SOKOINE UNIVERSITY OF AGRICULTURE
·1984·
ARDHI NI HAZINA

SIGIR 2025
Padova
ITALY

# Towards Enhanced Agricultural Information Access in Kiswahili

## Integrating Knowledge Graphs and Retrieval-Augmented Generation

Joseph P. Telemala, Neema N. Lyimo, Anna R. Kimaro, & Camilius A. Sanga

**Speaker:** Joseph P. Telemala

Sokoine University of Agriculture, Tanzania

Padua, Italy, 16 July 2025

# Content

1. Motivation
2. Research Objective
3. Methodology
4. Challenges & Open Questions
5. Next Steps

# The Problem Context

- Agriculture is Tanzania's backbone (livelihood and food security).
- Thus, access to agricultural information is vital for sustainable farming.
- Despite this, a language gap exists as most (agricultural) research is published in English, while consumers (mostly smallholder farmers and extension officers) primarily use Kiswahili.
- This disconnect limits the usability of agricultural knowledge, showing clearly the need for localized, Kiswahili-accessible content.

# The Mkulima Repository

- Part of the localized efforts is the Mkulima repository, developed by the Sokoine National Agricultural Library (SNAL).
- The repository currently hosts over 700 agricultural publications written in Kiswahili.
- However, as any other repository, its semi-structured format (PDFs) limits effective search and access leading to poor search & relevance, esp. in the error of generative AI.
- Users need quick, accurate, answers to field problems. Farmers do not need 1,000 paragraphs about "*viazi vitamu* - sweet potatoes" — they need specific, structured, explainable knowledge: "What to grow, how to grow it, and how to protect it."
- There is a clear need to either enhance the repository or develop an intelligent and user-friendly platform that utilizes the its content better information search and access.

# Research Objectives

- To utilize the Mkulima repository's content in developing a smart, conversational, platform using:

  Knowledge Graphs (KG)

  Retrieval-Augmented Generation (RAG)

- In order to empower farmers and agriculture stakeholders with AI-driven, natural language answers in Kiswahili.

# KG and RAG

- Imagine a user asks:

  "*Je, hali ya hewa huathiri magonjwa ya nyanya?*" – "Does the weather influence tomato diseases?"

- A RAG system might return unrelated paragraphs on *nyanya*, *hewa*, and *magonjwa*. But a KG can tell:

  *Magonjwa ya nyanya* → Fusarium wilt → *Inaathiriwa na* → *Unyevu mwingi*
  Tomato disease → Fusarium wilt → Is affected by → High moisture

- We believe, combining both means:

  RAG will fetch the narrative, whereas KG will give it structure and causality.
  RAG retrieves, KG reasons! e.g., RAG does not "know" how tomato, disease, and climate are connected, but KG does.
  English IR systems are flooded with models, data, and infrastructure. Kiswahili is not.

# Approach (Experimenting — RAG + KG)

- **Data Collection & Preparation**
  - Scrape Swahili agricultural documents from Mkulima repository and other open online sources
  - Clean, segment, and tokenize text

- **Knowledge Graph (KG) Construction**
  - Use NER and dependency parsing to extract entities and relations
  - Map results to a simple agriculture (domain-specific) ontology

- **Retrieval-Augmented Generation (RAG)**
  - Explore monolingual models (e.g., SwahBERT) for potential fine-tuning on agricultural text
  - Evaluate multilingual embeddings (e.g., LaBSE) for zero-shot retrieval in Kiswahili
  - Store document chunks in a vector database (e.g., FAISS) for fast semantic search

# Approach (Experimenting — RAG + KG)

- **Conversational Interface**

  Accept queries in Kiswahili

  Generate natural responses grounded in documents and KG

# Evaluation Strategy

- Compare performance of:
  - **RAG-only system** (document retrieval + generation)
  - **RAG + KG system** (document retrieval + generation augmented with KG facts)
- Participants
  - Real users, including extension officers, agriculture students, and agriculture experts
- Evaluation Criteria
  - Relevance, fluency, and accuracy of responses
  - Response clarity and usefulness
  - User satisfaction and preference
- Methodology
  - Controlled A/B testing on identical queries
  - Survey-based feedback (Likert scale and comments) and task success rate

# Challenges and Open Questions

## Challenge

- Existing RAG systems are optimized for English and general domains. Kiswahili and other low-resource languages lack annotated data, embeddings, and evaluation benchmarks.

## Open Questions

- How can Knowledge Graphs be effectively fused into RAG pipelines to improve not only retrieval, but the faithfulness and explainability of generated answers?

# Next Steps

- We believe information retrieval should work for everyone, in the languages they speak and in the domains they rely on.
- Our next step is to build this system, test it with the people it is meant for, and improve it with their input.
- If you are working on similar challenges, we would love to connect.