🇧🇮 Guide de Contribution – Ijwi ry'Ikirundi AI

🚀 Comment Contribuer

Ce document explique les **règles et standards exacts** pour contribuer au dataset.
Respecter ces directives garantit des données **propres, de haute qualité et utilisables** pour les modèles ASR/TTS/Traduction.

1. Structure du Fichier Maître (`metadata.csv`)

Le fichier **metadata.csv** doit respecter **exactement** le schéma suivant (12 colonnes).
Chaque contribution doit s'aligner sur cette structure.

Catégorie	Colonne	Description	Obligatoire
Contenu	Kirundi_Transcription	Phrase originale en Kirundi	OUI
Contenu	French_Translation	Traduction en français	OUI (ou English)
Contenu	English_Translation	Traduction en anglais	OPTIONNEL
Contenu	Domain	Catégorie linguistique (Conjugaison, Grammaire…)	OPTIONNEL
Contenu	Source	Source de la phrase (URL, document)	OPTIONNEL
Audio	File_Path	Nom exact du fichier audio	OUI si audio
Audio	Duration	Durée du clip audio en secondes	OUI si audio
Audio	Speaker_id	ID unique du locuteur	OUI si audio
Audio	Age	Âge du locuteur	OUI si audio
Audio	Gender	Genre du locuteur (M/F/O)	OUI si audio
Admin	Machine_Suggestion	Suggestion générée par IA	AUTOMATIQUE
Admin	Kirundi_Length	Nombre de caractères	AUTOMATIQUE

---

2. Guide de Contribution (Selon l’Action)

Le type de contribution détermine quelles colonnes doivent être remplies.
Les colonnes optionnelles ou gérées par l’ADMIN doivent rester **vides**.

A. Contribution de Texte et Traduction (Pull Request GitHub)

Pour ajouter de nouvelles phrases ou traductions :

Colonne	Obligatoire ?	Notes
Kirundi_Transcription	OUI	Clé principale de la ligne
French_Translation	OUI (ou English)	Fournir une traduction de qualité
English_Translation	OPTIONNEL	Peut être généré par IA si absent
Domain & Source	OPTIONNEL	Recommandé si connu
Colonnes Audio/Admin	NE PAS REMPLIR	Laisser vide

B. Contribution Audio (Soumise à l’ADMIN)

Pour enregistrer des clips audio pour des phrases existantes :

Colonne	Obligatoire ?	Notes
File_Path	OUI	Nom exact du fichier (ex : `2025_12_07_001.wav`)
Duration	OUI	Durée en secondes (ex : `2.45`)
Speaker_id	OUI	Format anonyme : `S01_M`
Age & Gender	OUI	Informations démographiques
Kirundi_Transcription	OUI	Doit correspondre EXACTEMENT au texte original

---

3. Gestion des Fichiers Audio (Processus Spécial)

⚠️ **Ne jamais envoyer les fichiers audio sur GitHub.**

Workflow correct :

1. Le contributeur enregistre le clip audio.
2. Le nomme exactement comme indiqué dans `File_Path`.
3. Envoie les fichiers audio (WhatsApp, email, Drive, etc.).
4. **ADMIN (Samandari)** vérifie et pousse les fichiers sur **Hugging Face**, configuré avec LFS.

4. Critères de Qualité et Normalisation

Pour garantir l'efficacité des modèles ASR/TTS, chaque phrase Kirundi doit respecter des normes strictes.

A. Normalisation de la Transcription (Kirundi_Transcription)

Règle	Description	Pourquoi ?
Pleine orthographe	Pas d'abréviations, de chiffres ou de symboles (&, @…).	L'IA apprend la prononciation des mots, pas des raccourcis.
Ponctuation finale	Chaque phrase se termine par `.`, `?` ou `!`.	Permet au TTS de détecter les limites et le ton des phrases.
Majuscules initiales	Majuscule uniquement sur le premier mot ou les noms propres.	Entraîne la capitalisation automatique.
Diacritiques obligatoires	Accents et voyelles longues (â, ū, é, í…) si présents dans la source.	Essentiel pour la prononciation et la prosodie correcte.
Limitation de longueur	Idéal : 4–25 mots (max 30).	Facilite l'enregistrement et l'alignement texte/audio.

B. Normalisation des Métadonnées Audio

Colonne	Format requis	Explication
Speaker_id	Format anonyme `S_01_M`	Assure confidentialité et cohérence.
Age	Catégories générales (`20s`, `30s`, `40s+`)	Diversité sans divulguer l'âge réel.
Duration	Nombre flottant (`3.15`)	Nécessaire pour aligner le texte et l'audio correctement.
Gender	Male/Female	Nécessaire pour les modeles

🤝 Remerciements

Merci de contribuer à la préservation et à la modernisation de la langue Kirundi.