

XÂY DỰNG TỪ ĐIỂN DỮ LIỆU LIÊN KẾT MỞ DÙNG LÀM NỀN TẢNG CHO CHUYỂN ĐỔI SỐ CHÍNH PHỦ Ở VIỆT NAM

TS Tạ Tuấn Anh

Công ty cổ phần Phát triển nguồn mở và Dịch vụ FDS

• **Tóm tắt:** Từ điển dữ liệu là tri thức dùng chung để phục vụ việc kết nối chia sẻ dữ liệu giữa các hệ thống thông tin. Mô hình dữ liệu liên kết mở hiện đang là xu hướng phát triển hiện đại dùng trong thiết kế và xây dựng các cơ sở dữ liệu có quy mô toàn cầu. Bài báo trình bày các nguyên tắc cơ bản và phương pháp phát triển một từ điển dữ liệu liên kết mở tạo ra sự liên thông về ngữ nghĩa cho một cơ sở hạ tầng dữ liệu mở được áp dụng vào quá trình chuyển đổi số chính phủ ở Việt Nam.

• **Từ khóa:** Hạ tầng dữ liệu mở; dữ liệu liên kết; liên thông dữ liệu; chính phủ điện tử.

DEVELOPING AN OPEN LINKED DATA VOCABULARY AS A FOUNDATION FOR DIGITAL GOVERNMENT TRANSFORMATION IN VIETNAM

• **Abstract:** A data vocabulary is defined as a semantic schema for data interoperability between information systems. Open linked data is using as the semantic model for designing and developing new databases on a global scale. This paper presents the principle and method of building an open linked data vocabulary to create semantic interoperability for an open data infrastructure, which would be deployed for digital government transformation in Vietnam.

• **Keywords:** Open data infrastructure; linked data; data interoperability; e-government.

1. CƠ SỞ HẠ TẦNG DỮ LIỆU MỞ TRONG CHUYỂN ĐỔI SỐ

Dữ liệu là nguồn nguyên liệu chính để tạo ra tri thức trong kỷ nguyên của chuyển đổi số. Xây dựng hệ sinh thái dữ liệu mở là một trong các chiến lược quan trọng của chuyển đổi số tại mỗi quốc gia trên thế giới [3]. Đi cùng với hạ tầng thiết bị về công nghệ thông tin, cơ sở hạ tầng về dữ liệu cũng đóng vai trò thiết yếu để thúc đẩy phát triển kinh tế - xã hội.

Tính mở của cơ sở hạ tầng dữ liệu được thể hiện ở mức độ sẵn sàng chia sẻ và tái sử dụng của các loại dữ liệu trên môi trường không gian mạng. Trước đây, dữ liệu thường chỉ được xây dựng cho mục đích sử dụng cục bộ trong một tổ chức. Ngay cả trong khu vực công, các cơ sở dữ liệu (CSDL) được tạo ra bởi các cơ quan nhà nước khác nhau cũng không được kết nối liên thông, chia sẻ với nhau. Chính vì vậy, nó tạo ra sự phân mảnh dữ liệu, thiếu tính đồng bộ, nhất quán về thông tin quản lý giữa các cơ quan trên phạm vi cả nước. Ngoài ra, cách tiếp cận xây dựng dữ liệu không đồng bộ như trên

cũng tạo ra sự lãng phí nguồn lực của xã hội khi cùng một dữ liệu có thể được thu thập xử lý trong nhiều dự án khác nhau hoặc cần phải chuyển đổi, tích hợp nếu muốn được tái sử dụng. Xây dựng một hạ tầng dữ liệu mở, có sự đồng bộ trong việc kết nối liên thông, chia sẻ dữ liệu giữa các tổ chức sẽ là một giải pháp căn bản của chuyển đổi số để hình thành một hệ sinh thái phát triển dựa trên dữ liệu.

Bộ quy tắc FAIR (Findable-Accessible-Interoperable-Reusable) [6] hiện đang được sử dụng như là các tiêu chí phổ quát dùng để đánh giá chất lượng của các nguồn dữ liệu dùng trong chuyển đổi số. Nó đưa ra các yêu cầu dữ liệu phải có để có thể dễ dàng tìm thấy, truy cập, tương hợp và tái sử dụng bởi cả con người và máy tính, cụ thể như sau:

• Khả năng tìm thấy (Findable): F1 - sử dụng định danh toàn cầu và vĩnh viễn cho dữ liệu và siêu dữ liệu; F2 - dữ liệu phải được mô tả đầy đủ với các thuộc tính siêu dữ liệu; F3 - siêu dữ liệu phải chứa tham

chiều tường minh tới định danh duy nhất của dữ liệu mà nó mô tả; F4 - dữ liệu và siêu dữ liệu được đăng ký và đánh chỉ mục trong một kho tìm kiếm.

- Khả năng truy cập (Accessible): A1 - có thể truy xuất dữ liệu và siêu dữ liệu thông qua một giao thức tiêu chuẩn; A2 - siêu dữ liệu vẫn phải có khả năng truy cập được ngay cả khi dữ liệu không còn tồn tại nữa.

- Khả năng tương hợp (Interoperable): I1 - sử dụng ngôn ngữ máy hiểu để biểu diễn dữ liệu và siêu dữ liệu; I2 - khai thác các từ điển thuật ngữ dùng chung tuân thủ bộ nguyên tắc FAIR; I3 - có thể chứa tham chiếu tới các bộ dữ liệu khác.

- Khả năng tái sử dụng (Reusable): R1 - xuất bản dữ liệu và siêu dữ liệu đi kèm với giấy phép truy cập mở; R2 - có mô tả chi tiết về nguồn cung cấp dữ liệu; R3 - thỏa mãn các tiêu chuẩn ngành của lĩnh vực áp dụng.

Theo đề xuất của Tim Berners-Lee, có thể thực hiện triển khai hạ tầng dữ liệu mở theo 5 bước đáp ứng các tiêu chí của bộ quy tắc FAIR như sau:

- Bước 1 - Cấp phép mở (Open License): chia sẻ dữ liệu (dưới định dạng bất kỳ kể cả dùng pdf hoặc html) để có thể truy cập

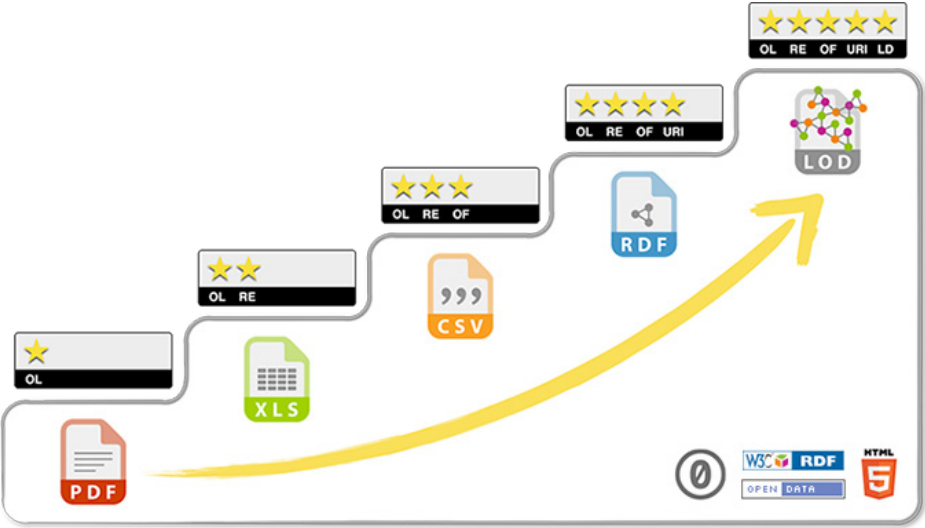
được trên internet và được cấp giấy phép truy cập mở.

- Bước 2 - Máy đọc được (Machine Readable): dữ liệu được chia sẻ dưới định dạng mà máy có thể đọc và xử lý được nội dung của nó mang theo.

- Bước 3 - Định dạng mở (Open Format): dữ liệu được chia sẻ dưới các định dạng theo tiêu chuẩn mở (không bị phụ thuộc vào chỉ một nhà cung cấp dịch vụ phần mềm ứng dụng).

- Bước 4 - Định danh URI (Uniform Resource Identifier): sử dụng các mã định danh toàn cầu URI (Uniform Resource Identification) để mô tả dữ liệu và siêu dữ liệu. Trong trường hợp này, dữ liệu cần phải được mô hình hóa theo một chuẩn được khuyến cáo bởi tổ chức W3C.

- Bước 5 - Dữ liệu liên kết (Linked Data): là cấp độ cao nhất thỏa mãn đủ các tiêu chí của tiêu chuẩn FAIR; nó cho phép các bộ dữ liệu có thể tham chiếu lẫn nhau thông qua các thuật ngữ dùng chung được định nghĩa dưới dạng của một từ điển dữ liệu. Đây là đặc điểm quan trọng nhất bởi nó cho phép dữ liệu được tạo ra trong một tổ chức có thể tham chiếu tới dữ liệu được tạo ra bởi một tổ chức khác (nghĩa là, không gian của dữ liệu sẽ không bị hạn chế ở trong một tổ chức).



Bảng 1. Mô hình 5 bước phát triển hạ tầng dữ liệu mở (Nguồn. 5stardata.info)

Phát triển cơ sở hạ tầng dữ liệu mở trong khu vực công là một trong những chính sách ưu tiên cho chuyển đổi số của các nước trên thế giới. Tổ chức Hợp tác và Phát triển kinh tế (OECD) thực hiện khảo sát đánh giá về mức độ phát triển về hạ tầng dữ liệu mở của các nước thành viên thông qua bộ chỉ số có tên là OURData (Open, Useful and Re-usable Data) Index [5]. Bộ chỉ số thực hiện đánh giá ở trên 3 khía cạnh chính là: khả năng sẵn có của dữ liệu được chia sẻ (data availability); khả năng truy cập và tính hiệu dụng của dữ liệu (data accessibility); và mức độ hỗ trợ của chính phủ trong việc tái sử dụng dữ liệu (data reusable). Theo kết quả đánh giá năm 2019, 5 nước đứng đầu trong bảng xếp hạng lần lượt là: Hàn Quốc, Pháp, Ireland, Nhật Bản và Canada.

Hiện nay, phần lớn dữ liệu mở ở các nước được cung cấp phổ biến theo các cấu trúc đã được chuẩn hóa với định dạng mở như: CSV, XML hoặc JSON (cấp độ 3). Tuy nhiên, trong khuyến cáo xây dựng dữ liệu mở của các nước đều đang hướng tới việc phải chia sẻ dữ liệu dạng có liên kết (cấp độ 5). Khi đó, dữ liệu không chỉ được cung cấp với định dạng mở mà còn phải được mô hình hóa với các định danh toàn cầu (URI) và sử dụng các từ vựng có ngữ nghĩa (dạng ontology) để mô tả dữ liệu.

2. TẠI SAO CẦN CÓ TỪ ĐIỂN DỮ LIỆU LIÊN KẾT MỞ?

Việt Nam vẫn còn đang ở trong giai đoạn đầu của việc xây dựng hạ tầng dữ liệu mở. Chính sách về kết nối và chia sẻ dữ liệu trong các cơ quan nhà nước mới được cụ thể hóa gần đây thông qua Nghị định số 47/2020/NĐ-CP của Chính phủ. Các đề án, dự án về phát triển ứng dụng công nghệ thông tin trong thực tế cũng đã hướng tới việc xây dựng CSDL dùng chung ở các quy mô khác nhau (quốc gia, bộ ngành, địa phương). Ví dụ, đề án xây dựng đô thị thông minh của thành phố Hồ Chí Minh đã chỉ rõ một mục tiêu tạo lập kho dữ liệu dùng chung và phát triển hệ sinh thái dữ liệu mở [2]. Cách tiếp cận của đề án là tích hợp các CSDL hiện hữu nằm rải rác tại các sở, ban, ngành, quận

huyện thành kho dữ liệu dùng chung của thành phố. Điều này giúp chia sẻ thông tin giữa tất cả các sở, ban, ngành, quận, huyện, người dân và doanh nghiệp có nhu cầu khai thác thông tin từ CSDL dùng chung này. Ở cấp độ quốc gia, có thể khai thác dữ liệu dùng chung từ các CSDL quốc gia về dân cư, đăng ký doanh nghiệp, đất đai,...

Để có thể kết nối, chia sẻ dữ liệu giữa các hệ thống thông tin thì dữ liệu cần phải được chuẩn bị ở mức tối thiểu đạt ở cấp độ 3, sử dụng một cấu trúc tiêu chuẩn dưới định dạng mở. Cấu trúc định dạng mở này thường chỉ được tiêu chuẩn hóa cho từng loại hình ứng dụng nghiệp vụ cụ thể. Chẳng hạn, cấu trúc gói tin dùng để trao đổi, chia sẻ dữ liệu quản lý văn bản được quy định trong quy chuẩn QCVN 102:2016/BTTTT. Cấu trúc dữ liệu về công dân được quy định trong quy chuẩn QCVN 109:2017/BTTTT. Cả hai quy chuẩn này đều sử dụng định dạng mở XML để định nghĩa lược đồ và mã hóa dữ liệu. Gần đây, quy chuẩn QCVN 120:2019/BTTTT mới được ban hành để quy định cấu trúc, định dạng dữ liệu gói tin phục vụ kết nối cổng dịch vụ công quốc gia với các hệ thống thông tin và CSDL khác trong hệ thống chính phủ điện tử. Quy chuẩn này cung cấp đặc tả mô hình dữ liệu mức logic cho các thông tin về hồ sơ, thủ tục hành chính, phản ánh kiến nghị, hỏi đáp trong lĩnh vực giải quyết dịch vụ công. Dữ liệu trao đổi có thể được mã hóa theo hai lựa chọn dùng chuẩn định dạng mở XML hoặc JSON.

Hiện nay, phần lớn các CSDL được xây dựng trên nhu cầu thực tiễn cụ thể tại đơn vị sử dụng và thường không được chuẩn hóa để chia sẻ cho các đơn vị bên ngoài. Ngay cả trong trường hợp dữ liệu chia sẻ được tiêu chuẩn hóa như các ví dụ ở trên thì cũng chưa đạt được ở cấp độ mở cao nhất theo phân loại của cơ sở hạ tầng dữ liệu mở. Chưa có sự thống nhất về sử dụng mã định danh URI và từ điển các thuật ngữ dùng trong mô tả dữ liệu. Từ đó sẽ gây ra rất nhiều khó khăn cho các đơn vị cần tích hợp khai thác dữ liệu từ nhiều nguồn, lĩnh vực khác nhau để phục vụ nhu cầu công việc, cụ thể như sau:

- Không có một mô hình dữ liệu thống nhất ở mức logic và vật lý cho các nguồn dữ liệu khác nhau. Người dùng sẽ phải xây dựng các ánh xạ dữ liệu khi cần tích hợp từ nhiều nguồn;

- Sử dụng nhiều loại từ vựng, ngôn ngữ khác nhau (Ví dụ: tiếng Việt, tiếng Anh) trong mô tả dữ liệu. Người dùng sẽ gặp nhiều trở ngại trong việc tiếp cận và hiểu dữ liệu;

- Thiếu sự nhất quán trong việc sử dụng các dữ liệu tham chiếu dùng chung. Người dùng sẽ phải thực hiện chuyển đổi, làm sạch dữ liệu về sử dụng cùng một bộ mã danh mục thống nhất;

- Dữ liệu từ nhiều nguồn không cùng sử dụng một mã định danh cho hai đối tượng dữ liệu giống nhau. Do đó, người dùng phải xây dựng các thuật toán phân tích dữ liệu để phát hiện ra sự trùng lặp của các đối tượng trên các CSDL khác nhau.

Chuyển đổi số có thể giúp khắc phục được các hạn chế nêu trên bằng phương pháp chuẩn hóa dữ liệu theo mô hình liên kết (linked data). Đây là mô hình dựa trên cấu trúc đồ thị RDF (Resource Description Framework) được sử dụng làm nền tảng dữ liệu của web ngữ nghĩa. Tất cả các từ vựng được dùng để định nghĩa các lớp, thuộc tính mô tả đối tượng dữ liệu đều phải được định danh duy nhất bằng URI để tránh được sự nhập nhằng về mặt ngữ nghĩa của dữ liệu. Bản thân đối tượng dữ liệu cũng được định danh duy nhất bằng URI nên tránh được sự trùng lặp khi trao đổi thông tin giữa các hệ thống. Chuẩn hóa dữ liệu để hướng tới cấp độ mở thứ 5 cũng sẽ là cách tiếp cận để tích hợp các nguồn dữ liệu đang sẵn có bằng cách thực hiện chuyển đổi dữ liệu theo cấu trúc của lược đồ cũ sang lược đồ dữ liệu liên kết.

Tại Việt Nam, các nguồn CSDL sẵn sàng để chia sẻ hiện có chưa nhiều. Do đó, chúng ta có lợi thế đi sau là có thể xây dựng mới các CSDL đáp ứng ngay chuẩn mô hình dữ liệu liên kết. Khi đó, chúng ta sẽ tiết kiệm được rất nhiều chi phí để thực hiện chuyển đổi, tích hợp các hệ thống nhằm đáp ứng đạt chuẩn cấp độ 5 của hạ tầng dữ liệu mở. Đây là xu thế không thể đảo ngược của tiến trình

chuyển đổi số đang diễn ra tại tất cả các nước trong đó có Việt Nam.

Xây dựng từ điển dữ liệu liên kết mở là quá trình thiết kế các từ vựng được định danh bằng URI và được dùng để mô hình hóa lược đồ ngữ nghĩa của dữ liệu. Lược đồ này được xây dựng dựa theo mô hình của ontology. Do đó, nó tạo ra sự thống nhất về mô hình dữ liệu ở mức logic được chia sẻ dùng chung giữa các CSDL. Tuy nhiên, từng CSDL có thể lựa chọn mô hình dữ liệu ở mức vật lý khác nhau để thực thi việc lưu trữ. Từ đó, định dạng dữ liệu dùng trong các gói tin trao đổi giữa các hệ thống thông tin cũng có thể sử dụng nhiều chuẩn biểu diễn khác nhau của mô hình dữ liệu liên kết, ví dụ như: RDF/XML, JSON-LD, RDFa, Turtle.

3. PHƯƠNG PHÁP XÂY DỰNG TỪ ĐIỂN DỮ LIỆU LIÊN KẾT MỞ

Trong kiến trúc thông tin, từ điển dữ liệu có vai trò như là một lớp nền tảng bảo đảm tính sẵn sàng của việc chia sẻ dữ liệu [4]. Từ lâu, chúng ta đã biết sử dụng danh mục các từ vựng có kiểm soát để tham chiếu trong các CSDL. Dữ liệu danh mục có thể là danh sách các từ khóa (thesaurus) hoặc bảng phân loại (taxonomy) được thống nhất dùng chung để quản lý thông tin trên các hệ thống khác nhau. Tiếp theo, từ điển dữ liệu được dùng để định nghĩa thống nhất các từ vựng cho phép mô tả siêu dữ liệu (metadata). Ví dụ như, bộ từ vựng Dublin Core bao gồm các trường thuộc tính cơ bản để mô tả thông tin chỉ mục của các tài liệu. Hiện nay, có rất nhiều bộ từ vựng dùng cho siêu dữ liệu được xây dựng để dùng trong các ngành, lĩnh vực khác nhau.

Ngoài ra, từ điển dữ liệu còn là công cụ dùng để đăng ký cấu trúc của lược đồ dữ liệu. Trong mô hình dữ liệu liên kết cấu trúc này được thể hiện bằng các từ vựng có định danh bằng URI. Có 3 dạng từ vựng cần phải định nghĩa trong mô hình dữ liệu liên kết là:

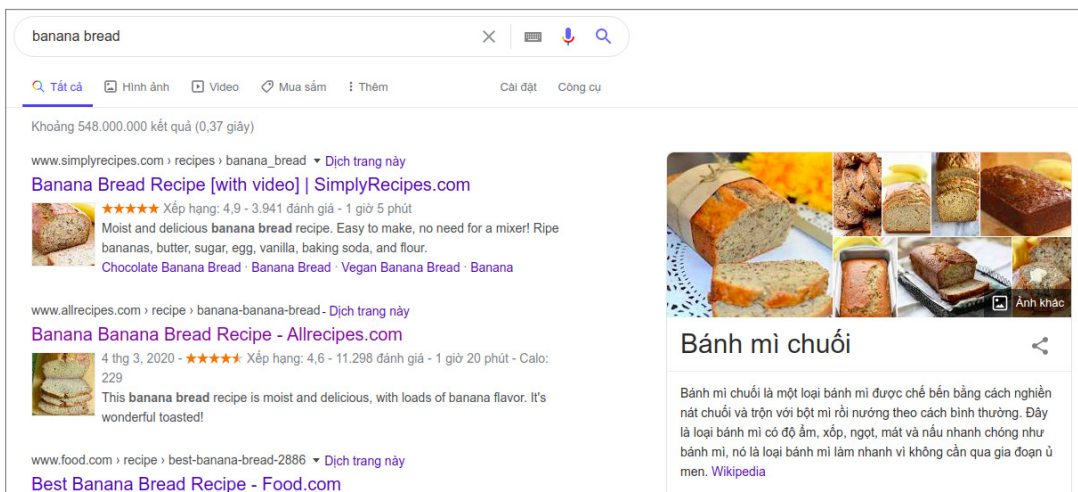
- Từ vựng lớp dữ liệu định nghĩa các kiểu đối tượng, ví dụ như: con người, tổ chức, địa điểm,...;

- Từ vựng thuộc tính dữ liệu định nghĩa các trường thông tin mô tả đối tượng, ví dụ như: tên gọi, năm sinh,...;

- Từ vựng thể hiện các giá trị phản ánh đối tượng cụ thể, ví dụ như: “nam”, “nữ” là các giá trị thể hiện nằm trong lớp dữ liệu mô tả về giới tính con người.

Mô hình dữ liệu liên kết đã được đưa vào thực tiễn áp dụng trong các máy tìm kiếm trên internet. Các trang web thông tin, ngoài

dữ liệu dùng để hiển thị dưới dạng HTML, có thể nhúng thêm dữ liệu có cấu trúc để mô tả ngữ nghĩa cho nội dung của nó. Định dạng sử dụng cho loại dữ liệu có cấu trúc này có thể là JSON-LD, RDFa hoặc Microdata. Nguồn dữ liệu có cấu trúc này giúp máy tìm kiếm có thể hiểu rõ và chính xác hơn thông tin có trên web. Từ đó, các kết quả tìm kiếm cũng sẽ được hiển thị với các thông tin có cấu trúc hơn chứ không còn dừng ở mức độ tìm kiếm các từ khóa.



Hình 2. Ví dụ tìm kiếm trên Google có kết quả được hiển thị với thông tin có cấu trúc

Ví dụ, khi gõ từ khóa “banana bread” trên trang tìm kiếm của Google thì kết quả nhận được sẽ là một bảng thông tin tổng hợp hiển thị ở phía bên phải của trang để mô tả về món bánh mì chuối. Đây là thông tin hoàn toàn có cấu trúc thể hiện được các giá trị, thành phần dinh dưỡng của thực phẩm. Danh sách kết quả tìm kiếm lúc này sẽ là các trang web nói về công thức nấu ăn của bánh mì chuối. Các kết quả được hiển thị theo một cấu trúc đặc biệt thể hiện cho món ăn gồm hình ảnh, đánh giá xếp hạng, lượng calo. Tất cả những thông tin này đã được máy tìm kiếm trích rút tự động từ dữ liệu có cấu trúc được nhúng kèm trong trang web.

Sử dụng dữ liệu liên kết sẽ làm cho các máy tìm kiếm trở nên thông minh hơn và có thể thực hiện chức năng của “trợ lý ảo” hỗ trợ hỏi đáp với người dùng. Ví dụ, nếu bạn gõ từ khóa “banana bread recipe” thì sẽ nhận ngay được kết quả là công thức hướng

dẫn thực hiện món ăn bánh mì chuối hiển thị ngay trên công cụ tìm kiếm mà không cần phải truy cập vào để xem chi tiết nội dung của trang web. Đây chính là một tính năng quan trọng nhằm hướng tới xây dựng một thế giới web có ngữ nghĩa.

Để các máy tìm kiếm có thể hiểu được ngữ nghĩa dữ liệu có cấu trúc trong các trang web thì cần phải sử dụng một từ điển dữ liệu mở thống nhất khi xuất bản các nội dung. Do đó, các nhà cung cấp dịch vụ lớn về tìm kiếm trên internet, bao gồm: Google, Bing, Yahoo, Yandex,... đã hợp tác cùng phát triển dự án schema.org. Mục tiêu của dự án là thiết lập bộ các từ vựng theo mô hình dữ liệu liên kết được sử dụng để mô tả các nội dung xuất bản trên web. Cách tiếp cận trong dự án là xây dựng một bộ từ vựng mới hoàn toàn từ các kiểu dữ liệu cơ bản nhất (vd., Text, Number, Date,...) cho đến các kiểu đối tượng thông tin thường xuất hiện trên web

(vd., Person, Organization, Place,...). Tổng cộng hiện nay, bộ từ vựng schema.org đã có tất cả 818 kiểu dữ liệu, 1326 thuộc tính, và 289 giá trị kiểu danh mục [7].

Để thiết lập chuẩn dữ liệu trao đổi trong phát triển chính phủ mở, dự án Popolo [8] lại đi theo cách tiếp cận sử dụng kế thừa từ vựng (URI) đã được tiêu chuẩn hóa ở nhiều dự án khác nhau trên internet. Dự án chỉ định nghĩa các từ vựng có ý nghĩa sử dụng mới mà không tìm thấy được sự tương đương từ các bộ từ vựng sẵn có. Sau đây là một số ví dụ về các bộ từ vựng thông dụng đã được tái sử dụng trong dự án, như: FOAF dùng để mô hình hóa thông tin các cá nhân và tổ chức; SKOS mô hình hóa các dữ liệu danh mục dùng chung; GeoNames mô hình hóa dữ liệu địa lý; DCMI Metadata mô hình hóa siêu dữ liệu; ... Tất cả các bộ từ vựng kế thừa đều tuân thủ đúng định dạng của chuẩn mô hình dữ liệu liên kết.

Xây dựng cơ sở hạ tầng dữ liệu mở ở Việt Nam, nhất là trong khu vực công, sẽ có rất nhiều tính đặc thù về nghiệp vụ theo yêu cầu quản lý riêng. Chính vì vậy, việc tìm kiếm, tái sử dụng các bộ từ vựng chuẩn trên internet để có thể đáp ứng được yêu cầu trong thực tiễn sẽ gặp rất nhiều khó khăn. Cách tiếp cận khả thi nhất là chúng ta sẽ thiết kế một bộ từ vựng hoàn toàn mới để áp dụng cho việc chia sẻ dữ liệu mở giữa các tổ chức ở Việt Nam. Quá trình xây dựng từ điển này sẽ có sự tham khảo từ các bộ từ vựng sẵn có trên thế giới và tạo ra một ảnh xạ tương đương hoặc gần giống giữa các khái niệm được sử dụng.

4. ỨNG DỤNG TỪ ĐIỂN DỮ LIỆU LIÊN KẾT MỞ TRONG PHÁT TRIỂN CHÍNH PHỦ ĐIỆN TỬ

Chuyển đổi số trong phát triển chính phủ điện tử để hướng tới một chính phủ số là chiến lược ưu tiên hiện nay ở các nước. Chính phủ số sẽ lấy khách thể (công dân, doanh nghiệp) làm trung tâm cho các kiến tạo để hình thành nên các chính sách mới. Toàn bộ điều hành của chính phủ sẽ được dẫn dắt bởi dữ liệu. Do đó, việc xây dựng hạ tầng dữ liệu mở có tầm quan trọng rất lớn để dữ liệu có thể sẵn sàng được chia sẻ dựa

trên sự thống nhất của một từ điển dữ liệu liên kết.

Phát triển từ điển dữ liệu mở dùng cho toàn bộ khối chính phủ sẽ bao trùm lên rất nhiều miền lĩnh vực nghiệp vụ. Đây là công việc đòi hỏi các kỹ năng thiết kế kiến trúc dữ liệu, đồng thời sử dụng nhiều tri thức sâu về các lĩnh vực chuyên ngành. Trong quá trình xây dựng từ điển, cần quan tâm tham khảo vận dụng các lược đồ dữ liệu mở đang được phổ biến áp dụng trên thế giới. Ví dụ, theo tài liệu [1], tổng hợp thông tin về một số tiêu chuẩn dữ liệu mở được dùng trong các ngành, lĩnh vực như sau:

- Kế toán và chống tham nhũng: dữ liệu của tổ chức minh bạch các tài trợ quốc tế (iatistandard.org), dữ liệu đấu thầu và hợp đồng (open-contracting.org).
- Tài chính công: cổng dữ liệu ngân sách và chi tiêu công (openspending.org), chuẩn dữ liệu liên kết mở tài chính công (openbudgets.eu).
- Đăng ký doanh nghiệp: tích hợp dữ liệu về thông tin doanh nghiệp (opencorporates.com), thông tin sở hữu doanh nghiệp (openownership.org).
- Tài trợ và hỗ trợ nhân đạo: chuẩn dữ liệu của tổ chức minh bạch các tài trợ quốc tế nhân đạo (hxlstandard.org).
- Môi trường: dữ liệu giám sát chất lượng không khí (openaq.org), dữ liệu giám sát chất lượng nguồn nước (gemstat.org), chuẩn dữ liệu cho đa dạng sinh học (tdwg.org).
- Tài nguyên: chuẩn dữ liệu của tổ chức minh bạch công nghiệp khai khoáng (eiti.org).
- Thông tin địa lý: dữ liệu mở bản đồ (openstreetmap.org), các chuẩn dữ liệu mở của tổ chức OGC (ogc.org).
- Đất đai: dữ liệu bản đồ đất đai (soilgrids.org), dữ liệu chuyển nhượng đất đai (openlandcontracts.org), dữ liệu mở đất đai của New Zealand (data.linz.govt.nz).
- Nông nghiệp: bộ từ vựng GACS Core (agrisemantics.org), cổng đăng ký từ điển dữ liệu nông nghiệp mở (agroportal.lirmm.fr).
- Giao thông vận tải: chuẩn dữ liệu quản lý

giao thông vận tải của châu Âu (transmodel-cen.eu), chuẩn dữ liệu biểu đồ vận tải công cộng của Google (gtfs.org).

Cần lưu ý rằng, khái niệm mở được dùng đối với từ điển dữ liệu có nội hàm ý nghĩa là chuẩn mở dùng để chia sẻ dữ liệu. Không phải tất cả dữ liệu được chia sẻ theo chuẩn mở có nghĩa mặc định sẽ là dữ liệu mở. Theo Luật Sở hữu trí tuệ, chỉ có thể có dữ liệu mở khi nó được phân phối cùng với một giấy cấp phép truy cập mở (ví dụ: Common Creative Attribution 4.0). Dữ liệu khi được chia sẻ đồng thời phải tuân thủ đầy đủ các quy định của Luật Công nghệ thông tin, Luật An toàn thông tin, trong đó có việc phải bảo vệ các dữ liệu cá nhân. Các thông tin liên quan đến danh tính và bí mật cá nhân của người khác sẽ chỉ được phép cung cấp cho bên thứ ba khi có quy định khác của pháp luật hoặc có sự đồng ý của người đó.

Việc triển khai xây dựng từ điển dữ liệu liên kết mở dùng cho phát triển chính phủ điện tử sẽ mang lại lợi ích, giá trị thực tiễn.

Thứ nhất, nó sẽ giúp hoàn thiện khung kiến trúc và kiến trúc chính phủ điện tử tại các bộ, ngành và địa phương. Khung kiến trúc chính phủ điện tử phiên bản 2.0 đã xây dựng một mô hình dữ liệu tham chiếu để áp dụng. Mô hình dữ liệu này hiện đang dừng ở đặc tả mức khái niệm. Từ điển dữ liệu liên kết mở sẽ thực hiện chi tiết hóa để hình thành mô hình dữ liệu tham chiếu ở mức logic. Qua đó tạo ra được một sự thống nhất về ngữ nghĩa dùng cho cơ sở hạ tầng dữ liệu của toàn bộ hệ thống chính phủ điện tử trong cả nước.

Thứ hai, từ điển dữ liệu liên kết mở có thể được dùng như là một cơ sở kỹ thuật để thực thi các chính sách về kết nối, chia sẻ dữ liệu của chính phủ. Nó bảo đảm tính phù hợp với các quy định của nhà nước, ví dụ như thông tư số 13/2017/TT-BTTTT quy định các yêu cầu kỹ thuật về kết nối các hệ thống thông tin, cơ sở dữ liệu với cơ sở dữ liệu quốc gia. Đồng thời, đáp ứng được tính hội nhập, bắt kịp với xu hướng phát triển của các công nghệ hiện đại trên thế giới.

Thứ ba, lược đồ tri thức có trong từ điển dữ

liệu giúp các nhà phát triển có định hướng đúng chuẩn công nghệ trong thiết kế xây dựng các hệ thống thông tin của chính phủ điện tử. Từ đó có thể rút ngắn thời gian triển khai, tiết kiệm được chi phí đầu tư không phải bổ sung nâng cấp hệ thống nhiều lần.

Thứ tư, cách tiếp cận xây dựng từ điển có tính tổng thể và hoàn toàn dựa trên nhu cầu bản địa. Qua đó phát huy được tất cả sức mạnh nội lực để phục vụ quá trình chuyển đổi số của chính phủ điện tử Việt Nam.

TÀI LIỆU THAM KHẢO

1. Davies, T., Walker, S., Rubinstein, M., & Perini, F. (Eds.) (2019). "The State of Open Data: Histories and Horizons", Cape Town and Ottawa: African Minds and International Development Research Centre, ISBN 978-1-928331-95-7.
2. Mai Anh. "Hệ sinh thái dữ liệu mở: Khởi đầu cho đô thị thông minh". Báo Sài Gòn Giải phóng Online, truy cập tại <https://www.sggp.org.vn/he-sinh-thai-du-lieu-mo-khoi-dau-cho-do-thi-thong-minh-442882.html>.
3. Tạ Tuấn Anh (2019). "Xây dựng hệ sinh thái dữ liệu mở cùng CMCN 4.0", Kỷ yếu Hội thảo khoa học Xây dựng và khai thác tài nguyên giáo dục mở, Hà Nội, tháng 10/2019.
4. Tạ Tuấn Anh (2019). "Chuyển đổi số ngành thông tin khoa học và công nghệ để hội nhập quốc tế", Tạp chí Thông tin và Tư liệu, Số 6/2019, ISSN 1859-2929.
5. "The OECD Open, Useful and Re-usable data (OURdata) Index: 2019". Truy cập tại <http://www.oecd.org/gov/digital-government/ourdata-index-policy-paper-2020.pdf>.
6. Wilkinson, M. D. et al. (2016). "The FAIR Guiding Principles for scientific data management and stewardship", Scientific Data. Vol 3.
7. <https://schema.org/docs/schemas.html>.
8. <https://www.popoloproject.com>,

(Ngày Tòa soạn nhận được bài: 10-4-2020; Ngày phản biện đánh giá: 14-9-2020; Ngày chấp nhận đăng: 15-11-2020).