

ĐỀ XUẤT VỀ TIÊU CHÍ DỮ LIỆU TRÊN HỆ TRI THỨC VIỆT SỐ HÓA

ThS Nguyễn Mạnh Duy
Bộ Khoa học và Công nghệ

Tóm tắt: Hệ tri thức Việt số hoá là một đề án của Chính phủ nhằm mục tiêu xây dựng nền tảng dữ liệu dùng chung trong các lĩnh vực khoa học, công nghệ và kinh tế - xã hội. Để lựa chọn được các dữ liệu đạt tiêu chuẩn để đưa lên Hệ tri thức Việt số hoá, cần xây dựng được bộ tiêu chí chuẩn cho các loại dữ liệu cụ thể nhằm tạo thuận lợi và thống nhất trong quá trình thu thập, xử lý, quản trị và chia sẻ dữ liệu. Bài báo này trao đổi và đề xuất các tiêu chí đối với một số loại dữ liệu đang được đưa lên chia sẻ trên Hệ tri thức Việt số hoá.

Từ khóa: Hệ tri thức Việt số hoá; dữ liệu mở; tiêu chí dữ liệu.

PROPOSAL ON STANDARD DATA CRITERIA OF VIETNAMESE DIGITAL KNOWLEDGE PLATFORM

Abstract: The Vietnamese Digital Knowledge Platform is a project of the Government aimed at building a shared data platform in the fields of science, technology and socio-economics. In order to select qualified data to upload to the digitized Vietnamese Digital Knowledge Platform, it is necessary to develop a set of standard criteria for specific types of data in order to facilitate and unify in the process of collection and processing, managing and data sharing. This article discusses and proposes criteria for some types of data being shared on the platform.

Keywords: Vietnamese Digital Knowledge Platform; open data; data criteria.

1. MỞ ĐẦU

Đề án “Phát triển Hệ tri thức Việt số hóa” được phê duyệt theo Quyết định số 677/QĐ-TTg ngày 18 tháng 5 năm 2017 của Thủ tướng Chính phủ. Đây là một trong các hoạt động cụ thể nhằm triển khai Chỉ thị 16/CT-TTg của Thủ tướng Chính phủ về tăng cường năng lực tiếp cận cuộc Cách mạng công nghiệp lần thứ tư. Mục tiêu của Đề án là xây dựng nền tảng dữ liệu và tri thức trong các lĩnh vực, nhất là những lĩnh vực liên quan trực tiếp đến đời sống của người dân như giáo dục, y tế, chăm sóc sức khỏe, nông nghiệp, văn hóa,... nhằm tạo điều kiện cho học tập, làm chủ tri thức, nghiên cứu và phát triển các ứng dụng công nghệ thông tin trên nền tảng AI và BigData phục vụ cho cộng đồng và xã hội.

Đề án đã thử nghiệm và triển khai thành công mô hình hợp tác giữa nhà nước - viện trường - doanh nghiệp - cộng đồng trong việc phát hiện và giải quyết những vấn đề cấp thiết của đất nước, thông qua việc ứng dụng công nghệ thông tin để tạo dựng các nền tảng dữ liệu dùng chung, huy động sự tham gia của cộng đồng đóng góp tạo lập, khai thác và sử dụng dữ liệu. Nhiều dự án đã được hình thành trong khuôn khổ Đề án, được triển khai thử nghiệm thành công và

hiện đã được chuyển giao cho các đơn vị có đủ năng lực để tiếp tục duy trì, phát triển.

Một số dự án về dữ liệu dùng chung đã được triển khai trên Hệ tri thức Việt số hoá gồm:

(1) *Dự án Dữ liệu khoa học dùng chung:* Đây là dự án với sự phối hợp của Viện nghiên cứu VinBigdata, Viện Công nghệ thông tin - Viện Hàn Lâm Khoa học và Công nghệ Việt Nam và một số đơn vị liên quan, nhằm xây dựng một nền tảng cho phép lưu trữ, quản lý và chia sẻ các dữ liệu khoa học mở (openscience) trên cơ sở ứng dụng công nghệ Cloud và Big Data. Các dữ liệu nghiên cứu khoa học (raw data) là nguồn tài nguyên khoa học quý giá nhưng lâu nay chưa được chia sẻ và sử dụng rộng rãi. Dự án sẽ giúp tạo ra môi trường để các tổ chức, cá nhân cùng nhau tham gia đóng góp, sử dụng những tài nguyên dữ liệu có giá trị để phục vụ cho hoạt động nghiên cứu khoa học, phát triển công nghệ và sản xuất, kinh doanh của mình.

(2) *Dự án Bản đồ số Việt Nam - Vmap:* Dự án nhằm tạo lập một nền tảng thu thập và chia sẻ dữ liệu bản đồ (tọa độ) của các địa danh, địa chỉ trên toàn quốc. Đây là cơ sở để các doanh nghiệp có thể sử dụng để tạo ra các ứng dụng trong nhiều lĩnh vực như: giáo dục, văn hóa, du lịch,... Thông

qua bản đồ số Việt Nam, người dùng cũng có thể dễ dàng tìm và được chỉ đường cụ thể đến tận lớp địa chỉ trong từng ngõ, hẻm hoặc các thôn, xã.

(3) *Dự án Hệ thống thông tin Nhân đạo số - iNhandao*: Dự án nhằm tạo lập hệ thống thu thập và cung cấp các địa chỉ nhân đạo trên toàn quốc, cho phép kết nối những người cần được trợ giúp với các nhà tài trợ và những người làm công tác thiện nguyện. Thông qua hệ thống sẽ giúp kết nối người cho, người nhận và những người làm từ thiện với nhau.

(4) *Dự án Giáo dục số - iGiaoduc*: Dự án với mục tiêu xây dựng nền tảng học liệu số (câu hỏi trắc nghiệm, bài giảng điện tử, học liệu STEM,...) cho các cấp học, góp phần thúc đẩy giáo dục trực tuyến và học tập suốt đời. Dự án đã thiết lập nền tảng Kho học liệu số với trên 7.000 bài giảng điện tử.

(5) *Dự án Bách khoa toàn thư số của Việt Nam*: Dự án với sự tham gia hợp tác giữa Đề án Hệ tri thức Việt số hóa và Đề án Biên soạn Bách khoa toàn thư Việt Nam, nhằm tạo ra nền tảng số nhằm kêu gọi cộng đồng cùng chung sức đóng góp các tài liệu và bài viết về tri thức nhân loại trong tất cả các lĩnh vực. Nền tảng hoạt động trên nguyên tắc mở, minh bạch và tôn vinh các cá nhân có đóng góp. Đến nay, dự án đã xây dựng xong nền tảng bách khoa toàn thư số và đưa hơn 60.000 mục từ vào hệ thống và đang kêu gọi cộng đồng tham gia đóng góp và chỉnh sửa bài viết.

2. TIÊU CHÍ DỮ LIỆU CỦA HỆ TRI THỨC VIỆT SỐ HÓA

Hệ tri thức Việt số hóa là một Đề án của Chính phủ với mục tiêu xây dựng một Hệ tri thức tổng hợp trong mọi lĩnh vực, góp phần thúc đẩy, tạo điều kiện để mọi người dân học tập suốt đời, làm chủ tri thức, tăng cường nghiên cứu sáng tạo, ứng dụng tiến bộ khoa học công nghệ, thúc đẩy phát triển đất nước.

Hệ tri thức Việt số hóa được xây dựng và cập nhật theo hình thức xã hội hóa, thu hút và khuyến khích mọi người dân và doanh nghiệp tham gia, với vai trò vừa khai thác vừa đóng góp để làm giàu các tài nguyên tri

thức số hóa của Việt Nam. Góp phần khơi dậy, lan tỏa niềm đam mê khoa học và công nghệ, khát vọng sáng tạo, cống hiến của mọi người trong việc tạo lập và phổ biến tri thức.

Mọi nội dung liên quan của Hệ tri thức Việt số hóa được lưu giữ và đăng tải trên website: itrithuc.vn

Chính sách đăng bài

Itrithuc.vn cho phép người dùng đăng bài viết và ý kiến của mình về các chủ đề tri thức. Mọi người dùng cần phải đăng ký và đăng nhập sử dụng tên thật của mình để có thể đăng bài hoặc thảo luận trên hệ thống. Bài viết của tác giả sẽ được hiển thị công khai. Do đó, website sẽ sử dụng chính sách kiểm duyệt để đảm bảo bài đăng phù hợp và không gây hại cho nhà nước, cộng đồng và những người khác.

Các quan điểm thể hiện trong các bài viết phản ánh những quan điểm của tác giả và không phản ánh quan điểm chính thức của Hệ tri thức Việt số hóa.

Về nguyên tắc, website sẽ đăng dữ liệu do người dùng cung cấp, nhưng các dữ liệu và thông tin đưa lên nói chung là lịch sử, khách quan, trung lập, có trích dẫn nguồn tài liệu tham khảo rõ ràng và tuân thủ các quy định về bản quyền sở hữu trí tuệ.

Website sẽ không hiển thị các bài đăng:

- chứa ngôn ngữ khiêu dâm, khiếm nhã, hoặc tục tĩu;
- chứa các mối đe dọa, hoặc tuyên bố phỉ báng;
- chứa các nội dung vu khống, bôi nhọ lãnh đạo Đảng, Nhà nước, xuyên tạc chủ trương, đường lối của Đảng, tuyên truyền chống Nhà nước Cộng hòa xã hội chủ nghĩa Việt Nam;
- chứa lời nói phân biệt về chủng tộc, màu da, giới tính, nguồn gốc quốc gia, dân tộc, tuổi tác, tôn giáo hoặc tàn tật;
- tiết lộ thông tin cá nhân/nhạy cảm của người khác mà không được phép (ví dụ: số điện thoại, CMT,...);
- chứa các thông tin được đăng vi phạm pháp luật, hoặc khuyến khích hoạt động bất

hợp pháp và tiết lộ thông tin mật của Nhà nước;

- quảng cáo các dịch vụ hoặc sản phẩm, bao gồm liên kết đến các trang web thương mại bên ngoài (lưu ý rằng các liên kết phi thương mại có liên quan đến chủ đề hoặc bài đăng là chấp nhận được);

- chứa nội dung nằm ngoài chủ đề, spam hoặc chủ yếu nhằm quảng bá liên kết.

Người dùng không được đăng các tài liệu có bản quyền hay thương hiệu lên website trừ trường hợp chúng minh được rằng mình được phép làm như vậy. Bằng cách gửi bài đăng hoặc tác phẩm của mình, tác giả cho phép bất kỳ ai xem website Hệ tri thức quyền sao chép, phân phối, tạo ra các dẫn xuất, hiển thị hoặc trình diễn tác phẩm của người đăng một cách công khai và miễn phí.

Đối với các bài đăng vi phạm một trong các quy định nêu trên, website sẽ có quyền yêu cầu tác giả chỉnh sửa nội dung cho phù hợp hoặc xóa bỏ bài đăng mà không cần phải thông báo trước cho tác giả.

Liên kết tới Hệ tri thức

Itrithuc.vn là trang web chính thức của Hệ tri thức Việt số hóa. Bạn có thể liên kết tới itrithuc.vn miễn phí. Khi bạn liên kết tới itrithuc.vn, vui lòng giới thiệu website như là một cổng tri thức miễn phí cho mọi người khi họ cần tìm dữ liệu, thông tin, kiến thức thuộc mọi lĩnh vực.

Dữ liệu mở

Kho dữ liệu mở thuộc đề án Hệ tri thức Việt số hóa là nguồn lực chiến lược quốc gia, với sự tham gia đóng góp bởi các cơ quan Nhà nước, tổ chức, doanh nghiệp và mỗi người dân. Dữ liệu mở là dữ liệu có thể được bất kỳ ai tự do sử dụng, sử dụng lại và phân phối lại, chỉ yêu cầu ghi nhận nguồn và chia sẻ tương tự.

Đối với các cơ quan nhà nước, dữ liệu mở giúp Chính phủ cải thiện được tính minh bạch và công khai, giúp cải tiến quy trình quản lý, tổ chức dữ liệu, tăng hiệu quả hoạt động trong các cơ quan. Các doanh nghiệp dựa vào nguồn dữ liệu mở để tối ưu hoá hoạt động kinh doanh, cải tiến tiếp thị và phát triển sản phẩm. Dữ liệu mở cũng mở

cũng giúp định hướng đầu tư kinh doanh, cải thiện cơ hội việc làm và thúc đẩy tăng trưởng kinh tế. Đối với người dân, dữ liệu mở giúp người dân có được các thông tin cần thiết để có thể so sánh và đối chiếu, tạo điều kiện thuận lợi cho thực hiện quyền dân chủ và tham gia quản lý.

Dữ liệu mở của các cơ quan nhà nước bao gồm các dữ liệu và thông tin được tạo ra, thu thập, xử lý, phổ biến bởi cơ quan nhà nước. Các dữ liệu này cần đảm bảo các nguyên tắc sau đây: Sẵn có, Công khai, Có cấu trúc, Cho phép người dùng tái sử dụng không hạn chế.

Mỗi cơ quan, tổ chức cử 01 đầu mối để cung cấp (upload) dữ liệu của đơn vị mình lên Hệ tri thức. Đồng thời hỗ trợ và hướng dẫn người dùng sử dụng dữ liệu đúng cách. Bản quyền các dữ liệu mở thuộc về tổ chức và cá nhân phát hành ra chúng, do đó, khi sử dụng dữ liệu người dùng cần trích dẫn nguồn phù hợp và tôn trọng quyền sở hữu trí tuệ.

Đảm bảo chất lượng dữ liệu

Hệ tri thức khuyến khích người dùng đăng các bài viết hoặc tác phẩm có giá trị và chất lượng cao, cung cấp các tri thức khoa học hữu dụng, văn phong mạch lạc sử dụng từ ngữ dễ hiểu, khách quan, trích dẫn các nguồn tư liệu rõ ràng.

Bên cạnh đó, website sử dụng các công cụ tự động và đội ngũ chuyên gia để đánh giá chất lượng của bài đăng. Các bài đăng đã được kiểm định chất lượng sẽ được gắn tem đánh dấu để người dùng nhận biết.

Yêu cầu đối với dữ liệu dùng chung

Dữ liệu dùng chung là các loại dữ liệu được bộ, ngành, địa phương thu thập và tạo ra trong quá trình hoạt động của mình và có thể được công khai cho công chúng truy cập và khai thác mà không có bất kỳ sự hạn chế nào.

Dữ liệu dùng chung bao gồm dữ liệu của các Bộ, ngành, địa phương và của các cơ quan, tổ chức trực thuộc.

Dữ liệu dùng chung phải là những dữ liệu có thể được phân tích, xử lý, khai thác để tạo ra các giá trị gia tăng. Tức là, không

bao gồm các loại thông tin, dữ liệu mang tính chất tin tức, thông báo, báo cáo,...

Dữ liệu dùng chung đảm bảo an toàn, an ninh thông tin, không tiết lộ bí mật nhà nước và thông tin cá nhân.

Dữ liệu dùng chung được cung cấp dưới dạng máy đọc được, tức là có gắn kèm các siêu dữ liệu mô tả (metadata) giúp người dùng có thể sử dụng các công cụ để đọc và khai phá dữ liệu.

Trách nhiệm của các Bộ, ngành, địa phương

Lập danh sách và mô tả tất cả các loại dữ liệu mà Bộ, ngành, địa phương có thể công khai chia sẻ không hạn chế truy cập. Danh sách này phải được đăng công khai lên trang dành riêng cho bộ, ngành, địa phương trên website Hệ tri thức Việt số hóa và bao gồm ít nhất các thông tin sau đây:

- Tên bộ dữ liệu
- Mô tả nội dung bộ dữ liệu
- Tần suất cập nhật
- Địa chỉ URL của bộ dữ liệu (link tới website chứa bộ dữ liệu hoặc webservice)
- Các điều khoản và điều kiện sử dụng dữ liệu
- Thông tin về đầu mối liên hệ trong trường hợp người dùng cần thêm thông tin hoặc chỉ dẫn về dữ liệu

Hằng năm, Bộ, ngành, địa phương có trách nhiệm cập nhật danh sách này.

Ngoài ra, các Bộ, ngành đăng tải (upload) các bộ dữ liệu dùng chung của bộ, ngành, địa phương mình lên Hệ tri thức Việt số hóa theo danh sách dữ liệu và tần suất cập nhật đã công bố công khai. Trong trường hợp dữ liệu đã có sẵn dưới dạng cơ sở dữ liệu đặt tại cơ quan chủ quản thì cung cấp địa chỉ URL liên kết tới nguồn dữ liệu hoặc webservice để khai thác dữ liệu.

3. KẾT LUẬN

Hệ tri thức Việt số hóa là một Đề án của Chính phủ với mục tiêu xây dựng một Hệ tri thức tổng hợp trong mọi lĩnh vực, góp phần thúc đẩy, tạo điều kiện để mọi người dân học tập suốt đời, làm chủ tri thức, tăng cường nghiên cứu sáng tạo, ứng dụng tiến

bộ khoa học công nghệ, thúc đẩy phát triển đất nước.

Hệ tri thức Việt số hóa được xây dựng và cập nhật theo hình thức xã hội hóa, thu hút và khuyến khích mọi người dân và doanh nghiệp tham gia, với vai trò vừa khai thác vừa đóng góp để làm giàu các tài nguyên tri thức số hóa của Việt Nam. Góp phần khơi dậy, lan tỏa niềm đam mê khoa học và công nghệ, khát vọng sáng tạo, cống hiến của mọi người trong việc tạo lập và phổ biến tri thức.

Hệ tri thức Việt số hóa cho phép người dùng đăng bài viết và ý kiến của mình về các chủ đề tri thức. Mọi người dùng cần phải đăng ký và đăng nhập sử dụng tên thật của mình để có thể đăng bài hoặc thảo luận trên hệ thống. Đề án khuyến khích người dùng chia sẻ kiến thức và suy nghĩ về các chủ đề được đưa ra thảo luận hoặc các câu hỏi. Bài viết của bạn sẽ được hiển thị công khai. Do đó, website sẽ sử dụng chính sách kiểm duyệt để đảm bảo bài đăng của bạn phù hợp và không gây hại cho nhà nước, cộng đồng và những người khác.

Về nguyên tắc chung, Đề án mong rằng dữ liệu được đưa lên từ phía người dùng phải đảm bảo yếu tố khách quan, trung lập, có trích dẫn nguồn tài liệu tham khảo rõ ràng và tuân thủ các quy định về bản quyền và sở hữu trí tuệ.

TÀI LIỆU THAM KHẢO

1. Quyết định số 677/QĐ-TTg ngày 18/5/2017 của Thủ tướng Chính phủ phê duyệt Đề án “Phát triển Hệ tri thức Việt số hoá”
2. <http://opendatahandbook.org/guide/en/how-to-open-up-data/>
3. <https://yourdatastories.eu/the-eu-open-data-handbook/>
4. <https://opendatacharter.net/principles/>
5. <https://project-open-data.cio.gov/principles/>
6. EU open dataThe basics for EU data providers (2016 EDITION)

(Ngày Tòa soạn nhận được bài: 15-12-2022; Ngày phản biện đánh giá: 6-01-2023; Ngày chấp nhận đăng: 15-01-2023).