

Faculty of Digital Engineering  
Department of Computer Science  
Specialized in Data Science  
Generation 09

## **Capstone II Project Proposal**

Student's Name : Sang Haksou, Sovan Chandara, Chork Theasov,  
Oung Chhunheng, Chea Virakbott, Chhin Menghour  
Group : 20  
Project Title : Multi-Agent Text Extraction System  
Started Date : 30<sup>th</sup> September 2025  
Ended Date : 26<sup>th</sup> December 2025

### **I. Project Description**

Multi-Agent Text Extraction System is an advanced artificial intelligence system designed to automatically collect, process, and organize Khmer-language content from Cambodian digital sources. The system uses a smart multi-agent architecture where specialized AI agents collaborate to perform different tasks including web scraping, Khmer language validation, content categorization, and dataset organization. A central Manager Agent orchestrates the entire workflow, ensuring smooth operation from data acquisition to structured storage. The platform will be developed using Python with modern AI frameworks with a simple web dashboard for data exploration and export.

#### **1. Problem Statement**

The Cambodian research and development community faces significant challenges in accessing structured Khmer-language datasets, slowing down development in natural language processing, social science research, and AI innovation. Key challenges include:

- **Manual Data Collection:** Researchers and developers waste significant time manually gathering Khmer text from online sources
- **Lack of Categorization:** Existing data is not organized into meaningful categories, making it difficult to retrieve specific information
- **No Centralized Khmer Dataset Repository:** Researchers have no central place to find organized Khmer dataset

#### **2. Solution**

The proposed solution is to develop a **Multi-Agent System** that automates the entire data collection and organization process for Khmer-language news:

- **Automated Data Collection:** Specialized agents automatically gather Khmer content from Cambodian news sites, government portals, and educational platforms

- **Khmer-Specific Processing Pipeline:** Custom pipelines clean, validate, and categorize Khmer text using AI models trained for local language patterns
- **Easy Data Access:** A user-friendly platform lets researchers browse, filter, and download organized Khmer datasets in multiple formats

### 3. Benefits

This project provides several key benefits:

- **Researchers & Students:** Get access to ready-to-use Khmer datasets by saving time and effort for academic projects and data analysis
- **Developer & Tech Community:** Can build better Khmer language tools and AI applications using high-quality and organized data
- **Society & Culture:** Help to preserve and modernize the Khmer language in the digital age by supporting Cambodia's technological development

## II. Project Objective

### 1. Primary Objective

To design and implement an intelligent multi-agent system capable of autonomously collecting, processing, categorizing, and distributing high-quality Khmer-language text datasets from diverse Cambodian digital sources.

### 2. Specific Objectives

- **Build specialized AI agents** that work together to automatically collect, process, and organize Khmer text data
- **Create Khmer-optimized systems** for web scraping, text cleaning, and content categorization that understand local language patterns
- **Develop a user-friendly platform** where anyone can easily search, filter, and download Khmer datasets for their projects

## III. Methodology

The methodology of this project follows a structured, multi-phase process that combines **AI-driven automation**, **web scraping**, and **intelligent data organization**. The system is designed as a **multi-agent architecture**, where each AI agent performs a specialized task and collaborates with others to achieve the overall goal: collecting, cleaning, classifying, and delivering Khmer-language news data to users in an organized format.

### 1. Agile Development Approach

The project will follow an **iterative agile methodology** with 2-week sprints:**Sprint Planning:**

- **Sprint 1 (Weeks 1-2):** Core agent framework & basic scraping
- **Sprint 2 (Weeks 3-4):** Khmer NLP processing pipeline
- **Sprint 3 (Weeks 5-6):** Multi-agent orchestration & database integration
- **Sprint 4 (Weeks 7-8):** Web dashboard & API development

- **Sprint 5 (Weeks 9-10):** Testing, optimization & deployment
- **Sprint 6 (Weeks 11-12):** Documentation & final presentation

## 2. System Tools and Technologies

Category	Technologies
Programming Language	Python 3.9+
Agent Framework	LangGraph, LangChain, CrewAI
Web Framework	FastAPI, Streamlit
Database	PostgreSQL
Search Engine	Elasticsearch
Task Queue	Celery with Redis

## IV. Project Milestone

The project will be completed in several key phases, each with its own milestone. These milestones represent major steps in the development process, ensuring that the project stays on track and meets its goals on time. Each milestone is designed to deliver tangible progress and build toward the final working prototype of the AI system.

Project Timeline												
Task	W1	W2	W3	W4	W5	W6	W7	W8	W9	W10	W11	W12
Project Planning												
Requirement Analysis												
Literature Review												
Foundation & Setup												
Database Development												
Agent Development												
Web Interface												
Testing & Optimization												
Documentation												

## VI. Project Organization

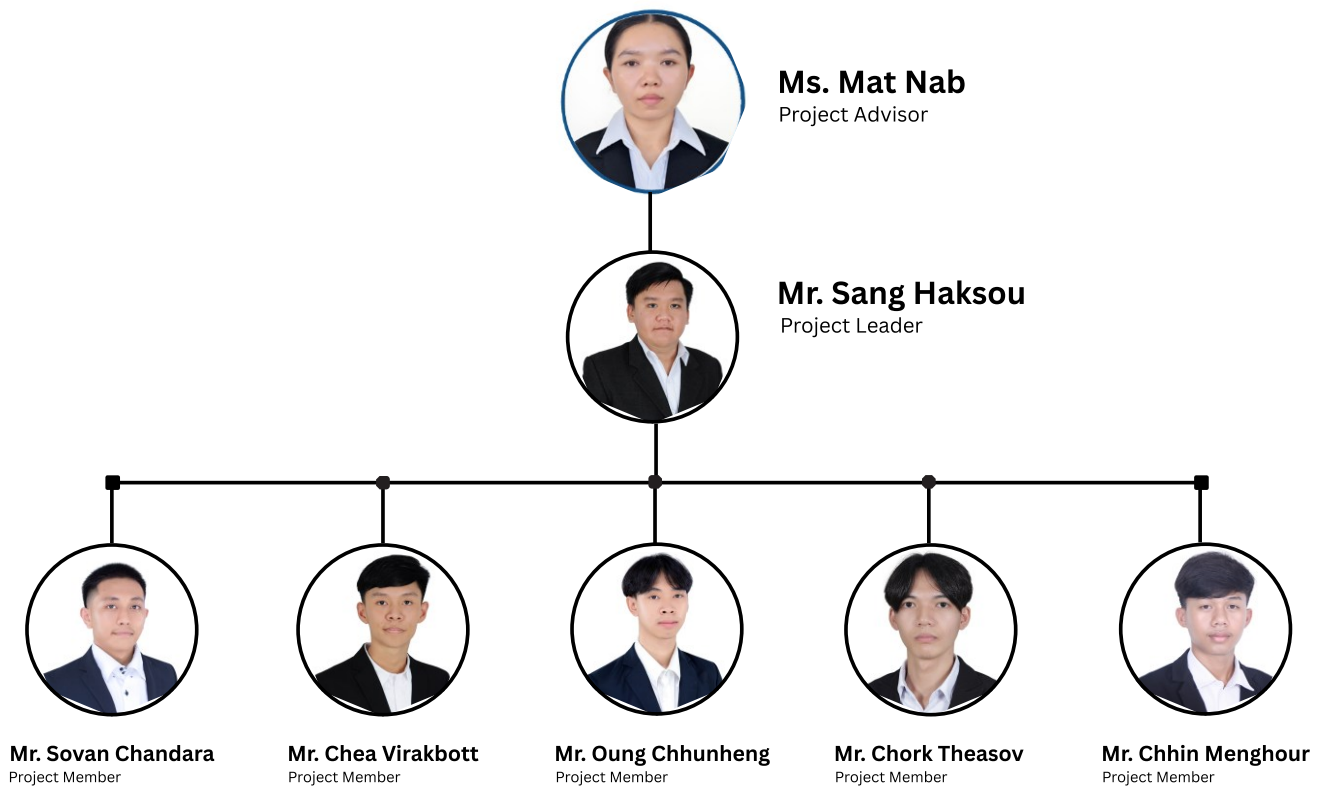
This project is carried out under the guidance of an experienced advisor who provides both technical and academic support to ensure successful development. The team consists of six members working together to build the Multi-Agent System.

- **Project Advisor:** Ms. Mat Nab
 

Provides both academic support and technical and also advice on project direction, methodology, and documentation to align the project with academic requirements.

- **Project Members:**

1. Mr. Sang Haksou: Project Lead,
2. Mr. Sovan Chandara: AI/ML Specialist
3. Mr. Chork Theasov: AI/ML Specialist
4. Mr. Chea Virakbott: Backend Developer
5. Mr. Oung Chhunheng: Database Architect
6. Mr. Chhin Menghour: Frontend Developer



Seen and Approved by Advisor

Ms. Mat Nab