



Observatorio de la diversidad lingüística y cultural en la Internet
Observatoire de la diversité linguistique et culturelle dans l'Internet
Observatory of the Linguistic and Cultural Diversity in the Internet

<https://OBDILCI.ORG/>

LEÇONS TIRÉES DE LA SÉRIE D'EXPÉRIENCES MECILDI

RAPPORT TECHNIQUE N° 1

La Toile désertique des langues autochtones, vernaculaires, locales ou minoritaires

OBDILCI, septembre 2026

Après avoir appliqué la plateforme d'observation multilingue MECILDI pour mesurer la distribution linguistique du contenu web dans 30 séries web différentes (gTLD, ccTLD ou Tranco), il est temps de tirer les enseignements de cette expérience. Le présent rapport technique est le premier d'une nouvelle série de rapports techniques élaborés par l'Observatoire de la diversité linguistique et culturelle sur Internet (OBDILCI).

Le cadre de calcul développé par OBDILCI, qui mesure la répartition des langues et le multilinguisme sur un ensemble de sites web, a été appliqué à la série Tranco (un million de sites web parmi les plus visités), à 8 gTLD associés aux langues de France, à 10 ccTLD de pays membres de l'Organisation de la Francophonie (OIF) et à 11 autres ccTLD, sélectionnés pour la présence de communautés linguistiques importantes. Ces mesures ont permis de confirmer des données antérieures sur la répartition des langues sur le web et de calculer de nouveaux indicateurs de multilinguisme.

Cette première étude se concentre sur les résultats concernant les langues **autochtones, vernaculaires, locales ou minoritaires** et dresse un tableau particulièrement préoccupant de leur très faible présence sur le Web, avec des données bien inférieures aux prévisions les plus pessimistes. Dans la suite de ce document, nous utiliserons le terme « local » pour désigner l'une quelconque des catégories de langues mentionnées.

Le tableau suivant présente la compilation des données obtenues lors de l'exécution de MECILDI, complétée par les indicateurs appropriés afin d'établir des comparaisons et un classement équitables. Les colonnes du tableau ont les définitions suivantes :

LANGUES : le nom français de la langue locale

ISO : Le code ISO 639-3 associé à cette langue

TLD : Le domaine de niveau supérieur mesuré par MECILDI.

LIEU : Le lieu (pays ou région) correspondant à ce TLD

%SP : % de locuteurs L1+L2 de cette langue dans le lieu. (Source : Ethnologue)

L1+L2 : Nombre de locuteurs de cette langue dans le lieu ou, si indication contraire, dans le monde ou un pays donné. (Source : Ethnologue)

%P : % des pages web dans cette langue par rapport au total dans la zone TLD. (MECILDI)

%S : % des sites web ayant une configuration linguistique dans cette langue. (MECILDI)

WRate : Rapport entre %SP et %S, indicateur de prévalence, 1 étant une proportionnalité stricte.

Le tableau est trié par WRate décroissant .

LANGUES	ISO	TLD	LIEU	%SP	L1+L2	%P	%S	WRate	N
Basque	eus	.eus	Pays Basque	36%	1M	36,37%	94,02%	2,612	1
Catalan	cat	.ad	Andorre	37%	30900	20,75%	61,68%	1,672	2
Breton	bre	.bzh	Bretagne	5%	200K (M)	2,91%	2,19%	0,438	
Papiamentu	pap	.cw	Curaçao	92%	136K	11,52%	18,47%	0,201	
Tahitien	tah	.pf	Polynésie française	19%	56K	1,69%	2,67%	0,144	
Malgache	mlg	.mg	Madagascar	58%	18M	4,61%	7,58%	0,131	
Corse	cos	.corsica	Corse	35%	120K (M)	3,55%	3,68%	0,105	
Créole haïtien	hat	.ht	Haïti	95%	13,8M (M)	3,41%	7,89%	0,083	3
Bislama	bis	.vu	Vanuatu	34%	115K (M)	0,99%	1,60%	0,047	
Swahili	swc	.cd	Rép. Dém. du Congo	25%	32M	-	0,53%	0,021	4
Yiddish	yid	.il	Israël	5%	443K (M)	0,02%	0,03%	0,006	
Rundi	run	.bi	Burundi	98%	13M (M)	-	0,56%	0,006	
Guarani	grn	.py	Paraguay	88%	6M	0,13%	0,26%	0,003	5
Quechua	que	.pe	Pérou	21%	7,3M	0,04%	0,06%	0,003	
Créole guadeloupéen	gcf	.gp	Guadeloupe	90%	400K (M)	0,07%	0,18%	0,002	6
Tamazight	zgh	.ma	Maroc	25%	9M	-	0,04%	0,002	
Kabuverdianu	kea	.cv	Cap-Vert	82%	492K	ε	ε	0,000	7
Ladino	lad	.il	Israël	0,5%	43K	ε	ε	0,000	8
Alsacien	gsw	.alsace	Alsace	35%	900K (FR)	0	0	0	
Aymara	aym	.bo	Bolivia	14%	1.7M	0	0	0	
Jula	dyu	.ci	Côte d'Ivoire	27%	8.5M	0	0	0	
Langues Kanak		.nc	Nouvelle Calédonie	36%	71K	0	0	0	9
Kimbundu + Divers		.ao	Angola	90%	33M	0	0	0	10

NOTES

- 1 : Bilinguisme quasi parfait avec l'espagnol
2. Un partage plus ou moins équilibré entre le catalan, l'anglais, l'espagnol et le français.
- 3 : 80 % de la population ne parle que le créole
- 4 : Parmi de nombreuses autres langues autochtones telles que le lingala (lin) 40M, toutes absentes
- 5 : Le guarani est la langue officielle du Paraguay
- 6 : Ce qui a été détecté est du créole haïtien
- 7 : Un site unique a été trouvé parmi plus de 1100 sites traités.
- 8 : Quatre sites web contenant du ladino (lad) ont été trouvés
- 9 : Drehu (12K), wallisien (18K) et 32 autres langues kanak.
- 10 : Kimbundu 3,7 M, Umbundu 5,9 M, Chokwe 2,4 M, Kikongo 2,4 M, Nhaneka 1,4M, Nganguela 700K, Kwanhama 1M, Fiote 391K, Muhumbi 253K, Luvala 180K

Que nous apprennent ces données ?

À l'exception du basque et du catalan, qui font preuve d'une vitalité remarquable sur le web, toutes les langues locales étudiées présentent une présence et une vitalité extrêmement faibles. Le WRate, indicateur de la prévalence de la langue locale au sein du TLD correspondant, affiche des valeurs nulles ou quasi nulles pour un tiers des langues, des valeurs critiques (inférieures à 0,01) pour plus de la moitié des langues, et des valeurs très faibles (entre 0,1 et 0,5) pour 20 % d'entre elles.

Une trentaine de lieux analysés ne suffisent évidemment pas à tirer des conclusions statistiques définitives, mais cela suffit pour susciter une préoccupation légitime.

D'après ses études précédentes, OBDILCI s'attendait à trouver la plupart des résultats dans la catégorie « valeurs très faibles », et non dans les valeurs critiques et nulles.

Des exemples comme les suivants permettent de constater l'ampleur du problème auquel on est confronté et l'urgence de mettre en place des politiques publiques appropriées :

- **Le guarani** est la langue officielle du Paraguay et bénéficie d'une attention croissante de la part des politiques publiques. Il est parlé par 88 % de la population paraguayenne, pourtant seulement 0,26 % des sites web ont une configuration en guarani, soit très peu d'unités.
- En Haïti, 80 % de la population ne comprend que **le créole** ; cette énorme proportion de la population haïtienne a un accès linguistique à moins de 8 % de l'écosystème web national.
- Depuis la fin des années 80, un scientifique bolivien a créé et développé un algorithme de traduction remarquable et tout à fait original utilisant **l'aymara** comme langue pivot¹; pourtant, il n'existe absolument aucun site web dans cette langue, bien qu'elle soit parlée par 14 % de la population.

¹ <https://www.obdilci.org/blog/ivan-guzman-de-rojas-and-the-aymara-language-2/>

- **Le papiamento** est la langue dominante de Curaçao, île particulièrement multilingue ; plus de 90 % de la population l'utilise quotidiennement, pourtant le web de Curaçao ne contient qu'un peu plus de 10 % de pages dans cette langue.
- **Le kabuverdien** est la langue dominante du Cap-Vert, parlée par plus de 80 % de la population, contre 60 % pour le portugais ; pourtant, le seul site web recensé est celui de l'Institut du patrimoine culturel.
- **Le rundi** est la langue principale du Burundi, comprise pratiquement par tous les habitants du pays, pourtant elle n'est présente que dans moins de 1% des sites web du domaine .bi.

Nous approchons du milieu de la *Décennie des langues autochtones* promue par l'UNESCO ; de nombreuses organisations, telles que l'OIF, organisent des ateliers d'experts pour relever les défis posés par l'IA afin d'intégrer pleinement les langues locales ; un nombre croissant de langues (peut-être 750) ont obtenu l'attention d'UNICODE pour recevoir la codification requise pour l'existence numérique, mais qu'en est-il des **contenus** dans ces langues ?

MÉTHODE

Nous avons appliqué la plateforme MECILDI à la liste complète des sites web identifiés sous le TLD des pays ou régions mentionnés. Dans certains cas, nous avons utilisé la zone officielle du TLD, lorsqu'elle était partagée par le centre d'information du réseau, via l'ICANN/CZDS ; dans les autres cas, nous avons utilisé un fournisseur commercial (zonestats.io). Dans ce dernier cas, la liste des domaines pourrait ne pas inclure certains sites web récents en langue locale, mais le résultat statistique serait de toute façon resté très proche.

MECILDI explore systématiquement tous les domaines situés sous le TLD ; une partie importante est cependant exclue de la mesure pour de nombreuses raisons possibles, notamment la volonté du propriétaire du site web de ne pas être traité par des robots, mais là encore, cette situation ne modifierait pas les résultats statistiques de manière significative.

Tous les résultats détaillés sont consultables publiquement, voir les références, à l'exception de la dernière série de ccTLD, qui n'a pas encore été publiée, l'exemple du Cap-Vert (.cv) est fourni en annexe pour permettre la compréhension du processus et du type d'indicateur obtenu.

RÉFÉRENCES :

- Pages MECILDI donnant accès à tous les résultats publiés : <https://obdilci.org/projets/mecildi/>
- Pimienta D., Field M., 2026, *MECILDI : une plateforme d'observation multilingue pour mesurer la distribution linguistique du contenu Web*.
<https://www.obdilci.org/wp-content/uploads/2026/08/MECILDI-V1-FR.pdf>

ANNEXE : Les résultats complets pour le Cap-Vert (.cv)

INDICATEURS	Définition de l'indicateur
%SITES (L)	Sites Web disposant d'une version linguistique en langue L
%PAGES (L)	Pourcentage de pages web en L (nombre de sites en L divisé par le nombre de versions linguistiques)
MultiR	Taux de multilinguisme (nombre de versions linguistiques divisé par le nombre de sites Web)
%Multi	Pourcentage de sites web multilingues
AvgL	Nombre moyen de langues par site web multilingue
ERR-TOT	Pourcentage total de sites web non traités
ERR-DNS	Domaine inaccessible
ERR-BLK	Sites valides mais l'accès par programme n'est pas autorisé.
ERR-CFG	Site stationné ou en construction
ERR-HTTP	Erreur du site Web
ERR-LANG	Site valide, mais échec de la détection de la langue
MonoL	Pourcentage de sites web monolingues
GooT	Pourcentage de sites web utilisant Google Traduction
HDEF	Facteur d'extrapolation dérivé par heuristique.
Lang=	Pourcentage de sites web utilisant l'attribut HTML Lang=
HrefLang =	Pourcentage de sites web utilisant l'attribut HTML HrefLang =
IDN	Pourcentage de sites avec des noms de domaine internationalisés
NON-ID	Langue non identifiée par l'algorithme de détection

Nombre de sites analysés : **2 429**

Nombre de sites validés pour ce processus : **1 212**

MultiR **1.79**

%Multi **23,98%**

AvgL **4,30**

MonoL **76,02%**

GooT **1,97%**

HDEF **2.62**

Lang= **79,51%**

Lang=Err **11,39%**

IDN **0,00%**

ERR-DNS **18,37%**

ERR-BLK **2,87%**

ERR-CFG **1,76%**

ERR-HTTP **6,03%**

ERR-LANG **53,10%**

ERR-TOT **82,13%**

DISTRIBUTION DES LANGUES	%PAGES	%SITES
TOTAL	100,00%	178,48%
Anglais	33,16%	59,18%
Portugais	25,12%	44,83%
Français	5,86%	10,45%
Indonésien	4,44%	7,93%
Vietnamien	3,91%	6,97%
Espagnol	2,10%	3,76%
Allemand	1,80%	3,21%
Chinois	1,68%	3,01%
Néerlandais	1,19%	2,12%
Japonais	1,15%	2,05%
Italien	1,03%	1,84%
Russe	0,96%	1,71%
Arabe	0,84%	1,50%
Coréen	0,80%	1,43%
Turc	0,73%	1,30%
Polonais	0,69%	1,23%
Thaïlandais	0,61%	1,09%
Bulgare	0,50%	0,89%
Tchèque	0,46%	0,82%
Malais	0,46%	0,82%
Roumain	0,46%	0,82%
Suédois	0,46%	0,82%
Danois	0,34%	0,61%
Grec moderne	0,34%	0,61%
Finlandais	0,34%	0,61%
Ukrainien	0,34%	0,61%
Latin	0,27%	0,48%
Persan	0,27%	0,48%
Bengali	0,23%	0,41%
Estonien	0,23%	0,41%
Hindi	0,23%	0,41%
Hongrois	0,23%	0,41%
Javanais	0,23%	0,41%
Letton	0,23%	0,41%
Lituanien	0,23%	0,41%
Mongol	0,23%	0,41%
Birman	0,23%	0,41%

Norvégien	0,23%	0,41%
Serbe	0,23%	0,41%
Tagalog	0,23%	0,41%
Non-ID	0,19%	0,34%
Basque	0,15%	0,27%
Irlandais	0,15%	0,27%
54 autres langues	0,11%	0,20%

Remarques :

- Le code ISO639-3 pour Kabuverdianu (kea) n'est pas présent dans les champs HTML Lang ou Hreflang.
- Le code pays (cv) a pu être utilisé comme code de langue dans de rares cas, comme par exemple <https://ipc.cv/cv/>.
- Les sites indonésiens apparaissent dans les redirections, principalement vers des sites de jeux vidéo.
- Le fichier .cv est vendu dans le monde entier et permet de stocker son CV professionnel au format JohnSmith.cv. Les administrateurs de ces sites les ont configurés pour interdire l'exploration par les robots, ce qui n'a donc pas influencé les résultats, hormis un pourcentage d'erreurs ERR-LANG très élevé, MECILDI étant dans l'impossibilité d'analyser leur contenu.
- Ces sites représentent plus de la moitié du nombre total de sites web.