WIKIMÉDIA FRANCE
LINGUA LIBRE

ÉTAT DES LIEUX DU PROJET LINGUA LIBRE

ENREGISTREMENT D'UNE SONOTHÈQUE LIBRE
POUR LA DIVERSITÉ LINGUISTIQUE

DÉROULÉ DE L'ARGUMENTAIRE

POURQUOI A-T-ON CRÉÉ LINGUA LIBRE ?

VISION INITIALE

TOUTES LES LANGUES DEVRAIENT ÊTRE DOCUMENTÉES SUR LES PROJETS WIKIMÉDIA

PROBLÈME

LES LOCUTEURS DE LANGUES PEU DOTÉES FONT FACE À DE GROS FREINS DE CONTRIBUTION À L'ÉCRIS PARCE QUE :

- IL EXISTE DE NOMBREUSES ORTHOGRAPHES DE LEUR LANGUE
- LES SOURCES TEXTUELLES MANQUENT
- LES LOCUTEURS ONT UNE CONNAISSANCE PARTIELLE DE LEUR LANGUE
- LEUR LANGUE N'EST PAS ORTHOGRAPHIÉE

SOLUTION

LES LOCUTEURS ONT DES CONNAISSANCES ORALES À PARTAGER, MÊME FRAGMENTAIRES ELLES SONT UN PATRIMOINE PRÉCIEUX. POUR LEVER UNE PARTIE DES FREINS DE CONTRIBUTION, ON LANCE UN OUTIL D'ENREGISTREMENT AUDIO : LINGUA LIBRE.

HISTORIQUE TECHNIQUE

2016	V1 LANCÉE
2018	V2 LANCÉE (CORPUS REDÉMARRÉ À ZÉRO)
2019-2020	REFONTE GRAPHIQUE ET AMÉLIORATION DE L'EXPÉRIENCE UTILISATEUR
2021	MISE À JOUR MEDIAWIKI, CRÉATION OUTIL SONOTHÈQUE

OBJECTIFS DE L'OUTIL EN 2018

ENREGISTRER DES LISTES DE MOTS DE MANIÈRE RAPIDE ET AUTOMATIQUE (300 À 800 MOTS PAR HEURE SELON LES PERSONNES)

STOCKER CES FICHIERS AUDIOS SOUS **LICENCE LIBRE** SUR WIKIMEDIA COMMONS, EN LES CLASSANT PAR LANGUE/UTILISATEUR/ORTHOGRAPHE

RÉUTILISER CES SONS SUR D'AUTRES PROJETS WIKIMEDIA GRÂCE AU LINGUA LIBRE BOT (EXTERNE) QUI ATTRIBUE UN ITEM WIKIDATA ET WIKIDATA LEXEME À CHAQUE ENREGISTREMENT, ET QUI ANCRE LE FICHIER AUDIO SUR LA PAGE DU WIKTIONNAIRE CORRESPONDANT À SON ORTHOGRAPHE.

PERMETTRE AUX UTILISATEURS DE S'ORGANISER EN **COMMUNAUTÉ** GRÂCE AUX PAGES DE DISCUSSION ET AU FONCTIONNEMENT EN WIKI

DOUBLE IMPACT VISÉ

PARTAGER DES CONNAISSANCES DANS ET AUTOUR DES LANGUES PEU DOTÉES OU MINORITAIRES

OFFRIR UNE PORTE D'ENTRÉE AUX LANGUES PEU DOTÉES VERS LES PROJETS WIKIMÉDIA

ÉTAT D'AVANCEMENT

AVANCÉES

QUANTITATIVES

687 000 ENREGISTREMENTS
150 LANGUES
782 PERSONNES Y ONT CONTRIBUÉ

QUALITATIVES

RETOURS TRÈS POSITIFS SUR LA TECHNOLOGIE DU RECORD WIZARD DE LA PART DE DÉVELOPPEURS AGUERRIS (WIKIMEDIA SUÈDE ET GOOGLE RESEARCH)

OUTIL ADAPTÉ POUR LES LANGUES AYANT PEU DE LOCUTEURS (EX : SURUI / PICARD)

CERTAINS LOCUTEURS DE LANGUES « MOINS DIFFUSÉES ET MOINS ENSEIGNÉES » (MODIMES) S'EN SONT EMPARÉS AVEC ENTHOUSIASME ...

Building a 50,000 pronunciation data repository in the Odia language

10 March 2022 by Subhashish Panigrahi

Last year, I started a small pilot under the OpenSpeaks project for building voice data as a foundational layer for speech synthesis research and application development. To test some of the learning in this field, I started building a wordlist by collecting words from multiple sources including Odia Wikipedia and Odia Wiktionary, and started recording pronunciations using Lingua Libre. Recently, the pilot hit a 55,000 pronunciation milestone. The repository also includes pronunciations of 5,600 words in Baleswaria, the northern dialect of Odia. All the recordings were also released under a Public Domain (Creative Commons CC0 1.0) release on Wikimedia Commons. These recordings make the largest repository of Public-Domain voice data in Odia, and add to another 4,000+ recordings of sentences in Odia on Mozilla Common Voice.

DÉFIS DIVERSITÉ

- LES 10 LANGUES LES PLUS DOCUMENTÉES POSSÈDENT DÉJÀ UNE VERSION DE WIKIPÉDIA
- SUR 150 LANGUES SEULES 35 ONT PLUS DE 1 000 MOTS ENREGISTRÉS
- LA RAPIDITÉ DU RECORD WIZARD SIGNIFIE QU'UN CORPUS PEUT ÊTRE CRÉÉ PAR UNE SEULE PERSONNE
- VOIX PRINCIPALEMENT MASCULINES

Langue	code ISO 639-3	Nombre d'enregistrements
français	fra	241836
polonais	pol	81777
bengali	ben	58982
oriya	ori	50956
espéranto	epo	33446
anglais	eng	22674
roumain	ron	19401
ukrainien	ukr	19091
allemand	deu	14956
marathi	mar	14407

LANGUES 1-10

Langue	code ISO 639-3	Nombre d'enregistrements
néerlandais	nld	1546
langues bicol	bik	1453
arabe marocain	ary	1286
vietnamien	vie	1231
finnois	fin	1174
malgache	mlg	932
persan	fas	883
gallois	cym	861
mandarin	cmn	829
arménien	hye	783

LANGUES 30-40

DÉFIS TECHNIQUES

TOUTES LES FONCTIONS DE LINGUA LIBRE ONT UN DÉFAUT DE FONCTIONNEMENT MAJEUR

VOIR SUR LINGUA LIBRE

LES BESOINS TECHNIQUES DE LA COMMUNAUTÉ AUGMENTENT TRÈS VITE

WISHLIST

LINGUA LIBRE EST TROP COMPLEXE POUR LE FORMAT HACKATHONS

DIFFÉRENTS FORMATS ONT ÉTÉ ESSAYÉS (DEMI-JOURNÉE / 7 JOURS / 3X5 JOURS)

Modular8951 06/06/2022

I decided to leave the Project forever. Here is some constructive feedback addressed to WM France @Adélaïde_Calais_WMFr @RémyWMFr @Michaël-WMFR & Lingua Libre Community :

1) Stop being so French-Centric ! Seriously ! Most of LL is designed for the exclusive benefits of French Speakers that benefit from recordings on the FR Wiktionary. 😤

2) The Statistics show that the project is dead. Almost no recordings in the last months, and French reached a plateau. The only way to attract contributors is to provide mutual benefit, but that does not happen now.

Currently, LinguaLibre just grabs people voices to make a better French Wiktionary. Useless for non Francophones. Therefore, nobody bothers to record on the website.🤭

3) The UI of LinguaLibre is USELESS for language learners. Nobody can use it to search a word and learn how to pronounce it. Without Language Learners, the website will continue to be dead as now.

4) No Full-time programmers = No future. The project will continue to be dead without WM Foundation paying full-time developers.

5) There should be more PR Commits on GitHub by PAID full-time developers and less cheap talk. Full period. GOODBYE.😕 (modifié)

👍 1 😔 1 🔥 1 🧡 1

Modular8951 06/06/2022

PS. Please do not see the above comment as a personal attack. It is meant to be constructive feedback. (modifié)

Poslovitch 06/06/2022

And he's gone 😢

DÉFIS STRATÉGIQUES

SUR LE TERRAIN

LA CONTRIBUTION QUI DOUBLAIT CHAQUE ANNÉE DE 2018 À 2021, S'EST IMMOBILISÉE EN 2022

LE DÉVELOPPEMENT AU COUP PAR COUP OCCASIONNE D'ÉNORMES PERTES DE CONNAISSANCES ET L'ÉPUISEMENT DES BÉNÉVOLES

LINGUA LIBRE N'A QU'UN SEUL MÉCÈNE, QUI EST AU MAXIMUM DE SON BUDGET

MÉTA

EN 4 ANS LES OBJECTIFS DE LA COMMUNAUTÉ SE SONT DIVERSIFIÉS - RISQUE DE PERTE D'IDENTITÉ

LINGUA LIBRE A UN POTENTIEL ÉNORME, QUI DEMANDE LE TRAVAIL D'UNE À DEUX PERSONNES À TEMPS PLEIN, DONC D'UN REPOSITIONNEMENT STRATÉGIQUE DE WIKIMÉDIA FRANCE

COMMUN NUMÉRIQUE POUR LA DIVERSITÉ LINGUISTIQUE

AVANCÉES SCIENTIFIQUES EN TRAITEMENT DU LANGAGE NATUREL

ENREGISTRER DES MOTS N'EST PAS SUFFISANT POUR DOCUMENTER UNE LANGUE : BESOIN DE CORPUS DE PHRASES POUR DOCUMENTER GRAMMAIRE ET SYNTAXE

ALIGNEMENT AVEC LA STRATÉGIE 2030 ET BESOINS DES COMMUNAUTÉS

DEDICATE A SIGNIFICANTLY LARGER AMOUNT OF MOVEMENT FUNDING TO SUPPORT EMERGING AND MARGINALIZED COMMUNITIES AND GROUPS BASED ON THEIR NEEDS ...

MOUVEMENT WIKIMEDIA EN RETARD SUR LA CONNAISSANCE ORALE

PAS DE PROJET DE PARTAGE DES CONNAISSANCES ORALES GLOBAL. LINGUA LIBRE (ET SURTOUT WIKIMEDIA FRANCE) NE DOIVENT PAS DEVENIR CE PROJET MAIS SE POSITIONNER DANS LES DISCUSSIONS DU MOUVEMENT À CE SUJET ET PRÉVOIR QUE LINGUA LIBRE S'IMBRIQUE DANS UNE FUTURE SOLUTION GLOBALE

RECENTRER LA MISSION DE LINGUA LIBRE

DOCUMENTER LA DIVERSITÉ LINGUISTIQUE

LE BUT DE LINGUA LIBRE N'EST PAS DE DOCUMENTER LES LANGUES HÉGÉMONIQUES

PRIORISER LES FONCTIONALITÉS ET PROJETS

LE CHAMP DES BESOINS AUXQUELS LINGUA LIBRE POURRAIT RÉPONDRE EST ÉNORME. AUJOURD'HUI WIKIMÉDIA FRANCE DOIT CHOISIR QUELS ASPECTS SONT PRIORITAIRES CAR ON NE POURRA PAS TOUT FAIRE

DIRECTION STRATÉGIQUE

BESOIN D'UN PLAN STRATÉGIQUE SUR 5 ANS POUR ASSURER LA PÉRENNITÉ DU PROJET

PÉRENNITÉ DE LINGUA LIBRE

STABILITÉ TECHNIQUE
BESOIN URGENT D'UN ÉQUIVALENT TEMPS PLEIN EN DÉVELOPPEMENT

INTERNATIONALISATION DU PROJET
BESOIN URGENT D'UN ÉQUIVALENT TEMPS PLEIN EN DÉVELOPPEMENT

ASSURER QUE LINGUA LIBRE NE DEVIENNE PAS OBSOLÈTE DANS 5 ANS
PRÉVOIR L'IMBRICATION DE LINGUA LIBRE DANS UN PROJET DE PARTAGE DE CONNAISSANCES ORALES PLUS LARGE

RÉFÉRENCES

Langues moins Diffusées et moins Enseignées (MoDiMEs)/Less Widely Used and Less Taught languages, Langues enseignées, langues des apprenants/Language learners' L1s and languages taught as L2s, de Fryni Kakoyianni-Doa, Monique Monville-Burston, Salomi Papadima-Sophocleous, Freiderikos Valetopoulos, 2020, Comptes-rendus de conférences 194 Pages.

Building Machine Translation Systems for the Next Thousand Languages, Ankur Bapna, Isaac Caswell, Julia Kreutzer, Orhan Firat, Daan van Esch, Aditya Siddhant, Mengmeng Niu, Pallavi Baljekar, Xavier Garcia, Wolfgang Macherey, Theresa Breiner, Vera Axelrod, Jason Riesa, Yuan Cao, Mia Xu Chen, Klaus Macherey, Maxim Krikun, Pidong Wang, Alexander Gutkin, Apurva Shah, Yanping Huang, Zhifeng Chen, Yonghui Wu, Macduff Hughes

Mining Training Data for Language Modeling across the World's Languages, Manasa Prasad Theresa Breiner Daan van Esch