Tạp chí Khoa học Trường Đại học Cần Thơ
Phần A: Khoa học Tự nhiên, Công nghệ và Môi trường
website: ctujsvn.ctu.edu.vn

DOI:10.22144/ctu.jvn.2021.170

GIẢI PHÁP QUY HOẠCH QUẢN LÝ DỮ LIỆU HỖ TRỢ NÔNG NGHIỆP THÔNG MINH

Trương Xuân Việt[1*], Nguyễn Hoàng Việt[1], Lê Hoàng Thảo[1], Cù Vĩnh Lộc[1], Trần Hoàng Việt[1], Lê Thành Phiêu[2] và Nguyễn Hiếu Trung[3]

[1]*Trung tâm Công nghệ Phần mềm, Trường Đại học Cần Thơ*

[2]*Phòng Quản trị Thiết bị, Trường Đại học Cần Thơ*

[3]*Viện Nghiên cứu Biến đổi Khí hậu, Trường Đại học Cần Thơ*

[*]*Người chịu trách nhiệm về bài viết: Trương Xuân Việt (email: txviet@ctu.edu.vn)*

Thông tin chung:
Ngày nhận bài: 16/08/2021
Ngày nhận bài sửa: 22/10/2021
Ngày duyệt đăng: 25/12/2021

Title:
Solution for data management plan applied in smart agriculture

Từ khóa:
Quy hoạch quản lý dữ liệu, data stewardship, nguyên lý FAIR, quản trị dữ liệu, chuẩn OGC, domain-based web service, web features service, kho dữ liệu, dữ liệu lớn

Keywords:
Data management plan (DMP), data stewardship, FAIR principles, data governance,OGC standards, domain-based web service, web features service, data warehouse, big data

ABSTRACT

The ability to make accurate and timely decisions in agriculture is directly related to the level of smart agriculture. From the perspective of scientific management, it can be seen that the problem of collecting, managing and sharing the necessary data sources for agricultural research has become urgent. This paper concerns the causes of the lack of agricultural research data sources in Vietnam, which are separated into two aspects: (1) the ability to manage and share public data sources and (2) the research community's capacity to exploit open data sources. On that basis, a global solution for data management planning based on FAIR principles (to be Findable, Accessible, Interoperable, Reusable) is proposed. Developing a Data Management Plan (DMP - Data Management Plan) is the main solution, applied on two sectors: (1) Government sector (public service) – which is directly affected by legal policies on building open data repositories, and (2) The sector of research and academic units (research centers, institutes, schools) – the data management mechanism on this sector is mostly voluntary, but it is very dynamic in the ability to exploit open data sources and high capacity in data analysis.

TÓM TẮT

*Năng lực ra quyết định chính xác và kịp thời trong nông nghiệp ảnh hưởng trực tiếp đến mức độ thông minh của nền nông nghiệp. Đứng ở góc độ quản lý khoa học, có thể nhận thấy bài toán thu thập, quản lý, chia sẻ các nguồn dữ liệu cần thiết cho các nghiên cứu nông nghiệp đã trở nên bức thiết. Bài viết tập trung vào việc đánh giá các nguyên nhân cốt lõi dẫn đến việc thiếu hụt nguồn dữ liệu nghiên cứu nông nghiệp ở Việt Nam, xét trên hai khía cạnh: (1) khả năng quản lý và chia sẻ các nguồn dữ liệu nghiên cứu, đặc biệt là dữ liệu công và (2) năng lực khai thác các nguồn dữ liệu mở của cộng đồng nghiên cứu. Dựa trên đó, bài viết đề xuất giải pháp tổng thể về quy hoạch dữ liệu dựa trên các nguyên lý FAIR (to be **F**indable, **A**ccessible, **I**nteroperable, **R**eusable).Cách thức tiếp cận là xây dựng một Quy hoạch Quản lý Dữ liệu (DMP – Data Management Plan) cho hai khối cơ quan chức năng: (1) Khối chính phủ (dịch vụ công) – là khối chịu ảnh hưởng trực tiếp bởi các văn bản quy phạm pháp luật trong lĩnh vực xây dựng nguồn dữ liệu mở; và (2) Khối nghiên cứu, học thuật (trung tâm nghiên cứu, viện, trường) – là khối mà cơ chế quản lý dữ liệu phần nhiều mang tính tự nguyện, nhưng lại rất năng động trong khả năng khai thác các nguồn dữ liệu mở và có năng lực cao vềphân tích dữ liệu.*

1. ĐẶT VẤN ĐỀ

Ra quyết định dựa trên phân tích dữ liệu là nhu cầu căn bản của quản trị thông minh. Một nền nông nghiệp thông minh nói chung phải đáp ứng hai tiêu chí quan trọng của xu hướng công nghệ, một là xu hướng *Chuyển đổi số*, và hai là xu hướng *Cách mạng Công nghiệp lần thứ tư* (4.0). *Chuyển đổi số* là xu hướng ra đời trước và đã phát triển mạnh mẽ ở các quốc gia phát triển, xu hướng này dẫn đến phát sinh một nguồn dữ liệu số khổng lồ trong quá trình xây dựng và phát triển. Trong khi đó, *Cách mạng Công nghiệp 4.0* mới được định nghĩa từ năm 2013, hướng đến những giải pháp thông minh nhằm khai thác các nguồn dữ liệu khổng lồ mà *Chuyển đổi số* mang lại. Tại Việt Nam, hai xu hướng này đều khởi động chậm hơn thế giới và đang phát triển song hành. Nông nghiệp thông minh liên quan đến cả hai, trong đó nguồn dữ liệu số là điểm giao thoa, mà nếu không giải quyết tốt được bài toán quy hoạch quản lý dữ liệu sẽ không thể đưa ra được các quyết định chính xác và có cơ sở khoa học trong quản lý nông nghiệp. Hay nói cách khác, đó sẽ là nền nông nghiệp vẫn còn dựa trên cảm tính, rời rạc và thiếu tính dự báo. Khái niệm thông minh trong bài viết này dựa theo cách phân loại bao gồm ba loại hình (và cũng là ba cấp độ) khác nhau của các kỹ thuật phân tích dữ liệu (Data analytics, được trình bày trong James (2012)), bao gồm: *Phân tích mô tả* (Descriptive analytics), *Phân tích dự báo* (Predictive analytics) và *Phân tích đề xuất* (Prescriptive analytics).

Ba nhóm đối tượng quan trọng có thể tác động trực tiếp hoặc gián tiếp đến khả năng thông minh của nền nông nghiệp, bao gồm: (1) Nhóm chuyên gia tư vấn chính sách khối chính phủ, (2) Nhóm chuyên gia thuộc khối doanh nghiệp/hợp tác xã và (3) Các nhà khoa học và lực lượng nghiên cứu thuộc lĩnh vực nghiên cứu/giáo dục. Sở dĩ ba nhóm đối tượng này được đề cập trong nghiên cứu vì đây là các nhóm đối tượng có cách tiếp cận khoa học nhất với dữ liệu so với các nhóm còn lại khi đưa ra các quyết định liên quan đến quản trị nền nông nghiệp.

Sự phân mảnh, thiếu nhất quán và thiếu tính tương tác giữa các chủ sở hữu các nguồn dữ liệu là nguyên nhân chính dẫn đến các khó khăn về khả năng tiếp cận dữ liệu trong nghiên cứu. Các khó khăn này chủ yếu liên quan đến việc Việt Nam chưa hoàn toàn sẵn sàng tiếp cận với Khoa học mở (Open Science), cũng như chưa đánh giá đúng và vận dụng hiệu quả các chính sách chung của nhà nước và các tiêu chuẩn quốc tế. Lấy một số ví dụ để minh chứng cho các khó khăn mà cộng đồng nghiên cứu thường gặp. Chẳng hạn, phần lớn các dữ liệu cơ bản như bản đồ hành chính cấp tỉnh/huyện/xã của Việt Nam vẫn phải lấy về từ nguồn DIVA-GIS (Hijmans, 2016) (không phải dữ liệu chính thức của Việt Nam); dữ liệu về hệ thống sông ngòi, kênh rạch hiện chưa có đơn vị cung cấp chính thức; dữ liệu về hệ thống đập thủy lợi, đê điều càng khan hiếm hơn. Trong khi đó, phần lớn các dự án nghiên cứu lớn đều mang tính liên ngành, xuyên ngành và chịu ảnh hưởng của liên vùng, nên nhóm các dữ liệu không gian nêu trên đều rất cần thiết. Một ví dụ khác liên quan đến sự thiếu hụt dữ liệu về nguồn giống cây trồng cũng cần phải được nhanh chóng giải quyết. Chẳng hạn, có bao nhiêu giống lúa đang được canh tác tại Đồng bằng Sông Cửu Long, các thông số kỹ thuật về điều kiện canh tác và chất lượng nông phẩm ra sao, hoặc các khuyến cáo về chuẩn GAP (Schreinemachers et al., 2012), năng lực đảm bảo chất lượng của nông dân và hợp tác xã như thế nào, … Do vậy, mục tiêu đầu tiên của bài viết này là đưa ra các đánh giá sơ bộ về hiện trạng quản lý dữ liệu tại Việt Nam, sẽ được trình bày cụ thể ở Mục 4.

Bên cạnh đó, khái niệm chia sẻ dữ liệu đang được hiểu rất khác nhau của các bên liên quan. Trong đó có sự khác biệt khá xa về các nội dung quản lý: định dạng dữ liệu (hiểu theo nghĩa đơn giản là loại tập tin mà hệ thống cung cấp), nội dung dữ liệu (các thông tin cụ thể bên trong của tập dữ liệu, bao gồm cả nguồn gốc dữ liệu), chuẩn dữ liệu (các thông tin tuân thủ theo các tiêu chuẩn kỹ thuật chung), quyền truy cập dữ liệu, và đặc biệt, tính đồng nhất về yếu tố không gian và thời gian. Mỗi yếu tố nêu trên đều ảnh hưởng trực tiếp đến chất lượng dữ liệu, tham số đầu vào của mọi phương pháp phân tích thông minh. Chẳng hạn, nếu muốn so sánh chất lượng hai giống lúa, mà dữ liệu nhận được từ hai nguồn dữ liệu khác nhau, thì điều kiện cơ bản đầu tiên là các trường dữ liệu phải đồng nhất và tuân thủ theo một tiêu chuẩn về mô tả giống lúa đó, chẳng hạn về tên giống, mã giống, loại giống và vùng sản xuất, được mô tả theo một quy chuẩn nhất định nào đó, chẳng hạn theo Giấy phép 28/2016/TT-BNNPTNT về Ban hành Danh mục bổ sung giống cây trồng được phép sản xuất kinh doanh ở Việt Nam (Bộ Nông nghiệp & Phát triển Nông thôn, 2016). Một đòi hỏi quan trọng khác là việc ban hành các quy chuẩn này cần thống nhất với các tiêu chuẩn quốc tế khác như Dublin Core hay DCAT (Lisowska, 2016). Một ví dụ khác mà các nghiên cứu về môi trường, quản lý đất đai hay gặp phải là định dạng dữ liệu không có khả năng tái sử dụng, các đơn vị quản lý dữ liệu quốc gia ở Việt Nam liên quan đến dữ liệu không gian thường cung cấp dữ liệu ở dạng "chỉ xem", ví dụ dạng ảnh, mà không

cung cấp dữ liệu thuộc tính như **.shp* (của ESRI), **.kml, .gml* hay GeoJSON. Trong khi đó, các trung tâm dữ liệu quốc tế thường quan tâm đến vấn đề này dựa trên các nguyên tắc *bình đẳng về khả năng khai thác các nguồn dữ liệu*, nhất là dữ liệu công. Các nguồn dữ liệu công có thể được hiểu là các nguồn dữ liệu được quy định phải chia sẻ theo các chính sách khác nhau (miễn phí, thương mại, bảo mật), hoặc các dữ liệu nghiên cứu nhận được tài trợ công, như các đề tài nghiên cứu khoa học cấp nhà nước và cấp tỉnh/thành.

Sự thiếu hụt dữ liệu không phải là vấn đề mới, mà là vấn đề lâu năm của các nhà khoa học, nhà kinh tế và nhà quản trị, nó trở thành yếu tố cản trở sự phát triển kinh tế cá nhân, kinh tế doanh nghiệp và kinh tế vùng. Rất nhiều đề nghiên cứu khoa học khó được triển khai hoặc được triển khai nhưng tính ứng dụng không cao vì lý do không có khả năng thu thập dữ liệu cần thiết. Tuy nhiên, để khắc phục vấn đề trên, không đơn thuần là nêu ra các yêu cầu và sử dụng các mệnh lệnh hành chính để giải quyết, nó đòi hỏi một kế hoạch hoàn chỉnh và sự phối hợp đồng bộ *theo các nguyên lý mang tính khoa học và thực tiễn*. *Quy hoạch Quản lý Dữ liệu (DMP – Data Management Plan)* (Burnette et al., 2016) là một tiếp cận phổ biến, tiếp cận này thường dựa trên các nguyên lý chung được đồng thuận trong cộng đồng, chẳng hạn nguyên lý FAIR (*to be Findable, Accessible, Interoperable, Reusable*) (Lamprecht et al., 2020; Wilkinson et al., 2016). FAIR là nguyên lý được đề xuất bởi nhóm G20 từ năm 2016, đã được áp dụng và ngày càng chứng minh tính hữu hiệu trong cộng đồng nghiên cứu ở các quốc gia phát triển. Các hệ thống áp dụng nguyên lý FAIR chứng tỏ tính hiệu quả trong việc thu thập, chia sẻ, nối kết và phân tích dữ liệu, qua đó hỗ trợ thiết thực cho quá trình quản trị thông minh, cụ thể hơn là giúp cho các nhà khoa học (gián tiếp) và tư vấn chính sách (trực tiếp) để có thể đưa ra các quyết định dựa trên dữ liệu (data-driven decision). Đề xuất Mô hình Quy hoạch quản lý dữ liệu dựa trên FAIR là mục tiêu quan trọng thứ hai và cũng là mục tiêu quan trọng nhất mà bài viết hướng đến, nội dung của đề xuất này sẽ được nêu trong Mục 5 và các phân tích liên quan được nêu trong Mục 3 và 4.

2. NGHIÊN CỨU LIÊN QUAN

2.1. Quản trị dữ liệu

Quản trị dữ liệu (Data governance) được định nghĩa bao gồm các chính sách, thủ tục và nguyên tắc (bao gồm các tiêu chuẩn) chi phối dữ liệu của bạn. Quản trị dữ liệu đảm bảo độ tin cậy của dữ liệu và có cơ chế kiểm soát để các bên liên quan phải chịu trách nhiệm về chất lượng của chúng (Steve, 2009). Có thể thấy, khía cạnh mà quản trị dữ liệu quan tâm là các yếu tố mang tính cơ chế. Để các cơ chế của quản trị dữ liệu trở thành hiện thực, đòi hỏi nhà quản lý phải áp dụng nội hàm thứ hai là *quản lý dữ liệu* (data stewardship) (Teperek et al., 2018). Quản lý dữ liệu được định nghĩa như cách thức triển khai các chính sách, thủ tục và nguyên tắc mà quản trị dữ liệu quy định. Quản lý dữ liệu thực hiện giám sát, đảm bảo chất lượng và tính phù hợp cho mục đích quản lý tài sản dữ liệu của tổ chức, bao gồm siêu dữ liệu (metadata) cho các tài sản dữ liệu đó. Quản lý dữ liệu quan tâm đến ba tiến trình hợp tác, nối kết và chia sẻ dữ liệu (Datavault, n.d.). Siêu dữ liệu (metadata) là một dạng dữ liệu dùng để mô tả dữ liệu khác, một ví dụ dễ hiểu cho siêu dữ liệu chính là dòng tiêu đề cho các bảng dữ liệu. Siêu dữ liệu đóng vai trò rất quan trọng trong việc xác định ngữ nghĩa (semantic) của dữ liệu. Việc định nghĩa các chuẩn dữ liệu, về bản chất cũng là định nghĩa các chuẩn siêu dữ liệu, hay nói các khác quá trình này xác định cấu trúc và định dạng cụ thể bên trong tập dữ liệu, tên và kiểu dữ liệu mà tập dữ liệu đó phải có. Trong quản trị dữ liệu, ngữ nghĩa của dữ liệu là yếu tố quan trọng bậc nhất, đặc biệt khi dữ liệu đó có nhu cầu dùng để chia sẻ và sử dụng trong các giao thức truyền thông.

Trong nội dung đánh giá hiện trạng quản lý dữ liệu ở Việt Nam (Mục 4), có một đặc điểm đáng lưu ý là Chính phủ Việt Nam khá quan tâm đến các chính sách quản trị dữ liệu, tuy nhiên việc triển khai quản lý dữ liệu của phần nhiều các cơ quan chức năng đang ở vạch xuất phát, hoặc dừng ở mức độ tự phát.

2.2. Nguyên lý FAIR trong chia sẻ dữ liệu

Các nguyên lý FAIR(Lamprecht et al., 2020; Wilkinson et al., 2016) (dịch từ tiếng Anh là CÔNG BẰNG) được định nghĩa từ 4 nhóm nguyên lý thành viên liên quan đến quản lý dữ liệu:Findable (Tìm kiếm được, gồm F1 đến F4), Accessible (Thâm nhập được, gồm A1.1, A1.2 và A2), Interoperable (Tương tác được, gồm I1 đến I3) và Reusable (Có khả năng tái sử dụng, bao gồm R1.1, R1.2 và R1.3). Các nguyên tắc FAIR nhấn mạnh khả năng hoạt động của máy (nghĩa là khả năng của các hệ thống tính toán có thể tìm kiếm, truy cập, tương tác và sử dụng lại dữ liệu mà không có hoặc hạn chế đến mức thấp nhất sự can thiệp của con người), bởi vì con người ngày càng dựa vào sự hỗ trợ của máy tính trong xử lý dữ liệu do sự gia tăng khối lượng, độ phức tạp và tốc độ tạo dữ liệu. Chính vì có nhiều nguyên lý cùng lúc được nêu trong FAIR nên quá trình quản lý dữ

liệu đòi hỏi nhiều tiêu chuẩn kỹ thuật khác nhau và thông thường rất ít trung tâm dữ liệu thỏa mãn cùng lúc tất cả các tiêu chí do FAIR đề xuất.

Trong các nguyên lý này, bài viết sẽ đi sâu vào phân tích các yêu cầu về mã định danh tài nguyên dữ liệu bằng DOI (Digital Object Identifier) hoặc URL/URI (nguyên lý F), siêu dữ liệu (metadata) (nguyên lý F); các giao thức mở, chuẩn dữ liệu phổ biến (nguyên lý A1.1); các định dạng dữ liệu đa dạng có thể truy vấn như JSON, TURTLE, RDF/XML, CSV, … (nguyên lý I1), khả năng sử dụng các bộ từ vựng chuẩn mực (hoặc cao hơn là sử dụng phân loại học(taxonomy)) trong quản lý dữ liệu (nguyên lý I2); và cuối cùng, dữ liệu cần được liên kết với xuất xứ chi tiết (nguyên lý R1.2).

Có khá nhiều nền tảng công nghệ công bố hỗ trợ FAIR và được ứng dụng rộng rãi trong giới khoa học và công nghệ, một số trong đó được giới thiệu trong (Wilkinson et al., 2016).Trong đó, có thể liệt kê một số ứng dụng tiêu biểu như trình bày trong Bảng 1 (Bao gồm thuộc tính FAIR và Dữ liệu mở). Một thuật ngữ có sự kết hợp giữa nguyên lý FAIR và Dữ liệu mở, dữ liệu FAIR/O, trong đó bên cạnh FAIR liên quan đến một số nguyên tắc về bản quyền dữ liệu mở (hay bản quyền miễn phí).

Bảng 1. Một số nguồn dữ liệu mở hữu ích có hỗ trợ nguyên lý FAIR/O (FAIR/Dữ liệu mở).

STT	Nguồn dữ liệu	Mô tả nguồn dữ liệu	Địa chỉ URL	Hỗ trợ
1	DIVA-GIS Free Spatial Data	Nguồn dữ liệu không gian miễn phí của dự án DIVA-GIS	https://www.diva-gis.org/Data	Dữ liệu mở
2	USGS Data	Dữ liệu viễn thám và Landsat của USGS	https://www.usgs.gov/products/data-and-tools/overview	FAIR/O
3	Google Earth Engine	Nền tảng quy mô toàn cầu hỗ trợ truy vấn, phân tích và dữ liệu khoa học trái đất của Google	https://earthengine.google.com/	FAIR, các giải thuật máy học
4	Google APIs Explorer	Cung cấp đa dạng các RESTful API cho phép truy vấn và xử lý dữ liệu trực tuyến liên quan đến các ứng dụng của Google	https://developers.google.com/apis-explorer	FAIR
5	Mekong River Commission (MRC)	Dự án thúc đẩy, phối hợp quản lý và phát triển bền vững nước của Ủy ban Mê Kông	https://portal.mrcmekong.org/home	FAIR/O
6	Catch-Mekong	Kho dữ liệu về giám sát môi trường và vận động của Đồng bằng sông Cửu Long và lưu vực sông Mê Kông	https://catchmekong.eoc.dlr.de/Elvis/	FAIR/O
7	USDA Scientific Data Services	Dịch vụ dữ liệu khoa học thuộc Thư viện Nông nghiệp Quốc gia (Mỹ)	https://www.nal.usda.gov/main/data	FAIR/O
8	DataSuds Dataverse (IRD)	Kho dữ liệu mở của Viện nghiên cứu Phát triển Pháp (IRD)	https://dataverse.ird.fr/	FAIR/O
9	Harvard Dataverse	Kho dữ liệu mở của Đại học Harvard (Mỹ)	https://dataverse.harvard.edu	FAIR/O
10	Open Development Mekong	Kho dữ liệu mở về tài chính/tài trợ, các tác động và các vấn đề liên quan đến phát triển cơ sở hạ tầng ở các quốc gia Tiểu vùng Mê Kông.	https://opendevelopmentmekong.net/	FAIR/O

2.3. Phân loại học, Thực thể học và Dữ liệu liên kết

Phân loại học (taxonomy) được định nghĩa là ngành học chuyên về xếp loại và phân chia các đối tượng theo cấu trúc cây. Ngành khoa học này được áp dụng rất phổ biến trong sinh học, nổi tiếng nhất có thể là Phân loại tiến hóa (Evolutionary taxonomy), trong đó các chủng loại sinh vật được phân loại theo Thuyết tiến hóa của nhà sinh học Darwin. Trên thực tế, phân loại học đã thâm nhập vào hầu hết các lĩnh vực khoa học khác, tiếp cận này

cho phép nhìn nhận các đối tượng theo hệ thống, rất hữu ích cho việc quản lý dữ liệu có tính hệ thống. Chính vì vậy, phân loại học được khuyến cáo trong nguyên lý I2 của FAIR, và trở thành một trong các tiêu chuẩn quan trọng của quản lý dữ liệu. Có thể thấy, trong nông nghiệp, phân loại học có thể áp dụng trong phân loại giống cây trồng, vật nuôi, nghiên cứu thiên địch, nghiên cứu thổ nhưỡng hay quản lý đất đai.

Trong khoa học máy tính, thực thể học/bản thể học (ontology) bao gồm cách biểu diễn, đặt tên chính thức và định nghĩa các danh mục, thuộc tính và mối quan hệ giữa các khái niệm, dữ liệu (Guarino et al., 2009). Nói một cách đơn giản hơn, thực thể học là một cách thể hiện các thuộc tính của các chủ thể và mối liên hệ giữa các chủ thể đó. Khái niệm thực thể học được đề cập ở đây vì nó liên quan trực tiếp đến khái niệm dữ liệu liên kết (Linked Data) (Berners-Lee, 2009; W3C, 2021) được sử dụng trong Web ngữ nghĩa, thuật ngữ nằm trong cả ba nguyên lý F, A và I của FAIR. Để dễ hình dung về thực thể học, chúng ta có thể xem xét cấu trúc khá phổ biến của nó trong trích dẫn của một ấn phẩm khoa học ở các hình thức phổ biến: BibTeX, RIS, RDF N-Triples, RDF/XML hay XML. Trong đó, mỗi ấn phẩm thường bao gồm các thông tin quan trọng như một mã DOI, thông tin tác giả, tiêu đề, nguồn và năm xuất bản. Quan trọng hơn, thông tin này được tái sử dụng bởi các nền tảng quản lý trích dẫn như Mendeley[1] và Zotero[2] để hỗ trợ việc lập danh mục tài liệu tham khảo trên Microsoft Word hoặc LaTeX. Cách thức tương tự như vậy được áp dụng trong nhiều nền tảng ứng dụng khác nhau để tạo thành một hệ thống thông tin nối kết mà *cơ sở chung ban đầu là chuẩn dữ liệu*. Các tập dữ liệu (dataset) hiện nay cũng cần được quản lý theo cách thức tương tự, nhằm hỗ trợ nhận dạng và tái sử dụng ở bất cứ đâu.

Bảng 2 trình bày một số nguồn dữ liệu quốc tế về sinh học, nông nghiệp và thủy sản có giá trị cao cần được tham khảo để áp dụng cho nông nghiệp Việt Nam. Các nguồn dữ liệu này được quản lý dựa trên nguyên lý phân loại học.

Bảng 2. Các nguồn dữ liệu phân loại học liên quan đến sinh học, nông nghiệp và thủy sản.

STT	Tên nguồn	Mô tả dữ liệu	Địa chỉ URL
1	Genesys	Chứa thông tin về ngân hàng gen thế giới của PGRFA (Plant Genetic Resources for Food and Agriculture)	https://www.genesys-pgr.org/
2	VEST: GODAN ACTION	Dữ liệu mở toàn cầu về nông nghiệp và dinh dưỡng	https://vest.agrisemantics.org/
3	AgroPortal	Thực thể học về nông nghiệp thực phẩm	http://agroportal.lirmm.fr/ontologies/AFEO
4	DBpedia ontology	Thực thể học đa ngành	https://www.dbpedia.org/resources/ontology/
5	Ffish.asia	Dự án ghi lại sự đa dạng sinh học cá nước ngọt của Đông Nam Á và Đông Á (Hợp tác giữa Campuchia, Trung Quốc, Lào, Mã Lai, Miến Điện, Đài Loan, Thái Lan, Việt Nam và Nhật Bản).	https://ffish.asia

Dữ liệu mở nối kết (Linking Open Data – LOD) là chủ đề rộng và cần được quan tâm nghiên cứu trong việc quy hoạch quản lý dữ liệu ở Việt Nam. Thuật ngữ này kết hợp hai thuộc tính quan trọng, gồm dữ liệu mở (Open Data) và Dữ liệu nối kết (Linked Data). Trong đó Dữ liệu mở nói đến thuộc tính chia sẻ miễn phí còn Dữ liệu nối kết liên quan đến khả năng cung cấp các chuẩn dữ liệu có khả năng tái sử dụng tự động bên trong các ứng dụng CNTT. Trong đó, các hệ thống dữ liệu như Wikipedia, Wikibooks, Geonames, MusicBrainz, WordNet, và thư viện học liệu DBLP là một trong các dự án áp dụng LOD trong quá trình tổ chức và quản lý nguồn dữ liệu khổng lồ của mình. W3C đưa ra hàng loạt các tiêu chuẩn định nghĩa các bộ từ vựng cũng như cơ chế ánh xạ từ các ứng dụng phần mềm đến các tiêu chuẩn này. Trong đó, DCAT (Data Catalog Vocabulary) là chuẩn chung dùng trong mô

[1]https://www.mendeley.com/

[2]https://www.zotero.org/

tả từ vựng cho các tập dữ liệu. Một số giao diện ứng dụng dựa trên DCAT bao gồm: GeoDCAT-AP[3] (EU ISA Programme), DDI v2 to Dublin Core[4] (DDI Alliance), ISO 19115 - DCAT - Schema.org mapping[5] (W3C SDW WG) và CiteDCAT-AP[6] (JRC).

Dữ liệu liên kết đòi hỏi phải sử dụng các tiêu chuẩn kỹ thuật. Đây là vấn đề mà Việt Nam cũng rất quan tâm và đề xuất áp dụng các tiêu chuẩn quốc tế trong những năm gần đây. Bảng 3 trích lọc một số tiêu chuẩn tiêu biểu được định nghĩa trong Thông tư 39/2017/TT-BTTTT về Ban hành Danh mục tiêu chuẩn kỹ thuật về Ứng dụng Công nghệ Thông tin trong cơ quan nhà nước (Bộ Thông tin và Truyền thông, 2017)(Thay thế cho thay thế Thông tư số 22/2013/TT-BTTTT trước đó).

Mặc dù vậy, dữ liệu mở và miễn phí chưa phải là trọng tâm thảo luận trong bài viết này, thay vào đó, chỉ các thuộc tính liên quan đến FAIR được phân tích sâu.

Bảng 3. Một số tiêu chuẩn tiêu biểu được định nghĩa trong Thông tư 39/2017/TT-BTTTT.

Mã tiêu chuẩn	Loại tiêu chuẩn	Ký hiệu tiêu chuẩn	Tên đầy đủ của tiêu chuẩn	Quy định áp dụng
1.14	Truy cập và chia sẻ dữ liệu	OData v4	Open Data Protocol version 4.0	Khuyến nghị
1.15	Dịch vụ Web dạng RESTful	RESTful web service	Representational state transfer	Khuyến nghị
2.1	Ngôn ngữ định dạng văn bản	XML v1.0 (5th Edition)	Extensible Markup Language version 1.0 (5th Edition)	Bắt buộc một trong hai tiêu chuẩn
		XML v1.1 (2th Edition)	Extensible Markup Language version 1.1	
2.6	Mô tả tài nguyên dữ liệu	RDF	Resource Description Framework	Khuyến nghị
		OWL	Web Ontology Language	Khuyến nghị
2.8	Khuôn thức trao đổi thông tin địa lý	GML v3.3	Geography Markup Language version 3.3	Bắt buộc
2.9	Truy cập và cập nhật các thông tin địa lý	WMS v1.3.0	OpenGIS Web Map Service version 1.3.0	Bắt buộc
		WFS v1.1.0	Web Feature Service version1.1.0	Bắt buộc
2.12	Bộ phần tử siêu dữ liệu Dublin Core	ISO 15836-1:2017	Dublin Core	Khuyến nghị
2.13	Định dạng trao đổi dữ liệu mô tả đối tượng dạng kịch bản JavaScript	JSON RFC 7159	JavaScript Object Notation	Khuyến nghị

3. VAI TRÒ CỦA NGUỒN DỮ LIỆU TRONG NÔNG NGHIỆP THÔNG MINH

3.1. Nguồn dữ liệu nông nghiệp

Nông nghiệp là ngành sản xuất vật chất cơ bản của xã hội, sử dụng đất đai để trồng trọt và chăn nuôi, khai thác cây trồng và vật nuôi làm tư liệu và nguyên liệu lao động chủ yếu để tạo ra lương thực, thực phẩm và một số nguyên liệu cho công nghiệp. Nông nghiệp là một ngành sản xuất lớn, bao gồm nhiều chuyên ngành: trồng trọt, chăn nuôi, sơ chế nông sản; theo nghĩa rộng, còn bao gồm cả lâm nghiệp, thủy sản. Dự thảo Quy hoạch vùng Đồng bằng Sông Cửu Long thời kỳ 2021-2030, tầm nhìn đến năm 2050 (Thực hiện theo Quyết định số 1163/QĐ-TTg của Thủ tướng Chính phủ) có thể trở thành cơ sở cho việc lập quy hoạch cấp tỉnh, quy hoạch đô thị, quy hoạch sử dụng đất, quy hoạch nông thôn, quy hoạch có tính chất kỹ thuật, chuyên ngành trên địa bàn vùng. Trong 08 chính sách liên quan đến "Công tác Quy hoạch", Bộ Kế hoạch Đầu tư đã đề cập trực tiếp đến nông nghiệp trong 05 chính sách. Trong đó, yếu tố về *phát triển bền vững, thích ứng với biến đổi khí hậu* và *"thuận thiên"* là

[3]GeoDCAT-AP
[4]DDI v2 to Dublin Core
[5]ISO 19115 - DCAT - Schema.org mapping
[6]CiteDCAT-AP

các nội hàm quan trọng được nhắc đến. Thông thường, các quy hoạch được xây dựng sẽ tác động đầu tiên và trực tiếp đến hoạt động của các cơ quan khối chính phủ, sau đó, bằng nhiều quy trình và cách thức khác nhau sẽ lan tỏa đến cộng đồng nghiên cứu học thuật và khối doanh nghiệp. Dù áp dụng quy trình và cách thức như thế nào, bài viết này hướng đến khuyến cáo rằng *cách tiếp cận ra quyết dựa trên dữ liệu là cách tiếp cận khoa học nhất và cần được chú trọng quan tâm trong quá trình hoạch định các chiến lược*, do đó, cũng cần phải chú trọng đến quy hoạch quản lý dữ liệu nông nghiệp. Theo đó, chúng tôi khuyến nghị áp dụng một Quy hoạch Quản lý Dữ liệu (DMP – Data Management Plan) trong nghiên cứu này theo theo bốn hướng tiếp cận: (1) Dữ liệu được chia sẻ từ các cơ quan khối chính phủ, cụ thể là từ các trung tâm dữ liệu cấp quốc gia, cấp vùng,cấp tỉnh, (2) Dữ liệu được chia sẻ từ cộng đồng nghiên cứu học thuật lấy từ các đề tài nghiên cứu khoa học, đặc biệt các đề tài nghiên cứu nhận tài trợ từ nguồn đầu tư công, (3) Nguồn dữ liệu trích xuất từ các kho dữ liệu quốc tế, đặc biệt là dữ liệu vệ tinh (4) Nguồn dữ liệu chia sẻ trên nguyên tắc tự nguyện giữa cộng đồng các chuyên gia, gồm 03 nhóm chuyên gia trình bày trong Mục 1.

Các dữ liệu phục vụ nông nghiệp có thể phân thành ba nhóm chính liên quan đến tính chất và cách thức tác động của chúng đến cộng đồng nghiên cứu:

Nhóm dữ liệu về điều kiện tự nhiên và quy hoạch, chẳng hạn: điều kiện tự nhiên như môi trường, khoa học đất (chẳng hạn thổ nhưỡng, độ mặn và độ phèn theo mùa), nguồn nước, khí tượng, thủy văn, biến đổi khí hậu; các quy hoạch nông nghiệp chung: đê điều, đập ngăn mặn,…; điều kiện về hệ thống giao thông, vận tải, logistics; điều kiện về năng lực hội nhập quốc tế; …

Nhóm dữ liệu về sản xuất, canh tác, chẳng hạn: năng lực cung ứng về giống cây trồng, vật nuôi,…; các tiêu chuẩn và quy trình canh tác (Chẳng hạn chuẩn GAP); năng lực cung ứng về phân bón, thuốc bảo vệ thực vật; năng lực xây dựng các khu nông nghiệp khép kín.

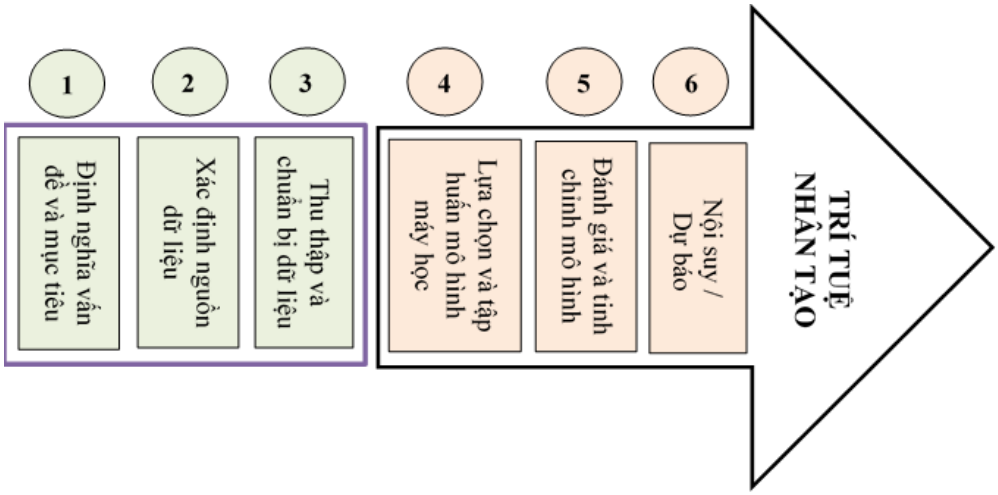
Nhóm dữ liệu về thị trường, chẳng hạn: Năng lực thu mua và xuất khẩu; Năng lực chế biến nông phẩm; Năng lực bảo quản nông sản; Năng lực định danh các nông sản, doanh nghiệp liên quan đến truy xuất nguồn gốc.

Mỗi nhóm dữ liệu nêu trên, tùy đặc thù và quy mô, sẽ có chủ sở hữu khác nhau, có chính sách bảo mật khác nhau và có ý nghĩa sử dụng khác nhau. Tuy nhiên, dù thuộc đặc thù gì, khuyến cáo trong bài viết này hướng tới việc yêu cầu dữ liệu đó phải được tổ chức theo nguyên lý FAIR (dựa trên các nền tảng các ứng dụng CNTT khác nhau), còn việc đây có thể là dữ liệu mở hay không thì không thuộc phạm vi của đề xuất của nghiên cứu này.

3.2. Nông nghiệp thông minh

Như đã trình bày bên trên, khái niệm thông minh trong nghiên cứu nông nghiệp được xây dựng trên 03 cấp độ khác nhau của phân tích dữ liệu. Trong cả 03 cấp độ phân tích này, *quy trình phân tích dữ liệu thông minh* có những nét tương đồng, bao gồm 06 bước (Hình 1); trong đó 03 bước đầu vừa đóng vai trò định hướng, vừa đóng vai trò quyết định sự thành công của toàn bộ quy trình, 03 bước còn lại phụ thuộc vào năng lực của đội ngũ chuyên gia (hoặc của các công cụ phân tích dữ liệu, bao gồm các phần mềm thuộc nhóm BI Tools như Stata, SPSS, SAS, Tableau, Power BI hoặc các ngôn ngữ chuyên phân tích dữ liệu như R hay Python).

Có thể nói, hai bước (2) và (3) trong quy trình phân tích dữ liệu thông minh là hai bước phụ thuộc nhiều vào yếu tố khách quan, trong đó phụ thuộc trực tiếp vào tính sẵn sàng của các nguồn dữ liệu. Bước (2) *Xác định nguồn dữ liệu* liên quan đến thuộc tính đầu tiên của FAIR là F (Tìm kiếm được), yêu cầu này chỉ được đáp ứng nếu nguồn dữ liệu có ít nhất hai sự hỗ trợ: (1) Có mã định danh DOI (Digital Object Identifier), URL/URI hoặc cách thức tương tự, và (2) Việc tổ chức dữ liệu đi kèm với siêu dữ liệu (metadata). Bên cạnh đó, việc hỗ trợ các công cụ tìm kiếm thông minh cho phép thực hiện các câu hỏi luận lý (AND, OR, NOT) cũng cần được chú ý (giống cách mà Google Search đang hỗ trợ). Một trong các phần mềm nguồn mở tuân thủ FAIR được khuyến cáo sử dụng là Dataverse (Datavese, 2021) có tích hợp Apache Solr (Apache Solr, 2021) bên trong nền tảng của mình để hỗ trợ cơ chế tìm kiếm thông minh này. Trong khi đó, bước (3) liên quan đến 03 thuộc tính còn lại của FAIR, đặc biệt liên quan đến các giao thức mở, chuẩn dữ liệu phổ biến (nguyên lý A1.1); các định dạng dữ liệu đa dạng có thể truy vấn như JSON, TURTLE, RDF/XML hay CSV (nguyên lý I1), khả năng sử dụng các bộ từ vựng chuẩn mực (hoặc cao hơn là sử dụng phân loại học) trong quản lý dữ liệu (nguyên lý I2); và cuối cùng, dữ liệu cần được liên kết với xuất xứ chi tiết (nguyên tắc R1.2). Một số giao thức mở phổ biến cũng được phân tích cụ thể trong Bảng 3.



Hình 1. Quy trình phân tích dữ liệu thông minh

4. ĐÁNH GIÁ HIỆN TRẠNG QUẢN LÝ DỮ LIỆU TẠI VIỆT NAM

Hiện trạng quản lý dữ liệu tại Việt Nam được xem xét đánh giá trên ba khía cạnh: (1) Liên quan đến định hướng chính sách tổ chức, quản lý và chia sẻ về dữ liệu; (2) việc vận dụng và triển khai các chính sách (liên quan đến năng lực quản lý dữ liệu); và (3) Năng lực khai thác các nguồn dữ liệu mở.

Thứ nhất, xét về định hướng chính sách tổ chức, quản lý và chia sẻ về dữ liệu, có thể nói Chính phủ Việt Nam rất quan tâm đến vấn đề quản trị nguồn dữ liệu quốc gia ở tầm vĩ mô. Có thể chia thành 02 nhóm chính sách sau:

Các hệ thống chỉ tiêu chuẩn mực và các quy định về dữ liệu đặc tả dùng chung hoặc chuyên ngành, chẳng hạn: Nghị định 97/2016/NĐ-CP về Quy định nội dung chỉ tiêu thống kê thuộc hệ thống chỉ tiêu thống kê quốc gia áp dụng từ 01/7/2016; Bộ chỉ số năng lực cạnh tranh cấp tỉnh hay PCI (viết tắt của Provincial Competitiveness Index); Chỉ số chất lượng không khí (AQI) theo Quyết định số 878/QĐ-TCMT.

Các chuẩn hỗ trợ nối kết/tương tác (Linked Data), chẳng hạn: Thông tư 39/2017/TT-BTTTT về Ban hành Danh mục tiêu chuẩn kỹ thuật về Ứng dụng CNTT trong cơ quan nhà nước; Thông tư 24/2011/TT-BTTTT về Quy định về việc tạo lập, sử dụng và lưu trữ dữ liệu đặc tả trên trang thông tin điện tử hoặc cổng thông tin điện tử của cơ quan nhà nước (Dublin Core), edXML (Quy chuẩn QCVN 102:2016/BTTTT); Thông tư 02/2021/TT-BTTTT Ban hành "Quy chuẩn kỹ thuật quốc gia về cấu trúc, định dạng dữ liệu phục vụ kết nối, tích hợp, chia sẻ dữ liệu giữa các hệ thống thông tin báo cáo trong Hệ thống thông tin báo cáo quốc gia"; Nghị định số 47/2020/NĐ-CP về quản lý, kết nối và chia sẻ dữ liệu số của cơ quan nhà nước.

Thứ hai, việc vận dụng và triển khai các chính sách (liên quan đến năng lực quản lý dữ liệu). Có thể nói đây là một hạn chế và là một rào cản rất lớn đối với cộng đồng chuyên gia, cũng chính là mục tiêu mà bài viết này muốn đề xuất tháo gỡ. Điểm sáng về chính sách gần đây là Nghị định số 47/2020/NĐ-CP về *quản lý, kết nối và chia sẻ dữ liệu số của cơ quan nhà nước* (Chính phủ, 2020), góp phần thúc đẩy một số Cổng dữ liệu mở cấp quốc gia[7], cấp tỉnh/thành như Cổng dữ liệu mở TP. Hồ Chí Minh[8] và Cổng dữ liệu mở Đà Nẵng[9]. Các tỉnh thành khác cũng đang xúc tiến các bước đi đầu tiên để thực hiện quá trình này, chẳng hạn Nghị quyết 70/NQ-HĐND về chủ trương đầu tư dự án *Xây dựng kho CSDL dùng chung tỉnh Sóc Trăng; Cổng dịch vụ dữ liệu mở của tỉnh; số hóa dữ liệu xây dựng chính quyền điện tử tỉnh Sóc Trăng* ngày 13/7/2021. Tuy nhiên, nếu so sánh với các hệ thống quản lý dữ liệu FAIR, các cổng dữ liệu mở bên trên vẫn còn nhiều hạn chế và cần được cải tiến. Các giải pháp cụ thể sẽ được đề xuất trong Mục 5.

Bên cạnh đó, phần lớn nguồn dữ liệu cấp tỉnh, hoặc cấp quốc gia hiện nay là các dữ liệu đã qua xử lý (thông thường là dữ liệu thống kê, tức là qua cấp độ Phân tích mô tả), trong khi cộng đồng nghiên cứu cần hơn nguồn dữ liệu cơ sở (hay dữ liệu thô). Điển

[7]http://data.gov.vn
[8]https://data.hochiminhcity.gov.vn/
[9]https://opendata.danang.gov.vn/

hình của dạng dữ liệu này là dữ liệu từ Tổng cục Thống kê và được nhiều Trung tâm thông tin/Trung tâm dữ liệu ở các tỉnh/thành tái chia sẻ và phân tích ở cấp địa phương (chẳng hạn Trà Vinh, Đồng Tháp và Bến Tre). Một hạn chế nữa, khi cần tiếp cận các nguồn dữ liệu cụ thể (theo hình thức thương mại), các thông tin mô tả về nội dung và cấu trúc của các tập dữ liệu này phần lớn vẫn chưa được rõ ràng. Một minh chứng cụ thể cho vấn đề này là khi tiếp cận các nguồn dữ liệu về khí tượng thủy văn. Còn lại, rất nhiều trung tâm chức năng cấp quốc gia, cấp vùng không có cơ chế chia sẻ dữ liệu chính thức cho cộng đồng nghiên cứu.

Thứ ba, liên quan đến năng lực khai thác các nguồn dữ liệu mở, đặc biệt các nguồn dữ liệu đã hỗ trợ FAIR. Như đã trình bày trong Bảng 1 và Bảng 2, các nguồn dữ liệu hỗ trợ nguyên lý FAIR đem lại rất nhiều lợi ích cho cộng đồng nghiên cứu và công nghệ, trong đó một số nguồn hỗ trợ cơ chế tìm kiếm và trích lọc thông tin rất hiệu quả. Đây là một số gợi ý để các trung tâm dữ liệu ở Việt Nam có thể học hỏi để có thể đóng góp tốt hơn vào nguồn dữ liệu đang thiếu hụt trong lĩnh vực nông nghiệp.

5. MÔ HÌNH QUY HOẠCH QUẢN LÝ DỮ LIỆU THEO NGUYÊN LÝ FAIR

Dựa trên các phân tích và đánh giá về thực trạng quản lý, chia sẻ dữ liệu tại Việt Nam nêu trên, Mô hình Quy hoạch Quản lý dữ liệu (DMP) được đề xuất nhằm áp dụng cho hai khối cơ quan chức năng liên quan đến quản lý dữ liệu nông nghiệp tại Đồng bằng Sông Cửu Long:

– Khối chính phủ: Nhóm các cơ quan thuộc khối chính phủ, bao gồm các trung tâm dữ liệu cấp quốc gia, cấp vùng hoặc cấp tỉnh/thành.

– Khối nghiên cứu và đào tạo: Nhóm các trung tâm nghiên cứu, viện nghiên cứu và các đơn vị đào tạo, đặc biệt là các đơn vị được giao trách nhiệm chính nghiên cứu về nông nghiệp (theo nghĩa rộng).

Trong mô hình đề xuất, hai nhóm công nghệ sau đây được tập trung phân tích:

Nhóm công nghệ máy chủ: Nhóm công nghệ này khó áp dụng do độ phức tạp cao, cần có lộ trình và quy hoạch trong thời gian dài. Đây là nhóm công nghệ đòi hỏi sự am hiểu sâu về CNTT, có thể được cài đặt bên trong các trung tâm dữ liệu lớn và tập trung. Đối với hệ thống này, cách thức cung cấp dữ liệu cho cộng đồng nên dựa trên việc cung cấp 03 nhóm công nghệ:

– Dịch vụ web chuẩn REST (RESTful WebServices) (Chaudhary & Bhise, 2013; Erl et al., 2012).

– Cơ chế OLAP (Online Analytical Processing) sử dụng kết hợp với Kho dữ liệu (Data Warehouse).

– Các hệ truy vấn dữ liệu Impala/Hive cho các hệ sinh thái dữ liệu lớn Apache Hadoop.

Các giải pháp nêu trên được các hệ thống dịch vụ của Google áp dụng rất thành công. Tại Đại học Cần Thơ, trong xây dựng bản đồ cơ sở hạ tầng Khu II, dịch vụ Domain-based Web Services (DWS) cũng được xây dựng theo cơ chế REST (Lê Thành Phiêu và ctv., 2020) (Hoang et al., 2021).

Nhóm công nghệ ứng dụng: Nhóm công nghệ này đơn giản hơn nên chỉ cần lập các kế hoạch ngắn hoặc trung hạn là có thể áp dụng. Đây là nhóm công nghệ đã được đơn giản hóa qua các phần mềm ứng dụng, dễ triển khai và có thể cài đặt ở các các đơn vị có quy mô từ lớn đến nhỏ. Ở giai đoạn đầu, 03 nhóm ứng dụng cụ thể được đề xuất nhằm cung cấp các dịch vụ sau:

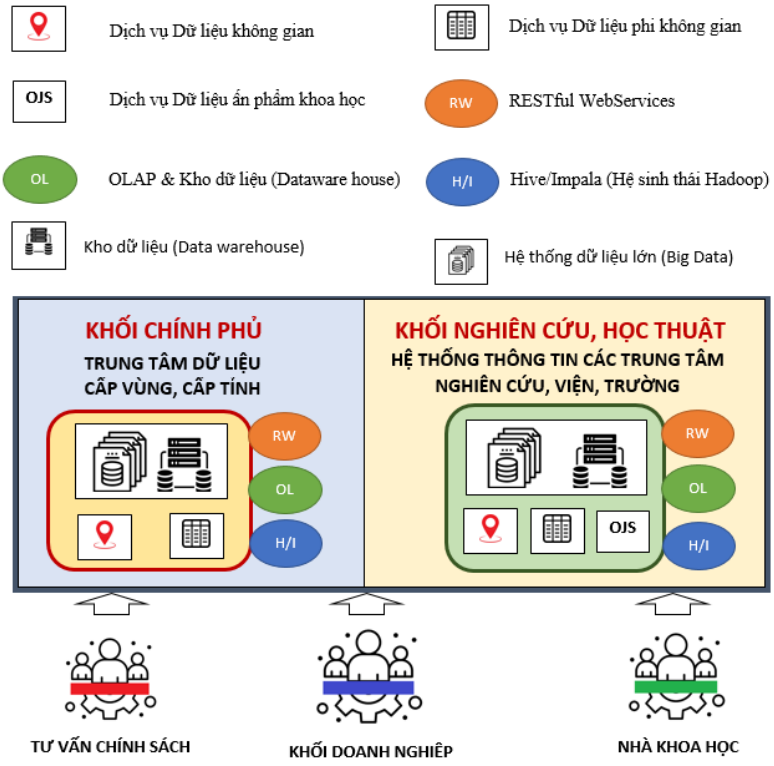
– **Dịch vụ Dữ liệu không gian (Spatial Data):** Trong nhóm này, các chuẩn dữ liệu của tổ chức OGC (OGC Standard, 2021) được khuyến cáo sử dụng. Cụ thể hơn các các dịch vụ cung cấp dữ liệu bản đồ, đặc biệt là phần mềm GeoServer (Geoserver.org, n.d.). GeoServer cũng chính thức tuyên bố tuân thủ nguyên lý FAIR trong các quy trình và chính sách quản lý dữ liệu của mình. Nhóm ứng dụng này được sử dụng khá phổ biến tại Đại học Cần Thơ và trong nhiều đề tài nghiên cứu khoa học cấp tỉnh/thành (Nguyen et al., 2014) (Nguyễn Văn Kiệt và ctv., 2011), tuy nhiên phần lớn các dịch vụ đều mang tính bảo mật, chưa chia sẻ dữ liệu cho cộng đồng.

– **Dịch vụ Dữ liệu phi không gian (Non-Spatial Data):** Đây là nhóm ứng dụng có hình thức chia sẻ dữ liệu tương tự các hệ thống DataSuds Dataverse, Harvard Dataverse, Kaggle, DMP Online, Open Development Mekong được giới thiệu trong Bảng 1. Hai nền tảng mã nguồn mở được đề xuất sử dụng là Datavese (2021) và Ckan (2021). Hiện tại, cộng đồng nghiên cứu học thuật tại Việt Nam chưa triển khai bất cứ hệ thống nào tương tự, một số hệ thống dữ liệu quốc tế đã cập nhật các nguồn dữ liệu tại Việt Nam trong phạm vi nghiên

cứu của họ, chẳng hạn DataSuds Dataverse[10] (IRD), Open Development Mekong[11]. Ở khối chính phủ, như đã phân tích, đã bắt đầu hình thành các trung tâm dữ liệu mở quốc gia đang phục vụ cho xu hướng này.

– **Dịch vụ Dữ liệu ấn phẩm khoa học (OJS)**: Ấn phẩm khoa học là yếu tố đặc biệt quan trọng trong nghiên cứu khoa học, do vậy các nguồn dữ liệu về thư viện học thuật liên quan đến ấn phẩm khoa học là rất cần thiết. Một trong các điển hình tiêu biểu là DBLP[12] và MDPI[13]. Trên thực tế, phần lớn các hội thảo, tạp chí khoa học đều được công bố rộng rãi, và phần lớn đều cho phép tải về nội dung bài viết. Tuy nhiên, hạn chế lớn nhất mà các tạp chí đang gặp phải là thiếu đi khả năng hỗ trợ trích dẫn mà các tạp chí chuẩn thường thấy, cụ thể là thiếu mục "Export Citations" cho phép trích xuất thông tin trích dẫn dưới các định dạng mở như BibTeX, RIS, RDF N-Triples, RDF/XML. Có hai cách thức để khắc phục hạn chế này, một là cải tiến phần mềm hiện có và bổ sung thêm chức năng trích xuất các thông tin trích dẫn, hai là sử dụng các nền tảng nguồn mở OJS (Open Journal System). Nền tảng nguồn mở được đề xuất trong nghiên cứu này là PKP (Public Knowledge Project) (PKP, 2021). Tại Đại học Cần Thơ, hệ thống Tạp chí Đại học Cần Thơ chưa cho phép trích xuất thông tin trích dẫn dưới các định dạng mở, tuy nhiên Trường cũng đang trên lộ trình thử nghiệm hệ thống OJS mới để hỗ trợ cơ chế này.

Hình 2 sơ đồ hóa toàn bộ giải pháp Quy hoạch Quản lý Dữ liệu (DMP) hỗ trợ nông nghiệp thông minh mà nghiên cứu này đề xuất, trong đó bao gồm các nhóm dịch vụ và ứng dụng cần phải có đối với hai Khối cơ quan: *Khối chính phủ* và *Khối nghiên cứu, học thuật*.



Hình 2. Giải pháp Quy hoạch Quản lý Dữ liệu (DMP) hỗ trợ nông nghiệp thông minh

Trong mô hình đề xuất, mặc dù *Khối nghiên cứu, học thuật* có nhiều thành phần dịch vụ hơn, cụ thể có thêm Dịch vụ Dữ liệu ấn phẩm khoa học (OJS), tuy nhiên, *Khối chính phủ* mới là đơn vị sở hữu nguồn dữ liệu phong phú và đa dạng nhất. Do vậy việc thúc đẩy triển khai DMP ở khối này là rất quan trọng và cần thiết.

Đối với việc tổ chức và phân loại dữ liệu, như đề cập trong Mục 3, cần phân chia dữ liệu thành 03

[10]https://dataverse.ird.fr/
[11]https://opendevelopmentmekong.net/
[12]https://dblp.org/
[13]https://www.mdpi.com/

nhóm độc lập, bao gồm: (1) *Nhóm dữ liệu về điều kiện tự nhiên và quy hoạch*, (2) *Nhóm dữ liệu về sản xuất, canh tác* và (3) *Nhóm dữ liệu về thị trường*. Việc phân chia này sẽ giúp việc quản lý được chặt chẽ và đồng bộ hơn. Bên cạnh đó, cơ chế nối kết và hỗ trợ qua lại giữa các nhóm dữ liệu trên cũng sẽ được chú trọng một cách đầy đủ, nhằm đảm bảo một nền nông nghiệp có nối kết và chia sẻ, tạo nền tảng ban đầu cho nông nghiệp thông minh.

Nói cách khác, Quy hoạch Quản lý Dữ Liệu (DMP) được trình bày trong Hình 2 là một giải pháp mang tính tổng thể. Trong đó, giải pháp này được trình bày theo lộ trình từ cấp độ đơn giản đến phức tạp, khuyến cáo áp dụng cho cả cho *Khối cơ quan chính phủ* và *Khối nghiên cứu, học thuật*. Nếu áp dụng quy hoạch này trên diện rộng, trong tương lai sẽ hình thành một hệ sinh thái các trung tâm và dịch vụ dữ liệu FAIR cung cấp đa dạng các nguồn dữ liệu cho cộng đồng chuyên gia, qua đó thúc đẩy sự phát triển bền vững của nền nông nghiệp nói riêng và khoa học nói chung. Trong đó, các cơ quan liên quan đến quản lý nông nghiệp, đặc biệt là các Sở Nông nghiệp và Phát triển Nông thôn, Sở Tài nguyên Môi trường, các trung tâm nghiên cứu và đào tạo đặc thù về nông nghiệp cần là các đơn vị tiên phong trong việc triển khai quy hoạch. Tiếp theo, là các cơ quan có liên quan đến nông nghiệp. Quan trọng hơn, đối với Quy hoạch vùng Đồng bằng Sông Cửu Long thời kỳ 2021-2030, tầm nhìn đến năm 2050, mối quan tâm đến quản trị dữ liệu cần là nội dung bắt buộc nếu nhà nước quyết tâm xây dựng một nền nông nghiệp thông minh, đáp ứng các tiêu chí về phát triển bền vững, thích ứng với biến đổi khí hậu và "thuận thiên".

6. KẾT LUẬN VÀ KIẾN NGHỊ

Tóm lại, bài viết hướng đến một đề xuất một Quy hoạch quản lý dữ liệu (DMP) mang tính tổng thể, áp dụng trong lĩnh vực nông nghiệp tuân thủ các nguyên lý FAIR (CÔNG BẰNG). Khái niệm nông nghiệp thông minh được đặt trong góc nhìn của khoa học dữ liệu, trong đó quy trình phân tích dữ liệu và các mức độ khác nhau của các kỹ thuật phân tích dữ liệu đóng vai trò quyết định đến mức độ thông minh của nền nông nghiệp. Trong quy trình phân tích dữ liệu, hai bước (2) và (3), *Xác định nguồn dữ liệu* và *Thu thập và chuẩn bị dữ liệu*, được xác định là yếu tố khách quan (đối với chuyên gia) ảnh hưởng đến chất lượng phân tích dữ liệu. Để đảm bảo được sự hỗ trợ tốt cho hai bước này, và để đảm bảo rằng các quyết định của các nhà quản lý nông nghiệp sẽ dựa trên dữ liệu, việc áp dụng nguyên lý FAIR là rất cần thiết.

Trong mô hình đề xuất, vai trò của việc áp dụng DMP được nhấn mạnh cho các cơ quan Khối chính phủ, đặc biệt là các cơ quan quản lý dữ liệu nông nghiệp và các lĩnh vực liên quan đến nông nghiệp (Như phân tích trong Mục 3.1). Nghị định số 47/2020/NĐ-CP về q*uản lý, kết nối và chia sẻ dữ liệu số của cơ quan nhà nước* dẫn đến việc hình thành các cổng dữ liệu mở cấp quốc gia là một dấu hiệu tích cực cho thấy sự quan tâm ở cấp chính phủ về vai trò của dữ liệu, hướng đến một chính phủ thông minh. Tất nhiên, sau nghị định trên sẽ cần thêm nhiều lộ trình thúc đẩy quá trình quản lý, kết nối và chia sẻ dữ liệu ở các cấp hành chính và lĩnh vực chuyên ngành cụ thể. Các đóng góp của bài viết về triển khai một DMP cấp vùng được hy vọng sẽ góp phần thúc đẩy xu hướng FAIR trong tương lai gần, qua đó hình thành một hệ sinh thái dữ liệu phong phú, đa dạng, thúc đẩy sự phát triển bền vững của nền nông nghiệp Đồng bằng Sông Cửu Long nói riêng, Việt Nam nói chung.

Tại Trường Đại học Cần Thơ, với quy hoạch trở thành Đại học vùng Đồng bằng Sông Cửu Long, trong đó nghiên cứu trọng điểm quốc gia về nông nghiệp, thủy sản và môi trường, cần đóng vai trò tiên phong trong việc triển khai DMP đại diện Khối nghiên cứu, học thuật. Một số kiến nghị cụ thể được đề xuất như sau:

Một là, trong kiến trúc Đại học thông minh mà Trường đang xây dựng, cần lưu ý nguyên lý FAIR trong việc cung cấp và chia sẻ dữ liệu từ Trung tâm Dữ liệu của Trường.

Hai là, Đại học Cần Thơ cần có máy chủ chính thức vận hành *Dịch vụ Dữ liệu không gian* trên nền GeoServer, bước đầu cung cấp nguồn dữ liệu không gian cho cộng đồng chuyên gia, trong đó đầu tiên là các nguồn dữ liệu lấy từ các đề tài nghiên cứu khoa học. Hiện tại, một số đơn vị đã vận hành dịch vụ này nhưng chủ yếu để phát triển các ứng dụng WebGIS (phục vụ nhu cầu cục bộ), chứ chưa phục vụ nhu cầu quản lý và chia sẻ dữ liệu.

Ba là, Đại học Cần Thơ cần xây dựng nền tảng cung cấp *Dịch vụ Dữ liệu phi không gian* sử dụng nền tảng Dataverse, đây là phần mềm nguồn mở áp dụng rất thành công tại Đại học Harvard (Hoa Kỳ) và Viện nghiên cứu Phát triển Pháp (IRD, Cộng hòa Pháp).

Bốn là, trong hệ thống OJS hiện đang xây dựng, cần hết sức chú trọng cơ chế trích xuất thông tin trích dẫn dưới các định dạng mở.

TÀI LIỆU THAM KHẢO

Berners-Lee, T. (2009). *Linked Data*. https://www.w3.org/DesignIssues/LinkedData.html

Bộ Nông nghiệp & Phát triển Nông thôn. (2016). *Giấy phép 28/2016/TT-BNNPTNT về Ban hành Danh mục bổ sung giống cây trồng được phép sản xuất kinh doanh ở Việt Nam*.

Burnette, M., Williams, S., Imker, H. (2016). From Plan to Action: Successful Data Management Plan Implementation in a Multidisciplinary Project. *Journal of EScience Librarianship, 5*(1), e1101. https://doi.org/10.7191/jeslib.2016.1101

Chaudhary, S., Bhise, M. (2013). *RESTful Services for Agricultural Recommendation System*.

Ckan. (2021, July 31). Ckan data repository software. https://ckan.org/

Chính phủ. (2020). *Nghị định số 47/2020/NĐ-CP về quản lý, kết nối và chia sẻ dữ liệu số của cơ quan nhà nước*.

Datavault. (2021, July 31). *What is Data Stewardship - Infographic*. https://www.data-vault.co.uk/what-is-data-stewardship-infographic/

Datavese. (2021, July 31). *Dataverse data repository software*. https://dataverse.org/

Erl, T., Carlyle, B., Pautasso, C., and Balasubramanian, R. (2012). *SOA with REST: Principles, Patterns & Constraints for Building Enterprise Solutions with REST*.

Geoserver.org. (2021, July 31). *GeoServer - Sharing Geospatial Data*. http://geoserver.org/

Guarino, N., Oberle, D., Staab, S. (2009). What Is an Ontology? In *Handbook on Ontologies* (pp. 1–17). https://doi.org/10.1007/978-3-540-92673-3_0

Hijmans, R. (2016). *Free Spatial Data | DIVA-GIS*. http://www.diva-gis.org/Data

Nguyen, H. V., Le, T. P., Ong, T. M. L., Cu, V. L., & Truong, V. X. (2021). Toward a Novel Architecture of Smart Campuses based on Spatial Data Infrastructure and Distributed Ontology. *(IntelliSys} 2021: Intelligent Systems Conference - 7th Virtual International Conference, Amsterdam, The Netherlands, September 2-3, 2021. Proceedings*.

James R. Evans, C. H. L. (2012). *Business Analytics: The Next Frontier for Decision Sciences*. http://faculty.cbpp.uaa.alaska.edu/afef/business_analytics.htm

Lamprecht, A.-L., Garcia, L., Kuzak, M., Martinez, C., Arcila, R., del Pico, E. M., Angel, V., van de Sandt, S., Ison, J., Martinez, P., McQuilton, P., Valencia, A., Harrow, J., Psomopoulos, F., Gelpi, J., Hong, N. P. C., Goble, C., and Capella-Gutiérrez, S. (2020). Towards FAIR principles for research software. *Data Sci., 3*, 37–59.

Lê Thành Phiêu, Phạm Thành Le, Nguyễn Hoàng Việt, Vũ Ánh Nguyệt, Trương Xuân Việt, Ông Thị Mỹ Linh, Phan Huy Phương, Võ Ngọc Giàu, Biện Công Nhật Trường, Trần Thị Phượng (2020). Quản lý dữ liệu không gian trong các hệ thống thông tin nền web: các vấn đề phát sinh và giải pháp chuẩn hóa. *Tạp Chí Khoa Học Trường Đại Học Cần Thơ, 56*(6A), 9–21. https://doi.org/10.22144/ctu.jvn.2020.139

Lisowska, B. (2016). *How can Data Catalog Vocabulary (DCAT) be used to address the needs of databases?* https://www.w3.org/2016/11/sdsvoc/SDSVoc16_paper_31

Nguyen, H. T., Nguyen, M., Cook, S., Moglia, M., Neumann, L. (2014). Approach for Climate Adaptation & Sustainable Development of Urban Water Systems – A Case Study in Can Tho City, Vietnam. *Journal of Science and Technology, Vietnam Academy of Science and Technology, 52* (3A), 343–348.

Nguyễn Văn Kiệt, Trương Xuân Việt, Lê Quyết Thắng (2011). Xây dựng Dịch vụ Bản đồ Tương tác với các WebServices dựa trên Kiến trúc SOA. *Kỷ Yếu Hội Nghị Tổng Kết 5 Năm Nghiên Cứu Khoa Học và Đào Tạo Khoa CNTT&TT− Đại Học Cần Thơ, Tháng 12/2011, Cần Thơ. Đại Học Cần Thơ*, 67–76.

OGC Standard. (2021, July 31). *Open Geospatial Consortium (OGC)*. https://www.ogc.org/standards

PKP. (2021, July 31). *PKP Open Journal Systems. PKP*. https://pkp.sfu.ca/

Schreinemachers, P., Schad, I., Tipraqsa, P., Williams, P. M., Neef, A., Riwthong, S., Sangchan, W., and Grovermann, C. (2012). Can public GAP standards reduce agricultural pesticide use? The case of fruit and vegetable farming in northern Thailand. *Agriculture and Human Values, 29*(4), 519–529. https://doi.org/10.1007/s10460-012-9378-6

Apache Solr. (2021, July 31). *Apache Solr enterprise search platform*. https://solr.apache.org/

Teperek, M., Cruz, M., Verbakel, E., Böhmer, J., and Dunning, A. (2018). *Data Stewardship – addressing disciplinary data management needs*. TU Delft Library. https://doi.org/10.31219/osf.io/5w9pj

Bộ Thông tin và Truyền thông. (2017). *Thông tư 39/2017/TT-BTTTT về Ban hành Danh mục tiêu chuẩn kỹ thuật về Ứng dụng Công nghệ Thông tin trong cơ quan nhà nước*.

W3C. (2021, July 31). *What is Linked Data?, from https://www.w3.org/standards/semanticweb/data*

Steve, S. (2009). The Data Governance Imperative. IT Governance Publishing. doi:10.2307/j.ctt5hh6sb

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific data, *3*(1), 1-9.