Concept Note: Evaluating OCR Capabilities for Digitizing African-Language Materials

1. Background

A significant amount of Africa's historical, cultural, educational, administrative, and linguistic knowledge remains contained in physical or scanned documents that are not readily searchable or machine-readable. These materials include newspapers, books, reports, magazines, government publications, educational resources, research documents, dictionaries, and historical archives.

While some of these materials have been scanned and preserved digitally, a scanned image is not equivalent to machine-readable digital content. Without reliable text extraction, the information remains difficult to search, analyze, translate, reuse, or incorporate into modern Artificial Intelligence systems.

Optical Character Recognition (OCR) provides an opportunity to transform these resources into structured digital text. However, OCR technologies have historically been optimized for languages and document collections with significant digital representation. Their effectiveness for African languages and African archival materials remains insufficiently understood.

This project proposes to **evaluate existing OCR and document-understanding technologies for digitizing African-language materials, using Rwanda as the primary case study**.

The project will particularly examine newspapers, books, and other Rwandan materials containing Kinyarwanda, English, and French.

Digital Umuganda will be responsible for gathering, selecting, preparing, and processing the documents required for the project. The researcher will therefore not be responsible for sourcing or manually preparing the underlying document collection. Instead, the researcher's work will focus on evaluating OCR technologies, benchmarking their performance, developing a practical digitization workflow, analyzing results, and demonstrating how these technologies can be integrated into a usable application.

The initial duration will be **three months**, with the possibility of extension based on the results.

2. Project Objective

The overall objective is to **evaluate the capabilities of existing OCR and document-understanding technologies for digitizing Rwandan materials and determine**

which technical approaches are most suitable for transforming these resources into accurate, searchable, and reusable digital content.

The project will evaluate materials such as:

- Newspapers;
- Books;
- Magazines;
- Government and institutional publications;
- Reports;
- Educational materials;
- Research publications;
- Historical archives; and
- Other printed or scanned documents.

The central research question is:

How effectively can current OCR and AI-based document-understanding technologies digitize Rwandan materials, particularly Kinyarwanda content, and which approaches provide the best combination of accuracy, efficiency, and usability?

3. Specific Objectives

The project will aim to:

1. Evaluate selected existing OCR technologies using Rwandan materials provided and processed by Digital Umuganda.
2. Compare different OCR approaches, including conventional OCR systems and emerging AI-based document-understanding technologies.
3. Assess OCR performance across different document types, layouts, print qualities, periods, and languages.
4. Conduct a focused evaluation of OCR performance on Kinyarwanda.
5. Establish appropriate metrics and methodologies for evaluating OCR performance on African-language materials.
6. Identify common recognition errors and technical limitations affecting Rwandan documents.
7. Evaluate approaches for improving OCR results through preprocessing and post-processing.
8. Evaluate the ability of OCR systems to preserve document structure and reading order.
9. Develop a reproducible technical workflow for digitizing Rwandan materials.
10. Develop a simple and user-friendly prototype demonstrating the digitization process.
11. Produce recommendations for scaling the approach to larger Rwandan collections and, subsequently, other African languages.

4. Roles and Responsibilities

Digital Umuganda

Digital Umuganda will provide the foundation required for the technical research and evaluation.

Digital Umuganda will be responsible for:

- Identifying and gathering relevant Rwandan materials;
- Selecting representative document samples;
- Digitizing or scanning physical materials where necessary;
- Preparing documents for evaluation;
- Processing and organizing the underlying data;
- Providing appropriate reference or ground-truth data required for evaluation;
- Defining the OCR technologies or technology categories to be investigated where appropriate;
- Providing access to previous Digital Umuganda work relevant to African-language digitization;
- Providing the required technical infrastructure and resources where available; and
- Providing technical supervision and guidance throughout the project.

This approach allows the researcher to concentrate primarily on the **technical evaluation and application of OCR technologies rather than data collection and preparation**.

Researcher

The researcher will focus on:

- Reviewing the provided document collection;
- Setting up and deploying selected OCR technologies;
- Developing the evaluation methodology;
- Running OCR experiments;
- Benchmarking different approaches;
- Analyzing recognition errors;
- Evaluating document structure and layout preservation;
- Investigating preprocessing and post-processing approaches;
- Developing the digitization workflow;
- Developing the prototype application;
- Documenting experiments and results; and
- Producing recommendations based on the findings.

5. Geographic and Language Focus

The primary focus will be **Rwandan materials**.

Language priority will be:

1. **Kinyarwanda**
2. **English**
3. **French**

The project will also consider multilingual documents containing combinations of these languages.

Kinyarwanda will receive particular attention because one of the primary objectives is to understand how well existing OCR technologies perform on African-language content compared with languages that have historically had substantially greater digital representation.

Although Rwanda will serve as the primary case study, the evaluation methodology should be designed so that it can subsequently be adapted to other African languages and document collections.

6. Document Categories

Digital Umuganda will provide representative documents across different categories and quality levels.

These may include:

Newspapers

Newspapers are particularly useful for testing OCR because they may contain complex layouts involving multiple columns, headlines, photographs, captions, advertisements, different font sizes, and multilingual content.

Books

Books may include literature, educational books, historical works, dictionaries, reference materials, and other publications.

Government and Institutional Documents

These may include reports, policy documents, administrative publications, statistical reports, research publications, and other institutional records.

Historical and Cultural Materials

Where available, the evaluation may include magazines, newsletters, historical publications, educational materials, and other archival resources.

The objective will be to test OCR technologies against a sufficiently diverse sample to understand how performance changes under realistic conditions.

7. OCR Technologies to Be Evaluated

The project will evaluate existing technologies rather than developing or training a new OCR model.

Potential technology categories may include:

- Traditional OCR engines;
- Open-source OCR systems;
- Cloud OCR services;
- Deep-learning-based OCR systems;
- Document understanding models;
- Vision-Language Models; and
- Multimodal Large Language Models.

One important research question will be whether newer AI-based approaches provide meaningful improvements over conventional OCR systems when processing African-language documents.

The comparison should consider not only recognition accuracy but also processing speed, computational requirements, cost, ease of deployment, scalability, and ability to understand complex document structures.

8. Evaluation Framework

The researcher will develop a standardized evaluation framework that allows different technologies to be compared using the same document samples.

Text Recognition

Core measurements may include:

- Character Error Rate;
- Word Error Rate;
- Missing text;

- Incorrect words;
- Incorrect characters;
- Word segmentation errors; and
- Language-specific errors.

Document Structure

The project will also examine preservation of:

- Reading order;
- Paragraphs;
- Columns;
- Headings;
- Captions;
- Tables;
- Lists; and
- Other structural information.

Operational Performance

Evaluation should additionally consider:

- Processing time;
- Hardware requirements;
- Cost;
- API dependency;
- Scalability;
- Ease of integration; and
- Batch-processing capabilities.

This will ensure that the project identifies not merely the most accurate system, but the most practical technologies for a future large-scale digitization programme.

9. Kinyarwanda-Specific Evaluation

A central component of the project will be understanding OCR performance specifically for **Kinyarwanda**.

The researcher will analyze recurring errors involving:

- Kinyarwanda vocabulary;
- Proper names;
- Historical spelling;

- Word boundaries;
- Capitalization;
- Punctuation;
- Character recognition; and
- Confusion with words from other languages.

The analysis should help identify whether errors are systematic and whether they can be corrected using preprocessing, post-processing, dictionaries, language models, or other existing technologies.

10. Preprocessing and Post-Processing

Using the documents and processed data provided by Digital Umuganda, the researcher will investigate how different technical processing approaches affect OCR quality.

Preprocessing experiments may include techniques such as:

- Deskewing;
- Image enhancement;
- Noise reduction;
- Contrast adjustment;
- Page segmentation; and
- Layout detection.

Post-processing may investigate:

- Language-aware correction;
- Dictionary-based correction;
- Rule-based correction;
- LLM-assisted correction;
- Formatting reconstruction; and
- Identification of potentially incorrect or low-confidence text.

The objective is not to train new OCR models but to determine the best **combination of existing technologies and processing techniques**.

11. Digitization Workflow

One of the project's main outputs will be a recommended technical workflow for Rwandan document digitization.

The overall process may follow:

Document Collection and Preparation (Digital Umuganda) → Scanning and Data Processing (Digital Umuganda) → Document Preprocessing → Layout Analysis →OCR / AI-Based Text Extraction → Post-Processing → Quality Evaluation → Human Validation where required →Structured Digital Output

The project will determine the most appropriate technologies and processes for the technical stages of this pipeline.

12. Prototype Application

The project should result in a simple **user-friendly digitization prototype** demonstrating how the recommended OCR workflow could eventually be operationalized.

A user should be able to provide a document and receive:

- Extracted text;
- Searchable content;
- Preserved document structure where possible;
- Confidence or quality indicators;
- Side-by-side comparison between the source and extracted content where useful; and
- Exportable machine-readable results.

Potential outputs may include searchable PDF, plain text, structured text, or JSON depending on the document and intended downstream use.

The prototype should demonstrate the practical application of the research rather than remaining solely an experimental benchmarking environment.

13. Project Duration

The initial project will run for **three months**.

A detailed workplan will be developed as one of the project deliverables.

The project may subsequently be extended depending on the findings of the initial phase.

An extension could explore:

- Larger document collections;
- Additional Rwandan archives;
- Improved automated correction;
- Human validation workflows;
- Searchable digital archives;
- Additional African languages;

- Production-scale processing pipelines; and
- Integration of digitized content with language models and other AI applications.

14. Expected Deliverables

The project is expected to produce:

1. **OCR Technology Assessment** – evaluation of selected OCR and document-understanding technologies.
2. **Evaluation Framework** – standardized methodology for assessing OCR accuracy, document structure, Kinyarwanda performance, efficiency, and operational considerations.
3. **Benchmark Results** – comparative results across OCR technologies and different categories of Rwandan materials.
4. **Kinyarwanda OCR Assessment** – detailed analysis of performance and recurring errors when processing Kinyarwanda.
5. **Processing Experiment Results** – analysis of how preprocessing and post-processing approaches affect OCR performance.
6. **Recommended Digitization Workflow** – a reproducible technical pipeline for processing Rwandan materials.
7. **Prototype Digitization Application** – a user-friendly demonstration of the recommended approach.
8. **Detailed Project Workplan** – developed as part of the project rather than prescribed in this concept note.
9. **Technical Documentation and Reusable Code** – enabling Digital Umuganda to reproduce and extend the experiments.
10. **Final Technical Report** – documenting methodology, experiments, findings, limitations, and recommendations.
11. **Scale-Up Recommendations** – recommendations for moving from experimentation toward larger-scale digitization of Rwandan materials and eventual expansion to other African-language collections.

15. Expected Outcomes

The project will establish practical evidence regarding the readiness of existing OCR and AI technologies for digitizing Rwandan materials.

It should determine:

- Which technologies perform best for Kinyarwanda;
- Which technologies perform best across different document categories;
- How document quality affects OCR performance;

- Whether modern Vision-Language Models provide meaningful improvements over traditional OCR;
- Which preprocessing and post-processing approaches provide measurable improvements;
- How much human verification may be necessary;
- What technical architecture is appropriate for large-scale processing; and
- What limitations must be addressed before scaling.

The project will also provide Digital Umuganda with a reusable evaluation framework and technical foundation for future digitization initiatives.

16. Strategic Importance

A significant portion of Rwanda's historical and contemporary knowledge was created before digital publishing became widespread. Newspapers, books, reports, educational materials, institutional publications, and historical archives contain valuable information about Rwanda's language, history, culture, institutions, and development.

Digital Umuganda's responsibility for **gathering and processing these materials** will provide the project with a structured data foundation, while the research component will establish how effectively modern OCR technologies can transform those materials into machine-readable knowledge.

This distinction allows the project to focus its research resources on the central technical question: **how well can current AI technologies digitize African-language knowledge?**

The resulting digital content can ultimately support searchable archives, educational applications, research, Machine Translation, Natural Language Processing, Retrieval-Augmented Generation, and locally grounded Large Language Models.

Starting with Rwanda provides a manageable environment in which to establish and validate the methodology. The longer-term opportunity is to develop a reusable approach that can contribute to the **large-scale digitization of African knowledge and African-language content for the AI era**.