Tạp chí Khoa học Đại học Cần Thơ
Số chuyên đề: Công nghệ chiến lược - Động lực phát triển Đồng bằng sông Cửu Long (SDMD)
ISSN 1859-2333 | e-ISSN 2815-5599

DOI:10.22144/ctujos.2026.251

XÂY DỰNG NGÂN HÀNG TRI THỨC THỰC VẬT HỌC PHÂN CẤP CHO GIỐNG CÂY TRỒNG DỰA TRÊN NỀN TẢNG MỞ WIKICROP

Trương Xuân Việt[1,2], Nguyễn Hữu Hòa[2], Nguyễn Đỗ Châu Giang[3], Huỳnh Quang Nghi[2], Trần Thị Minh Thảo[2], Lưu Trùng Dương[4,2] và Lưu Tiến Đạo[2,*]

[1]*Trung tâm Công nghệ Phần mềm, Đại học Cần Thơ, Việt Nam*
[2]*Trường Công nghệ Thông tin và Truyền thông, Đại học Cần Thơ, Việt Nam*
[3]*Trường Nông nghiệp, Đại học Cần Thơ, Việt Nam*
[4]*Trung tâm Chuyển đổi số và Truyền thông, Đại học Cần Thơ, Việt Nam*
**Tác giả liên hệ (Corresponding author): ltdao@ctu.edu.vn*

Thông tin chung (Article Information)

Nhận bài (Received): 09/08/2026
Sửa bài (Revised): 17/09/2026
Duyệt đăng (Accepted): 04/11/2026

Title: *Building a hierarchical botanical knowledge base for crop varieties on the WikiCrop open platform*

Author(s): *Truong Xuan Viet[1,2], Nguyen Huu Hoa[2], Nguyen Do Chau Giang[3], Huynh Quang Nghi[2], Tran Thi Minh Thao[2], Luu Trung Duong[4,2] and Luu Tien Dao[2,*]*

Affiliation(s): *[1]Cantho University Software Center, Can Tho University, Viet Nam; [2]College of Information and Communication Technology, Can Tho University, Viet Nam; [3]College of Agriculture, Can Tho University, Viet Nam; [4]Center for Digital Transformation and Communications, Can Tho University, Viet Nam*

TÓM TẮT

Bài báo được thực hiện nhằm giới thiệu nền tảng tri thức mở WikiCrop dựa trên lõi MediaWiki, giúp cấu trúc hóa dữ liệu phân cấp theo hệ thống phân loại thực vật học. Nhờ tích hợp cổng trí tuệ nhân tạo tại biên hỗ trợ mô hình ngôn ngữ lớn, nền tảng WikiLab, phân hệ phân tích dữ liệu thông minh và cơ chế kiểm duyệt đồng thuận từ hội đồng chuyên gia, hệ thống tự động hóa nhập liệu, hỗ trợ khai phá quy luật sinh học và bảo đảm độ tin cậy thông tin. Kết quả vận hành thực tế trên 300 giống cây trồng bước đầu cho thấy tính khả thi và sự ổn định, bất biến cấu trúc, đóng góp hạ tầng số thiết thực cho nông nghiệp Đồng bằng sông Cửu Long. Ngoài ra, ứng dụng di động WikiCrop đã được phát hành, cung cấp công cụ tra cứu trực tiếp và thuận tiện cho cộng đồng ngay tại thực địa.

Từ khóa: *Kiểm duyệt chuyên gia, MediaWiki, phân loại học thực vật, mô hình ngôn ngữ lớn, số hóa cây trồng, WikiCrop*

ABSTRACT

This paper introduces WikiCrop, an open knowledge platform powered by the MediaWiki core that structures agricultural data hierarchically based on the botanical classification system. By integrating an LLM-based edge artificial intelligence gateway, the WikiLab platform, an intelligent data analysis subsystem, and a consensus-based moderation mechanism by an Expert Panel, the system automates data ingestion, supports biological knowledge discovery, and ensures information reliability. Practical deployment on 300 crop varieties shows initial feasibility and structural stability, offering a pragmatic digital infrastructure for agriculture in the Mekong Delta. Furthermore, the WikiCrop mobile app has been released, providing a direct and convenient knowledge retrieval tool for the community in the field.

Keywords: *Botanical classification, crop digitalization, expert moderation, LLM, MediaWiki, WikiCrop*

1. GIỚI THIỆU

Giống cây trồng là yếu tố cốt lõi quyết định năng suất và an ninh lương thực, đặc biệt tại Đồng bằng sông Cửu Long (ĐBSCL) – vùng chiến lược đóng góp hơn 33% GDP nông nghiệp cả nước (Huyền, 2025). Tuy nhiên, nguồn thông tin khoa học về giống cây trồng, bao gồm đặc tính sinh học, kỹ thuật canh tác và các khả năng chống chịu thiên tai (mặn, hạn, ngập úng), hiện đang bị phân tán rải rác tại nhiều viện nghiên cứu, trung tâm giống và doanh nghiệp, gây khó khăn lớn cho sự phát triển chung của nền kinh tế nông nghiệp. Sự thiếu hụt một hệ thống dữ liệu đồng bộ không chỉ làm giảm hiệu quả ứng phó với biến đổi khí hậu theo từng tiểu vùng sinh thái, mà còn đối mặt với nguy cơ mai một nghiêm trọng các giống cây bản địa, đặc sản có giá trị cao về kinh tế và văn hóa do thiếu các chương trình bảo tồn phù hợp.

Minh chứng rõ nét nhất là cây lúa – ngành hàng chiếm hơn 90% kim ngạch xuất khẩu gạo cả nước (Tổng cục Thống kê, 2026) nhưng đang bị thu hẹp diện tích do đô thị hóa và thiên tai. Dù Đại học Cần Thơ đang lưu giữ hơn 1.900 mẫu lúa quý giá nhưng dữ liệu vẫn chủ yếu tồn tại ở dạng tệp cơ sở dữ liệu nội bộ và hồ sơ giấy khép kín, chưa thể khai thác nhanh chóng qua Internet. Các nghiên cứu số hóa nội địa trước đây thường gặp phải các hạn chế lớn: dữ liệu cục bộ thiếu tính thị trường và thông tin mang tính quá thiên về khoa học, ít chú trọng các yếu tố về kinh tế – xã hội. Ở phạm vi quốc tế, Viện Nghiên cứu Lúa gạo Quốc tế (International Rice Research Institute – IRRI) đã vận hành hiệu quả ngân hàng kiến thức trồng lúa và mở cổng chia sẻ hơn 3.000 bộ gen lúa toàn cầu (IRRI, as cited in Li et al., 2014).

Trước đây, Đạo và ctv. (2017) đã thực hiện số hóa cơ sở dữ liệu hoàn chỉnh về tất cả các giống lúa của ĐBSCL hiện đang được bảo quản tại Trường Nông nghiệp (thuộc Đại học Cần Thơ); sau đó, nhóm dữ liệu được gom để nhận diện sự tương đồng giữa các giống lúa. Tuy nhiên, nghiên cứu này cũng vướng phải một số giới hạn nhất định: (1) Dữ liệu mang tính khép kín trong các giống lúa của ngân hàng, thiếu các giống lúa được ưa chuộng trên thị trường (do thiếu cơ chế cập nhật thường xuyên), (2) nghiên cứu chỉ tập trung duy nhất vào cây lúa, thiếu các loại cây trồng khác, và (3) hệ thống cũng chỉ mang tính cung cấp thông tin một chiều và hiện tại cũng dừng ở mức học thuật, chưa đưa vào ứng dụng thực tiễn. Nhu cầu đưa các nguồn dữ liệu này vào thực tiễn, kết hợp với các nguồn dữ liệu giống cây trồng khác trở nên rất cần thiết.

Thông tư 28/2016/TT-BNNPTNT về Danh mục bổ sung giống cây trồng được phép sản xuất, kinh doanh tại Việt Nam được Bộ Nông nghiệp và Phát triển nông thôn (2016) công bố. Đây có thể được xem là một trong các cơ sở quan trọng để xác định các giống cây trồng có ý nghĩa lớn đối với kinh tế nông nghiệp tại Việt Nam. Về nhu cầu cơ bản, thông tin về các giống cây trồng này cũng cần được quản lý và chia sẻ trên các nền tảng dữ liệu mở, đảm bảo tính tương thích và đồng bộ với các tiêu chuẩn quốc tế như Dublin Core (Dublin Core Metadata Initiative, 2026) hoặc Data Catalog Vocabulary (World Wide Web Consortium, 2024) nhằm tối ưu hóa khả năng liên thông dữ liệu.

Trước yêu cầu cấp bách đó, sự ra đời của một nền tảng thông tin tin cậy về giống cây trồng, có tính kết nối là giải pháp căn cơ góp phần hạn chế các thông tin quảng cáo sai lệch trên Internet. Sự khác biệt và cần thiết của hệ thống này so với các hệ thống như Bách khoa toàn thư Wikipedia là khả năng chuẩn hóa dữ liệu sâu đến cấp độ giống (crop varieties) thay vì chỉ dừng lại ở cấp độ loài (species) – đây mới chính là thông tin cốt lõi mà nhà sản xuất và thị trường tiêu thụ tìm kiếm. Để đảm bảo tính khả thi và không bị dàn trải, nền tảng giới hạn phạm vi tập trung vào hai nhóm đối tượng then chốt: các giống cây trồng chủ lực đang phục vụ thiết thực cho nền kinh tế nông nghiệp hiện tại và các nhóm cây trồng quý hiếm thuộc Danh mục nguồn gen cần bảo tồn theo Quyết định 80/2005/QĐ-BNN của Bộ Nông nghiệp và Phát triển Nông thôn (2005b). Hơn nữa, bên cạnh thông tin về giống cây trồng, kỹ thuật canh tác liên quan đến từng giống cây cũng được cập nhật và chia sẻ, làm nền tảng tri thức quan trọng phục vụ công tác khuyến nông một cách thiết thực và hiệu quả.

Trong bối cảnh chuyển đổi số toàn diện ngành nông nghiệp, việc số hóa, lưu trữ và quản trị tri thức một cách khoa học về các giống cây trồng là hạ tầng thông tin cốt lõi, phục vụ đắc lực cho các giải pháp nông nghiệp chính xác và công tác bảo tồn đa dạng sinh học. Hiện nay, các nguồn tài liệu về đặc điểm hình thái thực vật học, chu kỳ sinh học và kỹ thuật canh tác tại Việt Nam đang tồn tại ở dạng phân tán, bất đồng bộ và thiếu tính kết nối. Phần lớn các cổng thông tin nông nghiệp hoặc trang tin khuyến nông hiện nay chỉ phân loại cây trồng một cách đại cương dựa trên giá trị công dụng kinh tế (như nhóm cây lương thực, nhóm cây ăn trái, cây công nghiệp), vô tình làm đứt gãy mối liên hệ tiến hóa, đặc tính sinh học và hình thái tự nhiên giữa các loài thực vật trong cùng một họ học thuật.

Để giải quyết những thách thức nêu trên, giải pháp xây dựng nền tảng bách khoa tri thức mở mang tên WikiCrop (được cài đặt tại địa chỉ https://wikicrop.ctu.edu.vn:8080) đã được công bố trong nghiên cứu. Cụ thể, những đóng góp chính của bài báo bao gồm: (1) đề xuất kiến trúc dữ liệu phân cấp sâu đến tận cấp độ giống cây trồng, (2) tích hợp công nghệ tự động hóa lai (Hybrid AI) để xử lý tri thức, (3) thiết lập cơ chế kiểm soát chất lượng dữ liệu nghiêm ngặt qua hội đồng chuyên gia và (4) phát triển hệ sinh thái ứng dụng thực tiễn với phân hệ phân tích trực tiếp trên máy khách và ứng dụng di động.

2. NGHIÊN CỨU LIÊN QUAN

Trong thực tế, nhiều quốc gia đã áp dụng công nghệ thông tin vào quản lý nông nghiệp nói chung và quản lý nguồn gen cây trồng và vật nuôi nói riêng. Viện IRRI (2026) đã xây dựng trang thông tin điện tử với tên gọi "Ngân hàng kiến thức trồng lúa" (với các phân hệ chính như sản xuất lúa gạo, sinh học cây lúa, kỹ thuật trồng lúa, quản lý dinh dưỡng, quản lý dịch hại, thu hoạch - bảo quản, kỹ thuật sản xuất hạt giống, các khóa học trực tuyến về lúa, bác sĩ lúa và các sản phẩm từ lúa gạo). IRRI cũng đã công bố cơ sở dữ liệu truy cập mở GigaDB chứa 3.000 bộ gen của các giống lúa đến từ 89 quốc gia (IRRI, as cited in Li et al., 2014). Do đó, việc xây dựng ngân hàng thông tin giống lúa vùng ĐBSCL là vô cùng cấp bách, góp phần giúp việc ứng dụng công nghệ thông tin vào sản xuất nông nghiệp của Việt Nam tiếp cận tiến bộ khoa học kỹ thuật của thế giới.

Tại Việt Nam, các công trình khảo cứu chuyên sâu về giống lúa đã được quan tâm từ sớm, tiêu biểu như tài liệu hệ thống hóa các giống lúa Việt Nam của Đáp (1980). Nối tiếp nền tảng đó, các kết quả chọn tạo của Viện Lúa Đồng bằng Sông Cửu Long – đặc biệt là nhóm giống lúa OM mang đặc tính năng suất cao và chống chịu điều kiện bất lợi – thường xuyên được công bố và cập nhật trên Tạp chí Omonrice (Tran & Pham, 2024; Nguyen et al., 2024). Bên cạnh đó, các công bố chuyên sâu về chọn lọc giống cây trồng chống chịu mặn hoặc kháng rầy cũng thường xuyên được đăng tải trên Tạp chí Khoa học Đại học Cần Thơ (Liên và ctv., 2012; Thanh và ctv., 2013; Thái & Dùng, 2013). Về phía cơ quan quản lý nhà nước, Bộ Nông nghiệp và Phát triển Nông thôn (2005a) đã chủ trì xuất bản nhiều tài liệu giá trị, tiêu biểu là ấn phẩm "575 Giống cây trồng nông nghiệp mới". Trong khi đó, Viện Nghiên cứu Cây ăn quả Miền Nam (2026) lại tiếp cận theo hướng cập nhật định kỳ các bản tin, thông tin giống và tài liệu kỹ thuật canh tác trực tiếp trên cổng thông tin điện tử của Viện. Có thể nói, nhu cầu về thông tin giống cây trồng và kỹ thuật canh tác liên quan là rất lớn, nhận được mối quan tâm của các cơ quan chức năng và các tổ chức chuyên môn khác nhau; tuy nhiên, việc hệ thống hóa giống cây theo phân loại chuẩn mực vẫn còn thiếu hụt hoặc chưa có tính mở. Đây chính là động lực lớn của nhóm tác giả trong việc xây dựng nền tảng tri thức mở WikiCrop dựa trên lõi MediaWiki, giúp cấu trúc hóa dữ liệu phân cấp theo hệ thống phân loại thực vật học.

Trong những năm gần đây, việc số hóa tri thức nông nghiệp và quản trị dữ liệu giống cây trồng đã thu hút sự quan tâm lớn của cả cộng đồng công nghệ thông tin lẫn khoa học nông nghiệp. Các hướng tiếp cận hiện tại có thể chia thành ba nhóm chính:

▪ Hệ thống cơ sở dữ liệu nông học truyền thống và đồ thị tri thức:

Các cơ sở dữ liệu lớn trên thế giới như AGROVOC của Food and Agriculture Organization of the United Nations (2026) hay Dbpedia (Kontokostas et al., 2012) đã xây dựng các bộ từ vựng thực thể học (ontology) để chuẩn hóa thuật ngữ nông nghiệp. Tại Việt Nam, các nghiên cứu của các viện, trường thường sử dụng cơ sở dữ liệu quan hệ để lưu trữ thông tin giống. Các hệ thống này thường mang tính một chiều, khép kín, khó đóng góp cho cộng đồng và thiếu tính tương tác thời gian thực. Việc truy vấn ngữ nghĩa đòi hỏi tài nguyên tính toán lớn và cấu trúc dữ liệu rất phức tạp, khó tiếp cận đối với người dùng phổ thông hoặc cán bộ khuyến nông tại địa phương.

▪ Các nền tảng bách khoa toàn thư mở dựa trên wiki:

Các dự án thuộc hệ sinh thái Wikimedia (như Wikipedia hay Wikidata) là những minh chứng thành công nhất về việc xây dựng ngân hàng tri thức mở từ sự đồng thuận của cộng đồng. Bằng cách sử dụng lõi MediaWiki, hệ thống cho phép liên kết dữ liệu đa chiều. Như đã phân tích, Wikipedia chỉ dừng lại ở việc định danh và phân loại sinh học đến cấp độ Loài. Trong sản xuất nông nghiệp thực tế, các thuộc tính thương mại, kỹ thuật canh tác đặc thù và các đặc tính nông học (như tính kháng mặn, kháng sâu bệnh) chỉ xuất hiện rõ ràng ở cấp độ Giống. Do đó, các nền tảng wiki phổ thông chưa thể đáp ứng được nhu cầu chuyên sâu của chuỗi cung ứng nông nghiệp.

Bảng 1. Đối sánh các đặc trưng cốt lõi của WikiCrop và các hệ thống liên quan

Nền tảng	Cấp độ quản lý dữ liệu	Cơ chế kiểm duyệt và cập nhật	Mục tiêu và định hướng ứng dụng
Wikipedia / Wikidata	Dừng ở định danh sinh học cấp độ Loài.	Cộng đồng mở	Bách khoa toàn thư chung, thiên về kiến thức bách khoa đa lĩnh vực.
AGROVOC / AgroPortal	Tập trung vào từ vựng, thuật ngữ nông nghiệp (Ontology)	Quản trị khép kín bởi tổ chức/chuyên gia	Chuẩn hóa ngữ nghĩa nông nghiệp chủ yếu phục vụ giới hàn lâm.
Genesys	Dữ liệu nguồn gen sinh học	Nguồn đóng góp từ các Viện/Ngân hàng gen quốc tế.	Phục vụ nghiên cứu bảo tồn đa dạng sinh học toàn cầu.
WikiCrop (Đề xuất)	Sâu đến cấp độ Giống	Cơ chế Lai (Hybrid): Trí tuệ nhân tạo (LLM + Exact Match) kết hợp hội đồng chuyên gia.	Phục vụ thực tiễn sản xuất, khuyến nông trực tiếp qua nền tảng Web/Mobile.

▪ Tự động hóa nhập liệu và chuẩn hóa dữ liệu bằng trí tuệ nhân tạo (TTNT)

Trong các nghiên cứu phân tích dữ liệu văn bản gần đây, việc sử dụng các công cụ tiền xử lý tự động như WikiExtractor (Attardi, 2015) đã trở thành một phương pháp phổ biến để phân tách các tệp dữ liệu XML Dump từ Wikipedia. Việc loại bỏ các thẻ định dạng phức tạp giúp trích xuất văn bản thuần túy, tạo thuận lợi cho quá trình chắt lọc và phân đoạn tập dữ liệu lớn. Đồng thời, với sự bùng nổ của các LLMs như GPT-4 (OpenAI) hay Llama (Meta), nhiều công trình đã ứng dụng TTNT để trích xuất thông tin từ các tài liệu phi cấu trúc trong y học và nông nghiệp. Có thể nói, thách thức lớn nhất của các phương pháp tự động hóa bằng LLM là hiện tượng "ảo giác" (hallucination), sinh ra các thông số kỹ thuật sai lệch hoặc không có thực. Nếu các dữ liệu này được nạp trực tiếp vào ngân hàng tri thức nông học mà không có bộ lọc kiểm duyệt thì các hệ thống sẽ mất đi tính độ tin cậy khoa học thiết yếu.

Khác với các nghiên cứu trước đây, WikiCrop khắc phục phần lớn các hạn chế trên bằng cách mở rộng chiều sâu dữ liệu thông qua hai cải tiến cốt lõi. Thứ nhất, hệ thống kế thừa cấu trúc phân loại học quốc tế của MediaWiki nhưng định nghĩa mới thẻ cấu trúc {{InfoCropPlant}} nhằm quản trị dữ liệu chi tiết đến tận cấp độ Giống. Thứ hai, nền tảng áp dụng cơ chế lai giữa TTNT và chuyên gia (Hybrid AI – Expert), sử dụng các LLMs (Manning, 2022) có trong phân hệ WikiLab để tự động hóa phần lớn quy trình bóc tách và chuẩn hóa dữ liệu thô từ các tài liệu phi cấu trúc, giúp giảm thiểu đáng kể khối lượng công việc nhập liệu thủ công trước khi chuyển sang bước kiểm duyệt của hội đồng chuyên gia.

Tóm lại, mặc dù các hệ thống nêu trên đã đóng góp đáng kể cho việc chuẩn hóa tri thức nông nghiệp và quản lý thông tin cây trồng, WikiCrop hướng đến mục tiêu ứng dụng khác biệt so với các nền tảng hiện hữu như Wikipedia/Wikidata, AGROVOC, AgroPortal hay Genesys. Bảng 1 tóm tắt đối sánh trực quan các đặc tính cốt lõi của WikiCrop so với các hệ thống này. Chi tiết sự khác biệt được phân tích cụ thể trên ba phương diện sau:

Thứ nhất, về độ sâu quản lý dữ liệu cây trồng: Các bách khoa mở như Wikipedia hay Wikidata chủ yếu dừng lại ở việc định danh phân loại đến cấp độ Loài (Species), trong khi các hệ thống từ vựng như AGROVOC hay AgroPortal tập trung vào việc chuẩn hóa bộ thuật ngữ chuyên ngành chung. Ngược lại, WikiCrop định nghĩa cấu trúc dữ liệu phân cấp sâu đến tận cấp độ Giống cây trồng (Crop Variety) thông qua thẻ {{InfoCropPlant}}, đáp ứng trực tiếp nhu cầu tra cứu đặc tính sinh học, kỹ thuật canh tác và thông tin thương mại trong sản xuất nông nghiệp.

Thứ hai, về cơ chế kiểm duyệt dữ liệu: WikiCrop kế thừa cơ chế biên tập cộng đồng của Wikipedia nhằm phát huy tính đóng góp mở nhưng khắc phục được điểm yếu về nhiễu thông tin. Đối với các dữ liệu được tự động hóa trích lọc bằng TTNT, WikiCrop thiết lập cơ chế kiểm duyệt lai chuyên sâu (Hybrid AI & Expert): LLM được ứng dụng để bóc tách dữ liệu thô, kết hợp bộ lọc Exact Match để tự động khóa cứng các thông số định lượng nông học nhạy cảm và phân quyền phê duyệt cuối cùng cho hội đồng chuyên gia theo đúng từng họ thực vật chuyên ngành.

Thứ ba, về định hướng ứng dụng thực tiễn: Khác với cơ sở dữ liệu Genesys vốn tập trung vào quản lý ngân hàng gen toàn cầu phục vụ mục tiêu bảo tồn đa

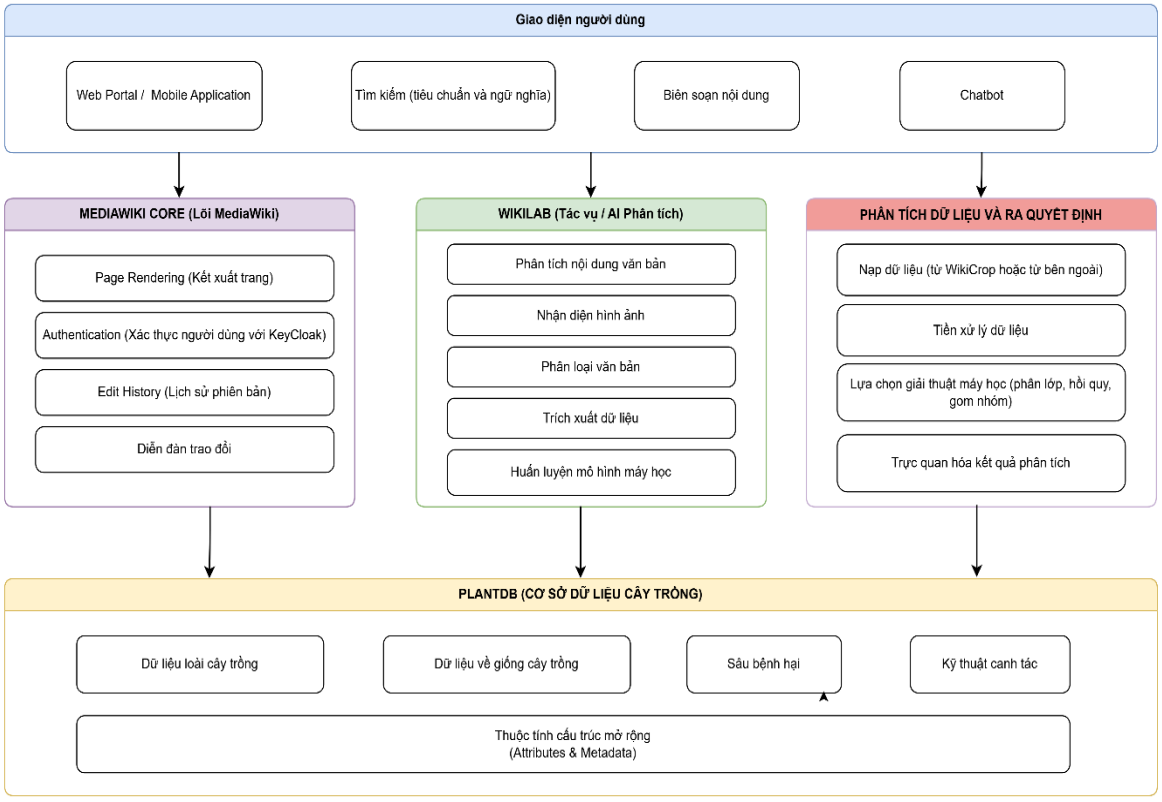
dạng sinh học cho giới học thuật, WikiCrop tập trung ưu tiên số hóa các giống cây trồng đang lưu thông phổ biến trên thị trường. Mục tiêu cốt lõi của nền tảng là cung cấp tri thức nông học chuẩn hóa, dễ tiếp cận cho cộng đồng sản xuất, người dân và cán bộ khuyến nông tại thực địa thông qua nền tảng Web và ứng dụng di động.

3. KIẾN TRÚC ĐỀ XUẤT CỦA NỀN TẢNG WIKICROP

Từ những phân tích về mặt hạn chế của các nền tảng lưu trữ hiện hữu, yêu cầu cấp thiết đặt ra là phải thiết kế một kiến trúc phần mềm đồng bộ và linh hoạt. Kiến trúc này không chỉ phải giải quyết được bài toán tổ chức cơ sở dữ liệu phân cấp sâu đến tận cấp độ giống cây trồng, mà còn phải tạo ra không gian tương tác mượt mà giữa các thuật toán trí tuệ nhân tạo (hỗ trợ nhập liệu) và hệ thống kiểm duyệt của con người (hội đồng chuyên gia). Để hiện thực hóa các định hướng lý thuyết đó, mô hình kỹ thuật của WikiCrop đã được nghiên cứu, thiết kế và triển khai. Hình 1 là kiến trúc tổng thể của hệ thống WikiCrop, được phân chia thành 05 thành phần chính:

(1) Thành phần giao diện người dùng (User Interface):

Giao diện người dùng là điểm chạm trực tiếp với người dùng, thành phần này đóng vai trò điều phối và thiết lập bộ khung cấu trúc cho toàn bộ cơ sở dữ liệu thực vật học mở. Các tính năng giao tiếp cốt lõi được xây dựng xoay quanh bốn phân hệ: Web Portal/Mobile App (cổng thông tin đa nền tảng), tìm kiếm (tiêu chuẩn và ngữ nghĩa), bộ công cụ biên soạn nội dung và Chatbot. Bên cạnh các tính năng giao tiếp cốt lõi, thành phần giao diện còn được tích hợp công cụ phân tích hành vi người dùng (như Microsoft Clarity (Microsoft, 2026)) nhằm theo dõi các luồng tương tác, từ đó liên tục tối ưu hóa trải nghiệm (UX/UI) cho cộng đồng khi tra cứu thông tin giống cây trồng.



Hình 1. Kiến trúc tổng thể của bách khoa cây trồng mở WikiCrop

(2) Thành phần lõi MediaWiki:

MediaWiki Core (MediaWiki, 2026) là nền tảng cốt lõi của Wikipedia, đại diện cho một hệ sinh thái tri thức mở toàn cầu được vận hành nhờ sự giao thoa giữa khoa học dữ liệu và công nghệ cộng tác phân tán. Trong kiến trúc của WikiCrop, các thành phần chính được kế thừa từ lõi này bao gồm: cơ chế kết xuất trang (hiển thị nội dung theo định dạng Wikitext), phân hệ xác thực người dùng (tích hợp

nền tảng KeyCloak (Keycloak Community, 2026)), hệ thống lưu trữ lịch sử phiên bản chỉnh sửa, và không gian diễn đàn trao đổi chuyên môn.

(3) Thành phần WikiLab (Tác vụ/AI Phân tích):

Được phát triển trên môi trường Jupyter Lab, thành phần này đóng vai trò là một hệ thống xử lý dữ liệu tự động khai thác sức mạnh của nhiều mô hình học máy khác nhau. Nòng cốt của phân hệ là các mô hình ngôn ngữ lớn (LLM), có khả năng tùy biến phương thức thực thi: hoặc xử lý cục bộ tại biên (thông qua Ollama) để bảo mật, hoặc xử lý trên đám mây (qua Gemini, ChatGPT) để tối ưu hiệu năng. Sự ra đời của phân hệ nhằm đáp ứng hiệu quả hai mục tiêu chiến lược trong chu trình quản trị tri thức:

Thứ nhất là *khai thác và chuẩn hóa tri thức ngoại biên* (External Knowledge Ingestion): WikiLab cung cấp các công cụ và thư viện chuyên dụng giúp thu thập, phân tích và xử lý các nguồn thông tin thô, phi cấu trúc về cây trồng từ các cơ sở dữ liệu hoặc tài liệu bên ngoài. Dữ liệu sau khi được bóc tách và chuẩn hóa thành cấu trúc Botany chuẩn hóa sẽ được chuẩn bị sẵn sàng để tích hợp vào lõi WikiCrop thông qua một cơ chế kiểm duyệt nghiêm ngặt và thực hiện quy trình trích dẫn nguồn gốc tài liệu một cách minh bạch, học thuật.

Thứ hai là *phân tích nội bộ và cơ chế tự học tiến hóa* (Internal Analysis & Self-Learning Loop): Phân hệ cho phép thực hiện các truy vấn ngữ nghĩa sâu trực tiếp vào cơ sở dữ liệu hiện hành của WikiCrop. Bằng cách phân tích mối liên hệ giữa các thuộc tính, các điểm đứt gãy thông tin sinh học hoặc xu hướng nông học tiềm ẩn đã được phát hiện, WikiLab rút trích ra các tri thức hữu ích mới. Những tri thức nâng cao này sau đó được tái cập nhật trở lại hệ thống lõi theo mô hình vòng lặp phản hồi, tạo nên một cơ chế tự học và tự tiến hóa liên tục cho ngân hàng dữ liệu WikiCrop mà không làm xáo trộn môi trường công cộng.

Cụ thể, chức năng tự động hóa nhập liệu bằng mô hình LLM trên WikiLab được thực thi chặt chẽ qua quy trình 3 bước:

(1) Đầu vào: Tiếp nhận các tài liệu nông học phi cấu trúc (như báo cáo khuyến nông, bài báo khoa học) hoặc văn bản thô mô tả đặc tính giống cây trồng;

(2) Xử lý bóc tách: LLM sử dụng các lời nhắc (prompt) chứa khung cấu trúc chuẩn (như thẻ {{InfoCropPlant}}) để tự động phân tích và chuyển đổi văn bản thô thành các cặp khóa - giá trị (key-value) đã được chuẩn hóa;

(3) Kiểm duyệt: Dữ liệu đầu ra của LLM bắt buộc phải đi qua bộ lọc cơ học (Exact Match) nhằm xác thực và khóa cứng các số liệu định lượng nhạy cảm; sau đó, dữ liệu được chuyển sang trạng thái chờ để hội đồng chuyên gia đối chiếu, thẩm định và phê duyệt trước khi tích hợp chính thức vào cơ sở dữ liệu lõi của WikiCrop.

(4) Thành phần PlantDB (Cơ sở dữ liệu Cây trồng):

Thành phần này chính là mục tiêu của nghiên cứu, và đóng vai trò nền tảng của toàn hệ thống. Dữ liệu cây trồng được lưu trữ chuyên sâu đến cấp độ Giống cây (sâu hơn một mức độ của Wikipedia, chỉ dừng ở mức độ Loài). Hơn nữa, dữ liệu liên quan đến kỹ thuật canh tác, các loài sâu bệnh gây hại cũng được thu thập và chia sẻ cho cộng đồng. Điểm cốt lõi giúp phân biệt hệ thống WikiCrop với các trang bách khoa toàn thư mở thông thường là việc cấu trúc hóa cây thư mục thực vật học thông qua thẻ cấu trúc `{{InfoPlant1}}` để biểu diễn Loài trên lõi MediaWiki. Một ví dụ cụ thể khi áp dụng cho loài lúa (*Oryza glaberrima*, *Oryza sativa*):

```
{{InfoPlant1
| TenCayTrong = Lúa
| Hinh = Lúa.png
| CacLoai =
<ul>
<li><i>Oryza glaberrima</i></li>
<li><i>Oryza sativa</i></li>
</ul>
|MoTaHinh=Cánh đồng lúa|Gioi=Plantae|KhongPhanLoai1=Angiospermae|KhongPhanLoai2=Monocots|KhongPhanLoai3=Commelinids|Bo=Poales|Ho=Poaceae|Chi=Oryza}}
```

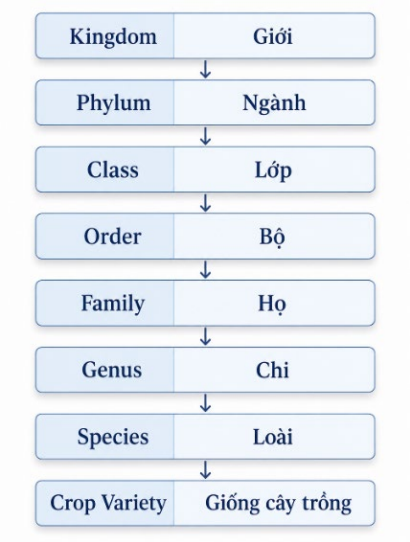
Quy trình nhập liệu trên hệ thống bắt buộc phải liên kết và gò ép chặt chẽ theo cấu trúc phân cấp phân loại học chuẩn quốc tế. WikiCrop định nghĩa mới một thẻ cấu trúc `{{InfoCropPlant}}` dựa trên `{{InfoPlant1}}` để bổ sung chiều sâu thông tin đến cấp độ giống với ba thuộc tính mới "TenGiong" (Tên giống), "LoaiGiong" (Loại giống: Dùng để phân biệt giống thuần chủng hoặc lai tạo)

và "XuatXu" (Xuất xứ). Một mẫu thông tin tiêu biểu cho các trường thông tin của {{InfoCropPlant}} như sau:

```
TenGiong = ST25 |
LoaiGiong=Lai tạo | XuatXu = Sóc
Trăng, Việt Nam
```

Để đảm bảo tính hệ thống và chặt chẽ trong quản trị dữ liệu giống cây trồng, điều quan trọng và mang tính cốt lõi nhất là căn cứ khoa học của phân loại. Trên thực tế, việc phân loại học chưa bao giờ là một vấn đề đơn giản và phân loại giống cây trồng cũng không phải ngoại lệ. Chẳng hạn, EOS (Earth Observation System) đề xuất giải pháp phân loại thành 06 nhóm: thực phẩm (Food), thức ăn gia súc (Forage), chất xơ (Fiber), dầu (Oil), cây cảnh (Ornamental) và công nghiệp (Industrial) (EOS Data Analytics, 2026). Tổ chức Lương thực thế giới (FAO) lại đề xuất một danh mục giống cây trồng quốc tế ICC (Indicative Crop Classification phiên bản 1.0) được phân loại thành 09 nhóm (Food and Agriculture Organization of the United Nations, 2005): ngũ cốc (Cereals), rau và dưa (Vegetables and melons), trái cây và các loại hạt (Fruit and nuts), cây lấy dầu (Oilseed crops), cây củ/rễ có hàm lượng tinh bột hoặc inulin cao (Root/tuber crops with high starch or inulin content), cây dùng làm đồ uống và gia vị (Beverage and spice crops), cây họ đậu (Leguminous crops), cây lấy đường (Sugar crops), các loại cây trồng khác (Other crops). Theo Sheaffer and Moncada (2012), khái niệm *giống cây nông nghiệp* (Agronomic crops) cũng được chia theo mục đích sử dụng thành 13 nhóm: cây ngũ cốc (Cereal Crops), cây lấy hạt (Grain Crops), cây họ đậu (Pulses), cây thức ăn gia súc (Forages), cây lấy sợi (Fiber Crops), cây lấy củ và rễ (Root and Tuber Crops), cây che phủ đất (Cover Crops), cây trồng xen canh (Companion Crops), cây cỏ phục vụ giải trí (Recreational Turf Crops), cây mía đường (Sugar Crops), cây lấy dầu (Oil Crops), cây lấy hạt giàu tinh bột (Starchy Seed Crops), cây lấy củ và rễ giàu tinh bột (Starchy Root and Tuber Crops). Đây là phân loại chi tiết dựa trên cả đặc tính thực vật học lẫn mục đích sử dụng.

Trong Bảng 2, chúng tôi đề xuất bám theo hệ thống *phân loại nông học dựa trên quan hệ họ hàng thực vật* để sử dụng cho WikiCrop, thực chất là sự kết hợp giữa hai nhóm phân loại cơ sở: (1) phân loại học thực vật (Botanical Classification (Sosef et al., 2020)) và (2) công năng kinh tế và mục đích sử dụng (Agronomic / Economic Classification). Trong đó, nhóm *cây ăn quả nhiệt đới* được đưa vào một phần dựa trên phân loại của ICC và vị trí địa lý vùng nhiệt đới của Việt Nam.



Hình 2. Phân loại học cây trồng theo phân loại học quốc tế (Botanical Classification (Sosef et al., 2020))

Bảng 2. Danh mục nhóm loài cây trồng trên hệ thống WikiCrop

STT	Nhóm loài cây trồng	Tên tiếng Anh
1	Ngũ cốc	Cereal Crops
2	Các loại kê	Millet Crops
3	Cây họ đậu	Pulse Crops
4	Cây dầu ăn	Edible Oilseed Crops
5	Cây dầu không ăn được	Non-edible Oilseed Crops
6	Cây lấy sợi	Fiber Crops
7	Cây thức ăn chăn nuôi	Forage Crops
8	Cây lấy đường	Sugar Crops
9	Cây củ	Tuber Crops
10	Cây kích thích	Stimulate Crops
11	Cây dược liệu	Medicinal Crops
12	Cây hương liệu	Aromatic Crops
13	Cây ăn quả nhiệt đới	Tropical fruit crops

▪ Cấp độ loài – đơn vị phân loại cơ bản của tự nhiên:

Trong phân loại sinh vật, Loài là bậc phân loại cơ bản và là một đơn vị tự nhiên tiến hóa. Một loài tập hợp các cá thể có các đặc điểm hình thái, giải phẫu và sinh lý tương đồng, có khả năng giao phối với nhau trong tự nhiên để sinh ra thế hệ sau có khả năng sinh sản và bị cô lập sinh sản với các nhóm sinh vật khác. Về tính chất, loài mang tính ổn định tự nhiên và đại diện cho một nhánh trên cây tiến hóa sinh học. Tên loài được định danh theo danh pháp

kép quốc tế (chẳng hạn, loài lúa là Oryza sativa, loài ngô là Zea mays).

Hình 3 cho biết loài lúa chia thành hai nhánh riêng biệt *Oryza glaberrima* và *Oryza sativa.* Thông tin ở cấp độ Loài cũng được Wikipedia quản lý và cũng được xem là cấp độ cuối cùng trong nhiều hệ thống phân loại sinh học. Tuy nhiên, như đã phân tích, thông tin này chưa đủ để phục vụ hoạt động kinh tế nông nghiệp, khi mà đơn vị quan tâm của họ là các giống cây trồng.



Hình 3. Bảng thông tin phân loại học về loài lúa

▪ Cấp độ Giống cây trồng – đơn vị cốt lõi của sản xuất và thương mại:

Giống cây trồng là một cấp phân loại dưới loài, là sản phẩm của quá trình chọn lọc tự nhiên lâu đời hoặc do con người chủ động chọn tạo, lai tạo và áp dụng công nghệ sinh học. Theo Luật Trồng trọt Việt Nam (Luật số 31/2018/QH14) (Quốc hội, 2018) và tiêu chuẩn quốc tế của International Union for the Protection of New Varieties of Plants (1991), một quần thể cây trồng chỉ được công nhận là một "Giống" khi đạt được 3 tiêu chí bắt buộc (tại Việt Nam được quy định tại Nghị định số 79/2023/NĐ-CP (Chính phủ, 2023)). Cụ thể, quần thể đó phải sở hữu *tính khác biệt* với ít nhất một đặc điểm hình thái hoặc nông học rõ ràng so với các giống đã biết; bảo đảm *tính đồng nhất* giữa các cá thể về những đặc tính cốt lõi (trừ các biến dị cho phép khi nhân giống); đồng thời duy trì được *tính ổn định*, đảm bảo các đặc tính này không bị thay đổi sau các chu kỳ nhân giống tiếp theo.

Về tính chất, Giống là đơn vị mang giá trị kinh tế trực tiếp. Tên giống thường là tên thương mại hoặc mã số do cơ quan chọn tạo đặt (chẳng hạn, Lúa ST25, Lúa OM18, Xoài cát Hòa Lộc). Hình 4 chỉ ra Giống ST25 là một biến thể của Loài Oryza sativa – một trong hai loài của cây Lúa. Đây cũng là giống lai tạo và có xuất xứ từ Sóc Trăng, Việt Nam.

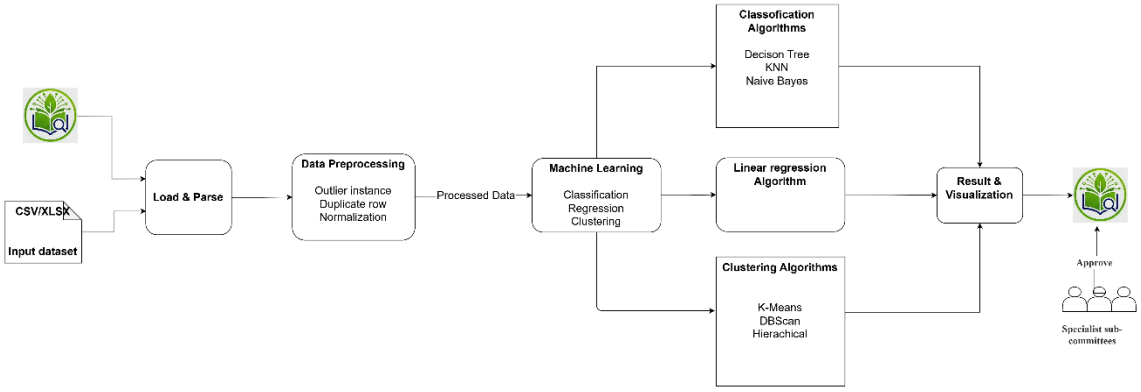


Hình 4. Bảng thông tin phân loại học về giống lúa ST25, thuộc loài lúa

(5) Thành phần phân hệ phân tích dữ liệu thông minh:

Thành phần phân hệ phân tích dữ liệu thông minh được thiết kế dưới dạng một công cụ hỗ trợ ra quyết định không cần viết các dòng mã lệnh, tích hợp vào mã nguồn của hệ thống thông qua cơ chế móc nối tự động và kế thừa lớp trang chuyên biệt. Thành phần này vận hành độc lập dưới dạng một ứng dụng đơn trang chạy hoàn toàn ở phía máy khách bằng ngôn ngữ JavaScript nhằm giảm thiểu tối đa rủi ro rò rỉ dữ liệu chọn giống nhạy cảm và giảm tải tối đa cho máy chủ. Để tối ưu hóa tính linh hoạt trong nghiên cứu thực tiễn, phân hệ được thiết kế để tiếp nhận dữ liệu đầu vào từ hai nguồn độc lập: Hệ thống cho phép truy xuất trực tiếp các tập dữ liệu giống cây trồng sẵn có trên nền tảng WikiCrop, hoặc cho phép các nhà khoa học tải lên các tập dữ liệu khảo nghiệm riêng biệt của cá nhân (dưới dạng tệp

tin bảng tính Excel hoặc các tập tin CSV) để tiến hành khai phá nội bộ mà không cần công bố ra công cộng. Phân hệ được thiết kế để vận hành bám sát theo quy trình khám phá tri thức từ cơ sở dữ liệu (Knowledge Discovery in Databases - KDD), bao gồm bốn giai đoạn tích hợp chặt chẽ (Hình 5):



Hình 5. Sơ đồ quy trình khám phá tri thức từ cơ sở dữ liệu áp dụng trong phân hệ phân tích dữ liệu thông minh

▪ **Nạp dữ liệu:** Tiếp nhận các tập dữ liệu từ WikiCrop hoặc tệp tin cá nhân của nhà nghiên cứu, tự động quét trùng lặp và chuyển đổi chúng thành dạng véc-tơ số học để chuẩn bị cho quá trình khai phá.

▪ **Tiền xử lý dữ liệu**: Tự động làm sạch dữ liệu, cảnh báo ngoại lai, xử lý giá trị bị thiếu bằng phương pháp nội suy trung bình hoặc loại bỏ; nhận diện và loại bỏ giá trị ngoại lai bằng phương pháp hàng rào Tukey dựa trên khoảng tứ phân vị kết hợp kỹ thuật giới hạn biên; chuẩn hóa các thuộc tính về cùng một thang đo bằng các thuật toán chuẩn hóa khoảng biến thiên, tiêu chuẩn hóa điểm số (Z-score).

▪ **Khai phá dữ liệu bằng máy học (Data Mining):** Hỗ trợ đồng thời ba nhóm bài toán học máy cốt lõi bao gồm: gom cụm dữ liệu khám phá liên kết ẩn (qua các thuật toán phân cụm phân cấp, K-Means, CLARA, hoặc cực đại hóa kỳ vọng), phân loại (qua thuật toán K-láng giềng gần nhất, cây quyết định và máy học véc-tơ hỗ trợ) và hồi quy đa biến nhằm mô hình hóa, dự báo các chỉ số sinh trưởng và năng suất cây trồng.

▪ **Diễn giải, trực quan hóa và cập nhật tri thức (Interpretation & Evaluation / Knowledge Integration):** Giao diện đồ họa tiêu chuẩn của trình duyệt được sử dụng để vẽ và hiển thị trực quan các biểu đồ (như biểu đồ phân tán hai chiều hoặc sơ đồ hình cây phân cấp) cùng các bộ chỉ số đánh giá hiệu năng thuật toán. Kết quả và tri thức mới khám phá từ quá trình phân tích được tự động cập nhật ngược trở lại hệ thống Wikicrop để làm giàu ngân hàng dữ liệu chung của cộng đồng.

4. KIẾN TRÚC KIỂM CHỨNG VÀ NGUYÊN TẮC PHÊ DUYỆT CHUYÊN GIA

Việc kiểm chứng nguồn dữ liệu trên WikiCrop là rất quan trọng, đảm bảo độ tin cậy của thông tin và quan trọng hơn là giá trị phục vụ cộng đồng. Đóng vai trò một hệ thống tri thức mở, WikiCrop chứa cùng lúc hai nguồn thông tin: (1) thông tin có sẵn, trích lọc từ các nguồn dữ liệu tin cậy và (2) thông tin mới, đóng góp mới từ cộng đồng người dùng. Dù thuộc nhóm thông tin nào, nguyên tắc chung là cần thực hiện kiểm chứng và phê duyệt trước khi công bố chính thức. Đối với các nền tảng tri thức mở (nơi cộng đồng có quyền đóng góp thông tin), thách thức lớn nhất là tính nhiễu, lỗi văn phong cảm tính và hiện tượng sai lệch dữ liệu. WikiCrop giải quyết bài toán độ tin cậy này bằng cách thiết lập một kiến trúc kiểm chứng chặt chẽ bảo vệ ranh giới dữ liệu:

(1) Bộ lọc khóa cứng dữ liệu định lượng (Exact Match Verification)

Đối với các trường thông tin số liệu định lượng nông học có tính chất nhạy cảm về kỹ thuật (chu kỳ sinh trưởng, chiều cao cây mẫu, ngưỡng chịu mặn, năng suất lý thuyết), hệ thống kích hoạt bộ lọc tự động Exact Match. Bộ lọc này tiến hành tiền xử lý cơ học bằng cách xóa sạch ký tự rác, đồng bộ hóa khoảng trắng, chuẩn hóa dấu gạch ngang dải số và thực hiện so khớp tuyệt đối chuỗi với hằng số kiểm thử được trích xuất từ các tài liệu chọn tạo giống chính thống của các Viện, Trường. Mọi đề xuất dữ liệu không đạt độ chính xác tuyệt đối đều bị hệ thống

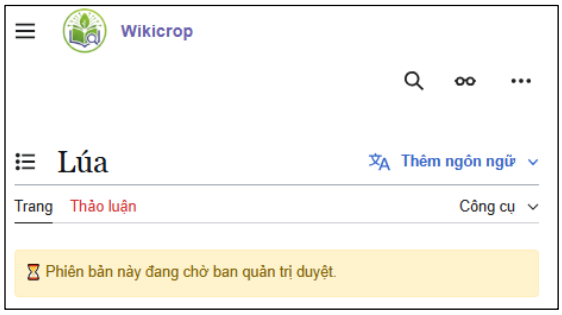
lập tức từ chối ghi nhận để bảo vệ tính nguyên bản của thông số kỹ thuật khuyến nông.

Cơ chế này chủ yếu được sử dụng khi người dùng muốn nạp dữ liệu vào hệ thống thông qua các mô hình LLM để tiết kiệm thời gian (thực hiện trên phân hệ WikiLab). Khi đó, WikiLab được sử dụng để thực hiện quy trình trích xuất dựa trên khung mẫu (Template-guided Extraction). Hệ thống tái sử dụng khung thuộc tính đã chuẩn hóa của một giống chuẩn để hỗ trợ bóc tách và khởi tạo trang cho giống mới có đặc tính tương đồng (ví dụ: dùng khung giống ST25 để tạo trang cho ST24). Đối với các thông số định lượng nhạy cảm (ví dụ như khả năng chịu mặn, thời gian sinh trưởng, năng suất), hệ thống áp dụng cơ chế xác thực cứng (Exact Match & Schema Validation) sau khi đã chuẩn hóa chuỗi. Dữ liệu trích xuất bắt buộc phải tuân thủ nghiêm ngặt khung cấu trúc chuẩn mới được nạp tự động, giúp giảm thiểu thông tin ảo giác và hạn chế tối đa các trường hợp nhận diện sai (False Positive). Cuối cùng, cũng tương tự như nhập liệu thủ công, mạng lưới các chuyên gia (con người) cũng tiến hành kiểm định dữ liệu trước khi thông qua nội dung trang dữ liệu giống cây trồng mới.

(2) Phân quyền kiểm duyệt theo miền tri thức chuyên sâu

Hệ thống phân hoạch tập hợp toàn bộ các chuyên gia thuộc hội đồng kiểm duyệt thành các tiểu ban chuyên môn độc lập dựa trên cấu trúc họ thực vật học thuật. Mỗi chuyên gia (là giảng viên, nhà khoa học, nhà quản lý) chỉ được hệ thống cấp quyền phê duyệt đối với miền tri thức chuyên sâu tương ứng của mình. Ví dụ, chuyên gia về cây lương thực ngắn ngày sẽ chịu trách nhiệm quản trị trực tiếp đối với họ Hòa thảo (Poaceae), trong khi chuyên gia về cây ăn trái lâu năm sẽ quản lý ranh giới bài viết thuộc họ Đào lộn hột (Anacardiaceae) hay họ Cam chanh (Rutaceae), bảo đảm thông tin nạp vào được thẩm định bởi người có chuyên môn phù hợp.

Hình 6 là một ví dụ về việc hệ thống cảnh báo cho người dùng biết rằng dữ liệu trang thông tin cây lúa chưa được phê duyệt ở cấp cao nhất sau khi được chỉnh sửa bởi một người dùng bất kỳ. Đây là cơ chế quan trọng cho phép kiểm soát dữ liệu ở mức độ cao nhất bởi cộng đồng chuyên gia.



Hình 6. Thông tin hệ thống cho biết dữ liệu cập nhật chưa được nhóm chuyên gia phê duyệt

5. KẾT QUẢ TRIỂN KHAI HỆ THỐNG VÀ THỰC NGHIỆM PHÂN TÍCH DỮ LIỆU

Bên cạnh các giá trị học thuật, việc phát triển hệ thống WikiCrop được đề ra với mục tiêu phục vụ thiết thực cho sự phát triển chung của nền kinh tế nông nghiệp. Với kiến trúc mở của MediaWiki và cơ chế tự động hóa quá trình thu thập và chuẩn hóa dữ liệu giống cây trồng, WikiCrop kỳ vọng sẽ được mở rộng cả về số lượng và chất lượng giống. Sau đây là một số kết quả vận hành thử nghiệm của hệ thống.

5.1. Kết quả triển khai và vận hành hệ thống

Số hóa dữ liệu giống cây trồng ở quy mô vùng nông nghiệp:

Để đánh giá năng lực quản trị dữ liệu và khả năng mở rộng quy mô của mô hình kiến trúc, nhóm nghiên cứu đã lựa chọn, cấu trúc hóa và nạp thành công thông tin của khoảng 300 giống cây trồng lên hệ thống. Tập dữ liệu này có độ bao phủ đa dạng trên nhiều loài thực vật, đại diện cho các nhóm cây trồng chủ lực mang ý nghĩa kinh tế trọng điểm tại vùng ĐBSCL.

Từng bước chuẩn hóa cấu trúc thông tin của các giống cây trồng:

Bên cạnh dữ liệu được các chuyên gia nông nghiệp tổng hợp trong quá trình nghiên cứu (Trường Nông nghiệp – Đại học Cần Thơ), một lượng dữ liệu mới được tự động trích lọc từ nguồn dữ liệu chỉ định, thông qua phân tích dữ liệu dựa trên mô hình ngôn ngữ lớn; sau đó, dữ liệu được kiểm chứng và ánh xạ chính xác vào cây thư mục phân loại học giống cây trồng.

Để đánh giá khả năng vận hành thực tế của kiến trúc, hệ thống đã khởi tạo và nạp thành công dữ liệu khởi điểm cho khoảng 300 giống cây trồng chủ lực tại vùng ĐBSCL. Con số 300 giống này đại diện cho

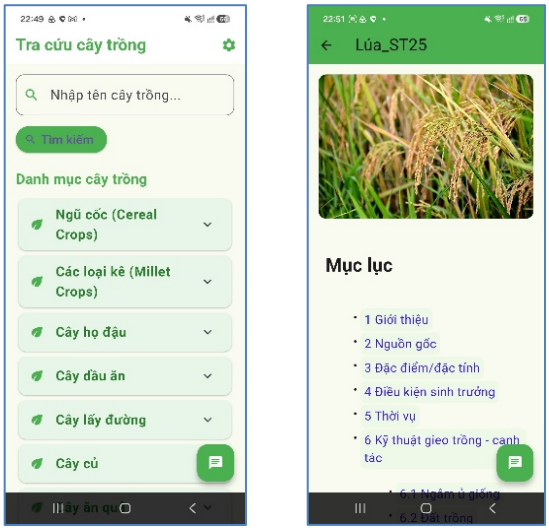
quy mô dữ liệu vận hành ban đầu nhằm chứng minh khả năng mở rộng của hạ tầng lưu trữ, chứ không phải là một tập mẫu cho thực nghiệm thống kê. Bên cạnh việc mở rộng quy mô dữ liệu, khả năng trích xuất chính xác của phân hệ WikiLab cũng được kiểm tra thông qua các trường hợp giống cây trồng có đặc tính tương đồng. Ví dụ, phân hệ có thể tái sử dụng khung cấu trúc thuộc tính (template) đã định dạng của giống lúa ST25 để hỗ trợ bóc tách dữ liệu cho giống lúa ST24, qua đó khả năng phân biệt được đánh giá và các thông số nông học riêng biệt giữa hai giống được ghi nhận chính xác mà không gây lẫn lộn dữ liệu.

Tính bất biến cấu trúc khung mẫu thông tin (Template Invariance) của hệ thống là yếu tố quan trọng mà WikiCrop hướng tới, nhằm giúp các trang tin có sự tương đồng cao nhất về nội dung và hình thức. Qua đó, có thể cung cấp cho cộng đồng nông nghiệp các nguồn tin mang tính khoa học và thực tiễn cao, giúp thúc đẩy hiệu quả hoạt động của hệ sinh thái nông nghiệp. Dù là một nền tảng mở hướng đến sự đóng góp và chia sẻ tri thức từ cộng đồng nhưng WikiCrop vẫn kiểm soát tốt tính chính xác khoa học nhờ cơ chế định khuôn dữ liệu chặt chẽ. Ngay cả khi hai giống thử nghiệm (ST24 và ST25) có sự tương đồng rất cao về mặt sinh học và phả hệ, hệ thống vẫn hỗ trợ người nhập liệu (hoặc các tác vụ TTNT tự động) bóc tách, định danh chính xác các thuộc tính nông học riêng biệt và tự động tách biệt cấu trúc cũ. Điều này giúp kiểm soát hiệu quả các giá trị nhiễu do nhầm lẫn thông tin giữa các giống quá gần gũi về thuộc tính, đảm bảo tính toàn vẹn dữ liệu cho ngân hàng tri thức mà không làm mất đi tính linh hoạt của một nền tảng bách khoa mở.

Ý nghĩa thực tiễn của giải pháp hình thức hóa kiến trúc:

Văn bản kỹ thuật nông nghiệp thu thập từ các nguồn trên Internet thường mang tính mô tả văn xuôi dài dòng, dàn trải và có mật độ nhiễu lớn. Khi được tích hợp vào hệ thống, WikiCrop thực hiện quá trình sàng lọc và chuẩn hóa để chuyển đổi thông tin thành các thuộc tính kỹ thuật cô đọng, phân tách rõ ràng bằng danh sách có cấu trúc. Việc định dạng cấu trúc này giúp tối ưu hóa hiệu năng truy vấn cơ sở dữ liệu và cung cấp giao diện trực quan, khoa học cho người nông dân, kỹ sư và cán bộ khuyến nông dễ dàng so sánh chéo đặc tính sinh học giữa các giống cây trồng để lựa chọn mô hình canh tác phù hợp nhất với thổ nhưỡng của địa phương. Bên cạnh đó, nhằm chuẩn bị cho giai đoạn triển khai mở rộng và đánh giá tính khả dụng trong thực tiễn, nền tảng đã tích hợp thành công công cụ phân tích Microsoft Clarity. Mặc dù hệ thống hiện đang ở giai đoạn thử nghiệm vận hành kiến trúc lõi và chưa tiến hành thu thập số liệu người dùng đại trà, việc tích hợp sẵn hạ tầng đo lường này (bao gồm bản đồ nhiệt, tỷ lệ cuộn trang và theo dõi luồng truy cập) đã tạo tiền đề kỹ thuật vững chắc cho các nghiên cứu đánh giá tương tác người dùng định lượng trong tương lai.

Đưa tri thức nông nghiệp ra thực địa qua nền tảng di động:



Hình 7. Giao diện ứng dụng di động WikiCrop trên hệ điều hành Android: màn hình giao diện chính hỗ trợ tra cứu (bên trái) và thông tin đặc tính sinh học, kỹ thuật canh tác của giống lúa ST25 (bên phải)

Để hiện thực hóa mục tiêu đưa ngân hàng giống cây trồng vào phục vụ kinh tế nông nghiệp, hệ thống đã hoàn thiện đóng gói và phát hành ứng dụng di động WikiCrop. Phiên bản di động cho phép đồng bộ hóa dữ liệu theo thời gian thực với lõi MediaWiki, mang đến trải nghiệm tra cứu thông tin giống và kỹ thuật canh tác mượt mà cho người dùng. Đặc biệt, ứng dụng đã vượt qua các quy trình thẩm định khắt khe của các nền tảng phân phối, tiêu biểu như việc hoàn tất quy trình xác minh nhà phát triển để được cấp quyền truy cập sản xuất chính thức trên cửa hàng ứng dụng Google Play Store; đồng thời, phiên bản dành cho hệ điều hành iOS cũng đang được triển khai thử nghiệm trên nền tảng TestFlight. Điều này minh chứng cho tính hoàn thiện về mặt công nghệ, mức độ an toàn dữ liệu và khả năng duy trì vận hành dài hạn của dự án trong thực tiễn. Giao diện vận hành thực tế của ứng dụng được minh họa trực quan thông qua màn hình trang chủ của hệ thống trên hệ điều hành Android (Hình 7, bên trái) cùng màn hình hiển thị chi tiết các đặc tính sinh học

và hướng dẫn kỹ thuật canh tác của giống lúa ST25 (Hình 7, bên phải).

5.2. Kết quả thực nghiệm phân tích dữ liệu

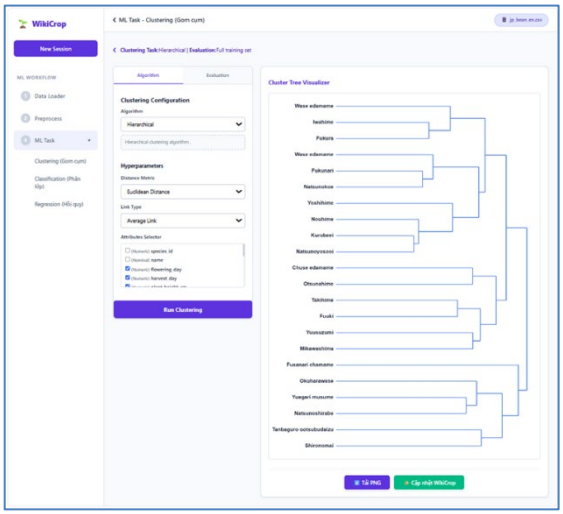
Nhằm kiểm chứng năng lực tính toán và độ ổn định của các giải thuật học máy khi thực thi trực tiếp ở phía máy khách, hệ thống đã tiến hành thực nghiệm khai phá dữ liệu trên bộ dữ liệu gồm 22 giống đậu nành rau Nhật Bản. Tập dữ liệu này chứa 15 chỉ tiêu đặc tả nông học và hình thái thực vật đa biến (bao gồm 8 thuộc tính định lượng và 7 thuộc tính định tính), đại diện cho một bài toán phân tích sinh học thực tế có tính chất dị đồng và phức tạp. Quá trình khai phá tri thức di truyền được thực hiện tuần tự bám sát theo quy trình KDD tiêu chuẩn:

Giai đoạn nạp và tiền xử lý dữ liệu: Hệ thống hỗ trợ nạp dữ liệu linh hoạt, giúp nhà nghiên cứu chủ động lựa chọn giữa việc khai thác nguồn dữ liệu sinh học sẵn có trên nền tảng WikiCrop và tải lên tệp dữ liệu thực nghiệm độc lập của cá nhân để phân tích riêng. Trong thực nghiệm này, tập dữ liệu 22 giống đậu nành rau Nhật Bản được nạp dưới dạng véc-tơ số học. Hệ thống tự động quét và hiển thị các thông tin thống kê mô tả tổng quát, giúp phát hiện sớm các giá trị ngoại lai. Sau đó, dữ liệu được làm sạch và chuẩn hóa thang đo thông qua giải thuật tiêu chuẩn hóa điểm số (Z-score) nhằm triệt tiêu độ chệch khoảng cách hình học giữa các thuộc tính nông học dị biệt trước khi đưa vào mô hình học máy.

Giai đoạn khai phá bằng máy học: Thuật toán phân cụm phân cấp tích lũy từ dưới lên (Agglomerative Hierarchical Clustering) sử dụng ma trận khoảng cách hình học Euclide và kỹ thuật liên kết trung bình (Average Link) được thực thi hoàn toàn bằng ngôn ngữ JavaScript trực tiếp trong bộ nhớ trình duyệt. Phương thức tính toán phân tán này giúp bảo vệ an toàn dữ liệu chọn giống nhạy cảm của chuyên gia, đồng thời giải phóng hoàn toàn tải trọng tính toán cho máy chủ.

Giai đoạn diễn giải và cập nhật tri thức: Kết quả phân nhóm di truyền được hệ thống vẽ và hiển thị trực quan tức thời dưới dạng sơ đồ hình cây phân cấp hai chiều (Dendrogram) thông qua công cụ đồ họa tiêu chuẩn HTML5 Canvas phía máy khách. Ngay tại giao diện trực quan này, chuyên gia có thể sử dụng chức năng "Cập nhật WikiCrop" để lưu trữ tri thức phân nhóm di truyền mới phát hiện này ngược trở lại cơ sở dữ liệu lõi, hoàn tất quy trình KDD (Hình 8).

Kết quả thực nghiệm này bước đầu khẳng định tính khả thi về mặt kiến trúc kỹ thuật của hệ thống, minh chứng rằng các luồng giải thuật máy học có thể thực thi trơn tru và trực tiếp ở phía máy khách. Tuy nhiên, do giới hạn quy mô tập dữ liệu thử nghiệm còn nhỏ (22 giống) và mô hình phân cụm hiện tại chưa được đối chiếu với các chỉ số đánh giá nội tại (như hệ số Silhouette Score hay Davies-Bouldin Index) nên các kết quả khai phá này chủ yếu mang ý nghĩa minh họa khả năng vận hành của phần mềm thay vì đưa ra các kết luận chuyên sâu về di truyền học. Mặc dù vậy, việc hiện thực hóa thành công luồng xử lý KDD trên trình duyệt đã cung cấp một công cụ nền tảng hỗ trợ ra quyết định không cần viết mã lệnh, tạo tiền đề để các nhà nông học tiếp tục mở rộng quy mô dữ liệu và kiểm chứng tri thức trong tương lai.



Hình 8. Ảnh chụp màn hình kết quả phân tích gom nhóm 22 giống đậu nành rau trên phân hệ Phân tích dữ liệu thông minh

6. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Nền tảng WikiCrop đã chứng minh tính khả thi trong việc xây dựng ngân hàng tri thức phân cấp sâu đến cấp độ giống cây trồng. Bằng cách tích hợp lõi MediaWiki, cơ chế kiểm duyệt lai và phân hệ phân tích trực tiếp trên trình duyệt, bài toán chuẩn hóa dữ liệu nông nghiệp chuyên ngành đã được giải quyết một cách hiệu quả trong nghiên cứu. Đặc biệt, việc phát hành ứng dụng di động đã hoàn thiện chuỗi truyền tải tri thức, cung cấp một hạ tầng số thiết thực đưa thông tin khoa học đến trực tiếp nông dân và cán bộ khuyến nông tại thực địa.

Mặc dù bước đầu vận hành hiệu quả nhưng nghiên cứu vẫn tồn tại một số hạn chế. Về quy mô, tập dữ liệu minh họa khả năng vận hành hiện mới dừng ở 300 giống cây trồng chủ lực tại ĐBSCL nhằm kiểm chứng hạ tầng, chưa bao phủ toàn quốc. Về kỹ thuật, dù bộ lọc cơ học giúp hạn chế tối đa rủi

ro thông tin ảo giác từ LLM nhưng hệ thống vẫn có thể bỏ sót dữ liệu và bắt buộc phải duy trì mạng lưới chuyên gia để thẩm định. Quy trình kiểm duyệt thủ công này có thể tạo ra rào cản về mặt thời gian khi khối lượng dữ liệu mở rộng nhanh chóng.

Trong giai đoạn tiếp theo, quy mô dữ liệu cần được mở rộng cho toàn hệ thực vật Việt Nam và đồ thị tri thức (Knowledge Graph) liên kết đặc tính sinh học với thổ nhưỡng cần được được xây dựng. Việc ứng dụng di động sẽ tiếp tục được nâng cấp tích hợp thị giác máy tính nhằm chẩn đoán sâu bệnh qua hình ảnh, hướng tới việc tự động tính toán và đề xuất các giải pháp canh tác cá nhân hóa, thích ứng với biến đổi khí hậu cho từng tiểu vùng sinh thái tại ĐBSCL.

LỜI CẢM TẠ

Nghiên cứu này được tài trợ bởi Đại học Cần Thơ, mã số: CTCS2025-01-01.

TÀI LIỆU THAM KHẢO

Attardi, G. (2015). *WikiExtractor*. GitHub. https://github.com/attardi/wikiextractor

Bộ Nông nghiệp và Phát triển nông thôn. (2005a). *575 Giống cây trồng nông nghiệp mới*. Nhà xuất bản Nông nghiệp.

Bộ Nông nghiệp và Phát triển nông thôn. (2005b). *Quyết định ban hành danh mục nguồn gen cây trồng quý hiếm cần bảo tồn (Số: 80/2005/QĐ-BNN)*. https://thuvienphapluat.vn/van-ban/Tai-nguyen-Moi-truong/Quyet-dinh-80-2005-QD-BNN-Danh-muc-nguon-gen-cay-trong-quy-hiem-can-bao-ton-8471.aspx

Bộ Nông nghiệp và Phát triển nông thôn. (2016). *Thông tư ban hành danh mục bổ sung giống cây trồng được phép sản xuất, kinh doanh ở Việt Nam* (Số 28/2016/TT-BNNPTNT). https://fcri.com.vn/thong-tu-so-28-2016-tt-bnnptnt-ngay-27-7-2016-v-v-ban-hanh-danh-muc-bo-sung-giong-cay-trong-duoc-phep-san-xuat-kinh-doanh-o-viet-nam/

Chính phủ. (2023). *Quy định chi tiết một số điều và biện pháp thi hành Luật Sở hữu trí tuệ về quyền đối với giống cây trồng* (Số 79/2023/NĐ-CP). https://vanban.chinhphu.vn/?docid=208976&pageid=27160

Dublin Core Metadata Initiative. (2026). *DCMI Metadata Terms*. https://www.dublincore.org/specifications/dublin-core/dcmi-terms/

Đạo, L. T., Tài, Â. T., Pháp, V. A., & Thư, T. N. M. (2017). Ứng dụng giải thuật gom nhóm dữ liệu để nhận diện sự tương đồng giữa các giống lúa. *Tạp chí Khoa học Đại học Cần Thơ*, (CĐ Công nghệ TT), 140–144. https://doi.org/10.22144/ctu.jsi.2017.019

Đáp, B. H. (1980). *Cây lúa Việt Nam*. Nhà xuất bản Khoa học Kỹ thuật.

EOS Data Analytics. (2026). *Types of crops: How to classify and manage different plants*. https://eos.com/blog/types-of-crops/

Food and Agriculture Organization of the United Nations. (2005). *A System of Integrated Agricultural Censuses and Surveys. Volume 1: World Programme for the Census of Agriculture 2010*. Food and Agriculture Organization of the United Nations. https://openknowledge.fao.org/handle/20.500.14283/a0135e

Food and Agriculture Organization of the United Nations. (2026). *AGROVOC Multilingual Agricultural Thesaurus*. https://agrovoc.fao.org/

Huyền, H.T.T. (2025). *Doanh nghiệp Đồng bằng sông Cửu Long và vai trò trong phát triển kinh tế nông nghiệp*. https://www.quanlynhanuoc.vn/2025/06/14/doanh-nghiep-dong-bang-song-cuu-long-va-vai-tro-trong-phat-trien-kinh-te-nong-nghiep/

International Rice Research Institute. (2026). *Rice Knowledge Bank*. http://www.knowledgebank.irri.org/

Keycloak Community. (2026). *Keycloak: Open Source Identity and Access Management*. https://www.keycloak.org/

Kontokostas, D., Bratsas, C., Auer, S., Hellmann, S., Antoniou, I., & Metakides, G. (2012). Internationalization of Linked Data: The case of the Greek DBpedia edition. *Web Semantics: Science, Services and Agents on the World Wide Web*, *15*, 51–61. https://doi.org/10.1016/j.websem.2012.01.001

Li, J.-Y., Wang, J., & Zeigler, R. S. (2014). The 3,000 rice genomes project: new opportunities and challenges for future rice research. *GigaScience*, *3*(1), 8. https://doi.org/10.1186/2047-217X-3-8

Liên, Q. T. Á., Nhung, N. T. H., & Thành, V. C. (2012). Đánh giá khả năng chịu mặn và phẩm chất của giống lúa Sỏi, Một bụi hồng và Nàng quớt biển. *Tạp chí Khoa học Đại học Cần Thơ*, (24a), 281–289. https://ctujsvn.ctu.edu.vn/index.php/ctujsvn/article/view/1314

Manning, C. D. (2022). Human Language Understanding and Reasoning. *Daedalus*, *151*(2), 127–138. https://doi.org/10.1162/daed_a_01905

MediaWiki. (2026). *MediaWiki*. https://www.mediawiki.org/wiki/MediaWiki

Microsoft. (2026). *Microsoft Clarity*. https://clarity.microsoft.com/

Nguyen, T. K. T., Tran, T. N., Chau, T. N., Nguyen, K. T., Vo, T. T., Tran, T. T., & Tran, N. T. (2024). Assessing the field trials of the salt stress-tolerant elite rice lines. *Omonrice*, 22, 54–68. https://clrri.org/ver2/uploads/OMR2023%20(Kieu%20Tien).pdf

Quốc hội. (2018). *Luật Trồng trọt Việt Nam 2018* (Số: 31/2018/QH14). https://vanban.chinhphu.vn/?docid=206097&pageid=27160

Sheaffer, C. C., & Moncada, K. M. (2012). *Introduction to Agronomy: Food, Crops, and Environment*. Cengage Learning.

Sosef, M., Degreef, J., Engledow, H., & Meerts, P. (2020). *Classification et nomenclature botanique - une introduction (version 1)*. Jardin botanique de Meise. https://doi.org/10.5281/zenodo.3706717

Thái, L. X., & Dũng, T. N. (2013). Chọn lọc giống lúa chống chịu mặn ở Đồng Bằng Sông Cửu Long. *Tạp chí Khoa học Đại học Cần Thơ*, (28), 79–85. https://ctujsvn.ctu.edu.vn/index.php/ctujsvn/article/view/1663

Thanh, T. N. T., Dũng, T. N., Mai, T. T. X., & Khang, Đ. T. (2013). Tuyển chọn và tái sinh một số giống lúa có khả năng chịu mặn thích ứng với biến đổi khí hậu ở Đồng Bằng Sông Cửu Long. *Tạp chí Khoa học Đại học Cần Thơ*, (26), 104–111. https://ctujsvn.ctu.edu.vn/index.php/ctujsvn/article/view/102

Tổng cục Thống kê. (2026). *Niên giám Thống kê 2025*. https://www.nso.gov.vn/default/2026/07/nien-giam-thong-ke-2025/

Tran, N. H., & Pham, T. K. V. (2024). The differential expressed genes in the PDH47 transgene-carrying rice under drought stress. *Omonrice*, *22*, 1–18. https://clrri.org/ver2/uploads/OMR2023%20%20(H%C3%A8%20).pdf

International Union for the Protection of New Varieties of Plants. (1991). *Công ước quốc tế về bảo hộ giống cây trồng mới*. https://www.upov.int/documents/d/upov/docs-en-publications-upov_pub_221.pdf

Viện Nghiên cứu Cây ăn quả miền Nam. (2026). *Trang chủ Viện Nghiên cứu Cây ăn quả miền Nam (SOFRI)*. https://sofri.org.vn/vi/tin-tuc.html

World Wide Web Consortium. (2024). *Data Catalog Vocabulary (DCAT)*. https://www.w3.org/TR/vocab-dcat-3/