audibility_Discussing_the_Dominant_Status_of_Yoruba_on_Social_Media
ccTLD: .ng

LANGUAGE FORM ISO639-3: lin (Lingala) ** @

NAME (English, local): Classification: Medium potential for language (**)
If macro language: no but use as L2 by speakers of more than 100 African languages
L1+L2: 40.5 M
L1+L2/L1: L2=L1
Connected L1+L2: 27%
Countries with speakers: 11 Democratic Republic of the Congo (40M), Congo (0.2), Angola (0.16), Belgium, Burundi, Canada, Central African Republic, United Kingdom, Rwanda, Uganda, United States
Virtual Presence Indicator: 0.30
Cyber-Globalization rank: medium
Wikimedia: Yes
GoogleTranslate: Yes
Comments: Its role of lingua franca inside the country plus its spreadin10 more countries (with minor proportion of speakers) may justify a TLD.
Pros: Relatively strong digital existence.
Cons: Low connectivity but could be boosted with TLD. No clear local language policies
Potential representation: Such a decision would be taken and managed by Ministry of Culture - https://culture.gouv.cd/ in association with Société Congolaise des Postes et Télécom. Who manages .cd. https://scpt.cd.
Potential funding:
References:
https://www.researchgate.net/publication/282553750_The_Making_of_Lingala_Corpus_An_Under-resourced_Language_and_the_Internet
https://localizationafrica.com/the-rise-and-rise-of-lingala/
https://gerflint.fr/Base/Afrique_GrandsLacs2/makomo.pdf
ccTLD: .cd

LANGUAGE FORM ISO639-3: wol (Wolof) *

NAME (English, local): Wolof
Classification: Low potential candidate for language
If macro language:
L1+L2: 22 646 100
L1+L2/L1: 3.17
Connected L1+L2: 60%
Countries with speakers: 13 Senegal (22M), Mali (0.08), Italy (0.04), France (0.03), United States (0.02), Mauritania (0.02), Côte d'Ivoire, Gabon, Canada, Guinea-Bissau, United Kingdom, Turkey, Belgium
Virtual Presence Indicator: 0.75
Cyber-Globalization rank: Fair
Wikimedia: Fair
GoogleTranslate: No
Comments: Lingua franca inside Senegal not really abroad. > 90% of speakers are in Senegal
Pros:
Cons: Do not see solid argument to justify a TLD for de facto lingua franca of a unique country except if .sn see otherwise.
Potential representation:
Potential funding:
References:
ccTLD: .sn

ANNEX 2: LANGUAGE MATRIX FOR LAST LEVELS LANGUAGES

In bold, those which could as well be treated as **.

Table 18: Language matrix for last levels candidates

ISO	LANGUAGE	COUNTRY	L1+L2	%C	M/F/G	Cw. Sp	W	GT	COMMENTS
ast	**Asturian**	**Spain**	**700K**	**95**		**2**	**3**	**N**	**10K in Portugal** **https://alladixital.org/**
lim	**Limburgish**	**Netherlands**	**1.3M**	**96**		**3**	**1**	**Y**	**Belgium (0.6)** **https://www.veldeke.net/** **https://hklimburg.nl/** **www.limburgsedialecten.nl**
gag	**Gagauz @**	**Moldova**	**200K**	**~64**		**4**	**1**	**N**	**Ukrania, Bulgaria, Russia, Romania** **www.gagauzia.md (regional)**
csb	**Kashubian @**	**Poland**	**99K**	**~86**		**2**	**2**	**N**	**Canada – Digitally active – Regional.** **https://www.kaszubi.pl/**
man	**Mandingo macro #**	**Guinea** **Gambia** **Mali** **Senegal +3**	**9M**	**43**	emk mwk mku mnk msc mlq	**7**	**0**	**N**	**Makes senses but challenging to setup.**
hrx	Hunsrik	Brazil	3M	84		1	0	Y	Not mature. Not enough Internet presence.
ext	Extremadurian	Spain	500K	95		2	1		1500 in Portugal
fls	West Flemish	Belgium	1.2M	~95		3	1	N	Netherlands+France https://anz.be/
sxu	Upper Saxon	Germany	2M	93		1	0	N	https://www.isgv.de/
gla	Scottish Gaelic	U. Kingdom	60K	95		1	1	Y	https://www.gaidhlig.scot/en/
pdt	Plautdietsch	Canada	362K			12	0	N	At difference with hrx, its population is spread in many countries from Latin America + Kazakhstan, Germany and USA, but does not appear mature neither. No unifying organization.
ryu	Okinawan, Central	Japan	1.2M	85		1	0	N	Regional but not digitally mature
mtq	Muong	Vietnam	1.5M	78		1	0	N	Regional -Ethnic - Christian
mey	Hassaniyya	Mauritania	5.2M	45		10	0	N	9 over 10 are African countries 40% speakers out of Mauritania.
yue	Chinese Yue (Cantonese)	China	87M	high		37	0	Y	Now that Hong Kong (6.7M) has lost autonomy such decision belongs 100% to China except if Australia, Canada and USA (2M together) joint effort for a diasporic Yue TLD…
fry	*Frisian*	*Netherlands*	*720K*	*97*		*2*	*2*	*Y*	*.frl exists for the concerned region*