Note de concept : Évaluation des capacités de l'OCR pour la numérisation des supports en langues africaines

1. Contexte

Une part importante du savoir historique, culturel, éducatif, administratif et linguistique de l'Afrique demeure contenue dans des documents physiques ou numérisés qui ne sont pas facilement consultables ou exploitables par des machines. Ces supports comprennent des journaux, des livres, des rapports, des magazines, des publications gouvernementales, des ressources éducatives, des documents de recherche, des dictionnaires et des archives historiques.

Bien que certains de ces supports aient été numérisés et conservés sous forme numérique, une image numérisée n'équivaut pas à un contenu numérique exploitable par une machine. Sans extraction fiable du texte, ces informations restent difficiles à rechercher, analyser, traduire, réutiliser ou intégrer dans les systèmes d'intelligence artificielle modernes.

La reconnaissance optique de caractères (OCR) offre l'opportunité de transformer ces ressources en texte numérique structuré. Cependant, les technologies OCR ont historiquement été optimisées pour des langues et des collections de documents bénéficiant d'une représentation numérique importante. Leur efficacité pour les langues africaines et les supports d'archives africains demeure insuffisamment comprise.

Ce projet propose d'évaluer les technologies OCR et de compréhension de documents existantes pour la numérisation de supports en langues africaines, en utilisant le Rwanda comme étude de cas principale.

Le projet examinera en particulier les journaux, livres et autres supports rwandais contenant du kinyarwanda, de l'anglais et du français.

Digital Umuganda sera responsable de la collecte, de la sélection, de la préparation et du traitement des documents nécessaires au projet. Le/la chercheur(se) ne sera donc pas responsable de la recherche ou de la préparation manuelle de la collection de documents sous-jacente. Ses travaux se concentreront plutôt sur l'évaluation des technologies OCR, l'analyse comparative de leurs performances, le développement d'un flux de travail pratique de numérisation, l'analyse des résultats et la démonstration de la manière dont ces technologies peuvent être intégrées dans une application utilisable.

La durée initiale sera de trois mois, avec une possibilité de prolongation en fonction des résultats obtenus.

2. Objectif du projet

L'objectif général est d'évaluer les capacités des technologies OCR et de compréhension de documents existantes pour la numérisation des supports rwandais, et de déterminer quelles approches techniques sont les plus adaptées pour transformer ces ressources en contenu numérique précis, consultable et réutilisable.

Le projet évaluera des supports tels que :

- Les journaux ;
- Les livres ;
- Les magazines ;
- Les publications gouvernementales et institutionnelles ;
- Les rapports ;

- Les supports éducatifs ;
- Les publications de recherche ;
- Les archives historiques ; et
- D'autres documents imprimés ou numérisés.

La question de recherche centrale est :

Dans quelle mesure les technologies OCR et de compréhension de documents basées sur l'IA peuvent-elles numériser efficacement les supports rwandais, en particulier le contenu en kinyarwanda, et quelles approches offrent la meilleure combinaison de précision, d'efficacité et d'utilisabilité ?

3. Objectifs spécifiques

Le projet visera à :

- Évaluer des technologies OCR existantes sélectionnées à l'aide de supports rwandais fournis et traités par Digital Umuganda.
- Comparer différentes approches OCR, y compris les systèmes OCR conventionnels et les technologies émergentes de compréhension de documents basées sur l'IA.
- Évaluer la performance de l'OCR selon différents types de documents, mises en page, qualités d'impression, périodes et langues.
- Mener une évaluation ciblée de la performance de l'OCR pour le kinyarwanda.
- Établir des indicateurs et des méthodologies appropriés pour évaluer la performance de l'OCR sur des supports en langues africaines.
- Identifier les erreurs de reconnaissance courantes et les limitations techniques affectant les documents rwandais.
- Évaluer des approches permettant d'améliorer les résultats de l'OCR grâce au prétraitement et au post-traitement.
- Évaluer la capacité des systèmes OCR à préserver la structure des documents et l'ordre de lecture.
- Développer un flux de travail technique reproductible pour la numérisation des supports rwandais.
- Développer un prototype simple et convivial démontrant le processus de numérisation.
- Produire des recommandations pour l'extension de l'approche à des collections rwandaises plus vastes puis, ultérieurement, à d'autres langues africaines.

4. Rôles et responsabilités

Digital Umuganda

Digital Umuganda fournira les fondations nécessaires à la recherche et à l'évaluation techniques.

Digital Umuganda sera responsable de :

- L'identification et la collecte des supports rwandais pertinents ;
- La sélection d'échantillons de documents représentatifs ;
- La numérisation ou la digitalisation des supports physiques, si nécessaire ;
- La préparation des documents en vue de l'évaluation ;
- Le traitement et l'organisation des données sous-jacentes ;
- La fourniture de données de référence appropriées nécessaires à l'évaluation ;
- La définition des technologies OCR ou catégories de technologies à examiner, le cas échéant ;
- L'accès aux travaux antérieurs de Digital Umuganda pertinents pour la numérisation en langues africaines ;
- La fourniture de l'infrastructure et des ressources techniques nécessaires, lorsque disponibles ; et

- La supervision et l'accompagnement techniques tout au long du projet.

Cette approche permet au/à la chercheur(se) de se concentrer principalement sur l'évaluation technique et l'application des technologies OCR, plutôt que sur la collecte et la préparation des données.

Chercheur(se)

Le/la chercheur(se) se concentrera sur :

- L'examen de la collection de documents fournie ;
- La mise en place et le déploiement des technologies OCR sélectionnées ;
- Le développement de la méthodologie d'évaluation ;
- La réalisation d'expériences OCR ;
- L'évaluation comparative des différentes approches ;
- L'analyse des erreurs de reconnaissance ;
- L'évaluation de la préservation de la structure et de la mise en page des documents ;
- L'étude des approches de prétraitement et de post-traitement ;
- Le développement du flux de travail de numérisation ;
- Le développement de l'application prototype ;
- La documentation des expériences et des résultats ; et
- La production de recommandations fondées sur les conclusions.

5. Focus géographique et linguistique

Le principal focus portera sur les supports rwandais.

La priorité linguistique sera :

- Le kinyarwanda
- L'anglais
- Le français

Le projet examinera également les documents multilingues combinant ces langues.

Le kinyarwanda fera l'objet d'une attention particulière, l'un des principaux objectifs étant de comprendre dans quelle mesure les technologies OCR existantes sont performantes pour le contenu en langues africaines, par rapport aux langues qui ont historiquement bénéficié d'une représentation numérique nettement plus importante.

Bien que le Rwanda serve d'étude de cas principale, la méthodologie d'évaluation devra être conçue de manière à pouvoir être ultérieurement adaptée à d'autres langues et collections de documents africaines.

6. Catégories de documents

Digital Umuganda fournira des documents représentatifs couvrant différentes catégories et niveaux de qualité.

Ceux-ci pourront inclure :

Journaux

Les journaux sont particulièrement utiles pour tester l'OCR, car ils peuvent comporter des mises en page complexes incluant plusieurs colonnes, des titres, des photographies, des légendes, des publicités, différentes tailles de police et du contenu multilingue.

Livres

Les livres peuvent inclure des œuvres littéraires, des manuels scolaires, des ouvrages historiques, des dictionnaires, des ouvrages de référence et d'autres publications.

Documents gouvernementaux et institutionnels

Ceux-ci peuvent inclure des rapports, des documents de politique, des publications administratives, des rapports statistiques, des publications de recherche et d'autres archives institutionnelles.

Supports historiques et culturels

Le cas échéant, l'évaluation pourra inclure des magazines, des bulletins d'information, des publications historiques, des supports éducatifs et d'autres ressources d'archives.

L'objectif sera de tester les technologies OCR sur un échantillon suffisamment diversifié afin de comprendre comment la performance évolue dans des conditions réalistes.

7. Technologies OCR à évaluer

Le projet évaluera des technologies existantes plutôt que de développer ou d'entraîner un nouveau modèle OCR.

Les catégories de technologies potentielles pourront inclure :

- Les moteurs OCR traditionnels ;
- Les systèmes OCR open source ;
- Les services OCR basés sur le cloud ;
- Les systèmes OCR basés sur l'apprentissage profond ;
- Les modèles de compréhension de documents ;
- Les modèles vision-langage ; et
- Les grands modèles de langage multimodaux.

Une question de recherche importante sera de déterminer si les approches basées sur l'IA plus récentes apportent des améliorations significatives par rapport aux systèmes OCR conventionnels lors du traitement de documents en langues africaines.

La comparaison devra tenir compte non seulement de la précision de la reconnaissance, mais aussi de la vitesse de traitement, des exigences en calcul, du coût, de la facilité de déploiement, de l'évolutivité et de la capacité à comprendre des structures de documents complexes.

8. Cadre d'évaluation

Le/la chercheur(se) développera un cadre d'évaluation standardisé permettant de comparer différentes technologies à l'aide des mêmes échantillons de documents.

Reconnaissance du texte

Les mesures principales pourront inclure :

- Le taux d'erreur sur les caractères ;
- Le taux d'erreur sur les mots ;
- Le texte manquant ;
- Les mots incorrects ;
- Les caractères incorrects ;
- Les erreurs de segmentation des mots ; et
- Les erreurs spécifiques à la langue.

Structure du document

Le projet examinera également la préservation :

- De l'ordre de lecture ;
- Des paragraphes ;
- Des colonnes ;
- Des titres ;
- Des légendes ;
- Des tableaux ;
- Des listes ; et
- D'autres informations structurelles.

Performance opérationnelle

L'évaluation devra également prendre en compte :

- Le temps de traitement ;
- Les exigences matérielles ;
- Le coût ;
- La dépendance aux API ;
- L'évolutivité ;
- La facilité d'intégration ; et
- Les capacités de traitement par lots.

Cela permettra de garantir que le projet identifie non seulement le système le plus précis, mais aussi les technologies les plus pratiques pour un futur programme de numérisation à grande échelle.

9. Évaluation spécifique au kinyarwanda

Une composante centrale du projet consistera à comprendre la performance de l'OCR spécifiquement pour le kinyarwanda.

Le/la chercheur(se) analysera les erreurs récurrentes concernant :

- Le vocabulaire kinyarwanda ;
- Les noms propres ;
- L'orthographe historique ;
- Les limites de mots ;
- L'usage des majuscules ;
- La ponctuation ;
- La reconnaissance des caractères ; et
- La confusion avec des mots d'autres langues.

Cette analyse devra permettre de déterminer si les erreurs sont systématiques et si elles peuvent être corrigées à l'aide du prétraitement, du post-traitement, de dictionnaires, de modèles de langage ou d'autres technologies existantes.

10. Prétraitement et post-traitement

À l'aide des documents et des données traitées fournis par Digital Umuganda, le/la chercheur(se) étudiera comment différentes approches de traitement technique influencent la qualité de l'OCR.

Les expériences de prétraitement pourront inclure des techniques telles que :

- La correction de l'inclinaison (deskewing) ;
- L'amélioration de l'image ;
- La réduction du bruit ;
- L'ajustement du contraste ;
- La segmentation de la page ; et
- La détection de la mise en page.

Le post-traitement pourra étudier :

- La correction tenant compte de la langue ;
- La correction basée sur un dictionnaire ;
- La correction basée sur des règles ;
- La correction assistée par LLM ;
- La reconstruction de la mise en forme ; et
- L'identification de texte potentiellement incorrect ou à faible confiance.

L'objectif n'est pas d'entraîner de nouveaux modèles OCR, mais de déterminer la meilleure combinaison de technologies et de techniques de traitement existantes.

11. Flux de travail de numérisation

L'un des principaux livrables du projet sera un flux de travail technique recommandé pour la numérisation des documents rwandais.

Le processus global pourra suivre l'enchaînement suivant :

Collecte et préparation des documents (Digital Umuganda) → Numérisation et traitement des données (Digital Umuganda) → Prétraitement des documents → Analyse de la mise en page → Extraction de texte par OCR / IA → Post-traitement → Évaluation de la qualité → Validation humaine, si nécessaire → Résultat numérique structuré

Le projet déterminera les technologies et processus les plus appropriés pour les étapes techniques de ce pipeline.

12. Application prototype

Le projet devra aboutir à un prototype de numérisation simple et convivial, démontrant comment le flux de travail OCR recommandé pourrait, à terme, être opérationnalisé.

Un utilisateur devrait pouvoir fournir un document et recevoir :

- Le texte extrait ;
- Un contenu consultable ;
- La structure du document préservée, dans la mesure du possible ;
- Des indicateurs de confiance ou de qualité ;
- Une comparaison côte à côte entre la source et le contenu extrait, lorsque utile ; et
- Des résultats exportables et exploitables par une machine.

Les résultats potentiels pourront inclure un PDF consultable, du texte brut, du texte structuré ou du JSON, selon le document et l'usage prévu en aval.

Le prototype devra démontrer l'application pratique de la recherche plutôt que de rester uniquement un environnement expérimental d'évaluation comparative.

13. Durée du projet

Le projet initial durera trois mois.

Un plan de travail détaillé sera élaboré comme l'un des livrables du projet.

Le projet pourra ensuite être prolongé en fonction des conclusions de la phase initiale.

Une prolongation pourrait porter sur :

- Des collections de documents plus vastes ;
- Des archives rwandaises supplémentaires ;
- L'amélioration de la correction automatisée ;
- Des flux de travail de validation humaine ;
- Des archives numériques consultables ;
- D'autres langues africaines ;
- Des pipelines de traitement à l'échelle de production ; et
- L'intégration du contenu numérisé avec des modèles de langage et d'autres applications d'IA.

14. Livrables attendus

Le projet devrait produire :

- Évaluation des technologies OCR — évaluation des technologies OCR et de compréhension de documents sélectionnées.
- Cadre d'évaluation — méthodologie standardisée pour évaluer la précision de l'OCR, la structure des documents, la performance pour le kinyarwanda, l'efficacité et les considérations opérationnelles.
- Résultats de l'évaluation comparative — résultats comparatifs entre les technologies OCR et les différentes catégories de supports rwandais.
- Évaluation OCR spécifique au kinyarwanda — analyse détaillée de la performance et des erreurs récurrentes lors du traitement du kinyarwanda.
- Résultats des expériences de traitement — analyse de l'impact des approches de prétraitement et de post-traitement sur la performance de l'OCR.
- Flux de travail de numérisation recommandé — un pipeline technique reproductible pour le traitement des supports rwandais.
- Application prototype de numérisation — une démonstration conviviale de l'approche recommandée.
- Plan de travail détaillé du projet — élaboré dans le cadre du projet plutôt que prescrit dans la présente note de concept.
- Documentation technique et code réutilisable — permettant à Digital Umuganda de reproduire et d'étendre les expériences.
- Rapport technique final — documentant la méthodologie, les expériences, les conclusions, les limitations et les recommandations.
- Recommandations pour le passage à l'échelle — recommandations pour passer de l'expérimentation à une numérisation à plus grande échelle des supports rwandais, puis à une extension à d'autres collections en langues africaines.

15. Résultats attendus

Le projet établira des preuves pratiques concernant la maturité des technologies OCR et d'IA existantes pour la numérisation des supports rwandais.

Il devra permettre de déterminer :

- Quelles technologies sont les plus performantes pour le kinyarwanda ;
- Quelles technologies sont les plus performantes selon les différentes catégories de documents ;
- Comment la qualité des documents influence la performance de l'OCR ;
- Si les modèles vision-langage modernes apportent des améliorations significatives par rapport à l'OCR traditionnel ;
- Quelles approches de prétraitement et de post-traitement apportent des améliorations mesurables ;
- Quel niveau de vérification humaine peut être nécessaire ;
- Quelle architecture technique est appropriée pour un traitement à grande échelle ; et
- Quelles limitations doivent être résolues avant tout passage à l'échelle.

Le projet fournira également à Digital Umuganda un cadre d'évaluation réutilisable et une base technique pour de futures initiatives de numérisation.

16. Importance stratégique

Une part importante du savoir historique et contemporain du Rwanda a été créée avant la généralisation de l'édition numérique. Les journaux, livres, rapports, supports éducatifs, publications institutionnelles et archives historiques contiennent des informations précieuses sur la langue, l'histoire, la culture, les institutions et le développement du Rwanda.

La responsabilité de Digital Umuganda dans la collecte et le traitement de ces supports fournira au projet une base de données structurée, tandis que la composante recherche établira dans quelle mesure les technologies OCR modernes peuvent transformer ces supports en connaissances exploitables par une machine.

Cette répartition permet au projet de concentrer ses ressources de recherche sur la question technique centrale : dans quelle mesure les technologies d'IA actuelles peuvent-elles numériser les connaissances en langues africaines ?

Le contenu numérique ainsi obtenu pourra à terme soutenir des archives consultables, des applications éducatives, la recherche, la traduction automatique, le traitement automatique du langage naturel, la génération augmentée par récupération et des grands modèles de langage ancrés localement.

Commencer par le Rwanda offre un environnement maîtrisable permettant d'établir et de valider la méthodologie. L'opportunité à plus long terme est de développer une approche réutilisable pouvant contribuer à la numérisation à grande échelle des connaissances africaines et du contenu en langues africaines pour l'ère de l'IA.