**BỘ THÔNG TIN VÀ TRUYỀN THÔNG**

**CỤC CHUYỂN ĐỔI SỐ QUỐC GIA**

---------------------------

**THUYẾT MINH**

**DỰ THẢO TIÊU CHUẨN QUỐC GIA**

**ĐẶC TẢ TẬP DỮ LIỆU MỞ CHO CỔNG DỮ LIỆU**

**DCAT-VN**

**Hà Nội, 2023**

**MỤC LỤC**

[1. Tên gọi và ký hiệu tiêu chuẩn 3](#_Toc126771490)

[2. Phạm vi và đối tượng áp dụng 3](#_Toc126771491)

[3. Đặt vấn đề 3](#_Toc126771492)

[3.1. Tình hình phát triển cổng dữ liệu cung cấp dữ liệu mở trên thế giới 3](#_Toc126771493)

[3.2. Tình hình phát triển cổng dữ liệu cung cấp dữ liệu mở tại Việt Nam 6](#_Toc126771494)

[3.3. Các quy định về phát triển dữ liệu mở 9](#_Toc126771495)

[3.4. Sự cần thiết và vai trò của tiêu chuẩn về mô tả tập dữ liệu mở 13](#_Toc126771496)

[4. Sở cứ xây dựng tiêu chuẩn 14](#_Toc126771497)

[4.1. Một số tiêu chuẩn, quy định trên thế giới 14](#_Toc126771498)

[4.1.1. Dublin Core 14](#_Toc126771499)

[4.1.2. DCAT 16](#_Toc126771500)

[4.1.3. DCAT-AP 18](#_Toc126771501)

[4.1.4. DCAT-US 19](#_Toc126771502)

[4.1.5. ISO 19115 21](#_Toc126771503)

[4.1.6. CKAN-API 22](#_Toc126771504)

[4.1.7. Chuẩn dữ liệu của cổng dữ liệu Singapore 24](#_Toc126771505)

[4.2. Đánh giá các tiêu chuẩn 24](#_Toc126771506)

[4.3. Thông tin tham chiếu của Asean và đánh giá chuẩn mô tả dữ liệu mở tại và Việt Nam 25](#_Toc126771507)

[5. Phương án xây dựng và giải thích nội dung tiêu chuẩn 26](#_Toc126771508)

[5.1. Đánh giá về việc khả năng chấp thuận các tiêu chuẩn thế giới 27](#_Toc126771509)

[5.2. Yêu cầu cho phương án đề xuất 28](#_Toc126771510)

[5.3. Phương án xây dựng chuẩn về mô tả tập dữ liệu mở 28](#_Toc126771511)

[5.4. Nội dung chuẩn mô tả tập dữ liệu mở DCAT-VN 29](#_Toc126771512)

[6. Bảng đối thiếu nội dung tiêu chuẩn với tài liệu tham khảo 31](#_Toc126771513)

[7. Khuyến nghị áp dụng tiêu chuẩn 32](#_Toc126771514)

[8. Kết luận 33](#_Toc126771515)

# Tên gọi và ký hiệu tiêu chuẩn

Tên tiêu chuẩn quốc gia**: Đặc tả tập dữ liệu mở cho cổng dữ liệu.**

Ký hiệu tiêu chuẩn quốc gia**: TCVN xxxx:2023.**

# Phạm vi và đối tượng áp dụng

Tiêu chuẩn này quy định cấu trúc dữ liệu đặc tả tập dữ liệu mở được công bố trên cổng dữ liệu của cơ quan nhà nước, tổ chức, cá nhân phục vụ mục đích công bố tập dữ liệu mở, kết nối các cổng dữ liệu với nhau để trao đổi thông tin về tập dữ liệu mở.

Tiêu chuẩn này áp dụng trong các hoạt động sau:

- Xây dựng cổng dữ liệu cung cấp dữ liệu mở: nội dung tiêu chuẩn này mô tả cấu trúc dữ liệu của tập dữ liệu mở phục vụ thiết kế cơ sở dữ liệu chứa dữ liệu đặc tả về tập dữ liệu mở. Thiết kế cơ sở dữ liệu của cổng dữ liệu được mở rộng từ mô hình cấu trúc dữ liệu được quy định trong tiêu chuẩn này.

- Xây dựng và công bố tập dữ liệu mở: dữ liệu mở được công bố phải kèm theo dữ liệu đặc tả. Tiêu chuẩn này có hướng dẫn để ghi các mô tả về dữ liệu mở khi được công bố.

- Kết nối, chia sẻ thông tin về dữ liệu mở giữa các cổng dữ liệu. Tiêu chuẩn này mô tả cấu trúc dữ liệu đặc tả tập dữ liệu mở khi được trao đổi giữa các cổng dữ liệu trong trao đổi thông tin về tập dữ liệu.

- Tra cứu, khai thác các tập dữ liệu mở được công bố. Dữ liệu mở khi được công bố phải kèm theo các thông tin đặc tả được thể hiện trong tiêu chuẩn này. Thông tin đặc tả cung cấp thông tin sơ bộ về dữ liệu mở trước khi người dùng khai thác dữ liệu chính thức.

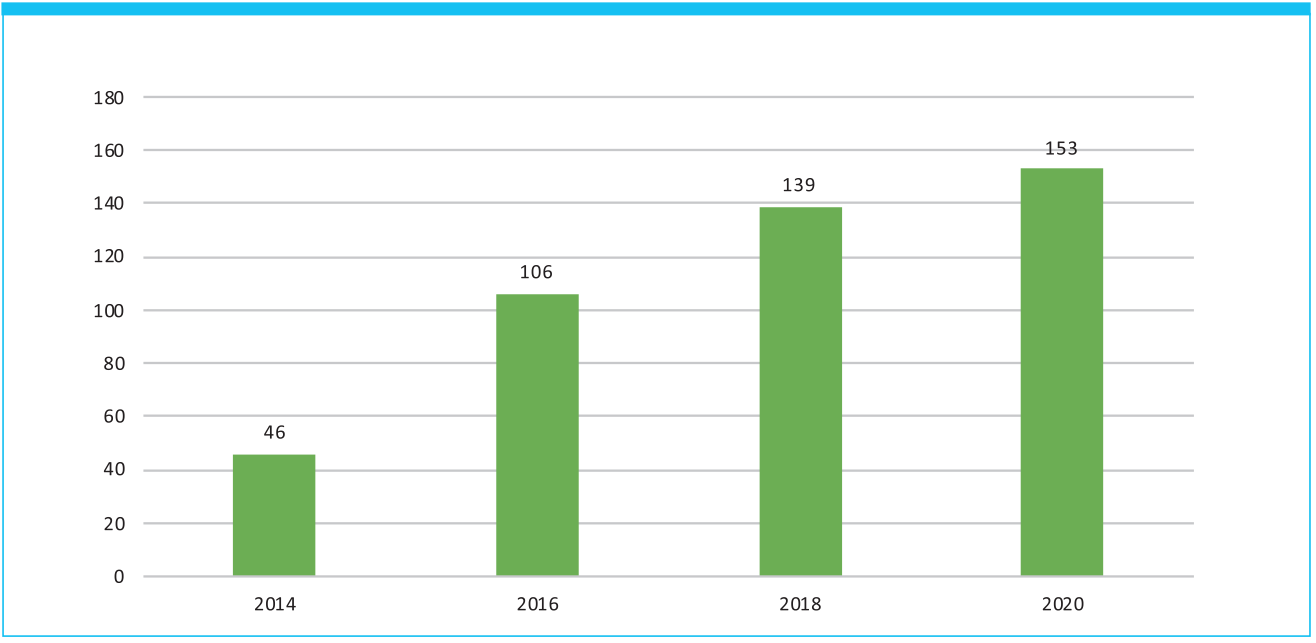
Các trường tin trong tiêu chuẩn này là tối thiểu cần đáp ứng.

# Đặt vấn đề

## Tình hình phát triển cổng dữ liệu cung cấp dữ liệu mở trên thế giới

Trong các năm gần đây, các quốc gia trên thế giới đã nhận thức được tầm quan trọng của dữ liệu mở và tích cực cung cấp dữ liệu mở cho cộng đồng. Theo thống kê báo cáo của Liên hợp quốc, từ năm 2014 số lượng quốc gia có cổng dữ liệu cung cấp dữ liệu mở đã tăng từ 46 quốc gia năm 2014 (khoảng 24%) lên 153 quốc gia vào năm 2020 đạt được khoảng 80%.

Trong số các quốc gia có cổng dữ liệu có 59% quốc gia có chính sách về dữ liệu mở, 62% quốc gia có đặc tả dữ liệu hoặc từ điển dữ liệu., 57% chấp nhận yêu cầu công khai dữ liệu đối với các tập dữ liệu mới, 52% có hướng dẫn sử dụng dữ liệu mở và 49% tham gia các nỗ lực quảng bá cho người dân, doanh nghiệp sử dụng dữ liệu mở. Ví dụ cho sự quảng bá này là tổ chức các cuộc thi hackathon về dữ liệu. Trong số 153 quốc gia có cổng dữ liệu thì có 114 quốc gia có chính sách về dữ liệu một cách rõ ràng.



**Hình 1 - Thống kê số nước có cổng dữ liệu cung cấp dữ liệu mở**

Về tổng hợp các cổng dữ liệu nói chung bao gồm cả các cổng dữ liệu của chính quyền địa phương, Tổ chức Open Knowlegde Foudation có sáng kiến thống kê tại dataportals.org toàn thế giới có hơn 597 cổng dữ liệu cung cấp dữ liệu mở đã được triển khai, đưa vào vận hành. Trong đó 197 cổng dữ liệu tại châu Âu, hơn 100 cổng dữ liệu ở Mỹ và chỉ 33 cổng dữ liệu ở châu Phi.

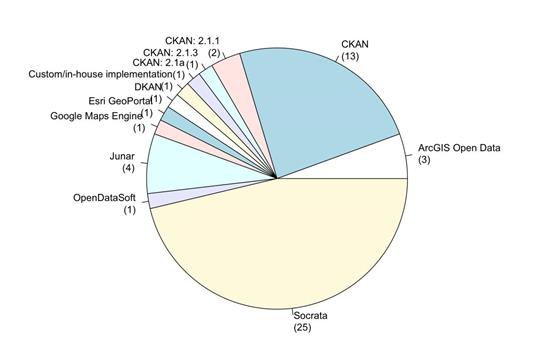
Một thông tin khác là có khoảng 118 cổng dữ liệu được phân loại thuộc nhóm Chính phủ, 12 cổng dữ liệu là cộng đồng, 5 cổng dữ liệu là tổ chức và 6 cổng dữ liệu là tổ chức nghiên cứu. Trong tổng số 141 cổng dữ liệu cung cấp dữ liệu mở được chỉ định một nhà xuất bản, 380 cổng dữ liệu cung cấp dữ liệu mở còn lại không được chỉ định trên cơ sở phân loại nhà xuất bản.

Một số cổng dữ liệu cung cấp dữ liệu mở cũng nhận thấy là không hoạt động.

Tuy nhiên không phải các cổng đều công bố công nghệ mà mình sử dụng. Điều đó khó xác định được thị phần các nền tảng xây dựng cổng.

Theo tổng hợp của Beata Lisowska trong dự án xây dựng tiêu chuẩn kết nối cổng dữ liệu cung cấp dữ liệu mở được thực hiện năm 2016 được đánh giá trên 55 cổng dữ liệu thì các nền tảng CLAN, Socrata, Junar là các nền tảng thông dụng và sử dụng phổ biến nhất thế giới.

Cụ thể thị phần được thể hiện theo hình 2 dưới đây.



**Hình 2 - Tỉ lệ sử dụng nền tảng theo khảo sát của dataportals.org**

Trong khuôn khổ nghiên cứu, đã thực hiện đánh giá chi tiết hơn về công nghệ của các cổng dữ liệu quốc gia trên thế giới cung cấp dữ liệu mở để xác định công nghệ sử dụng. Cụ thể thu được kết quả sau:

|  |  |  |
| --- | --- | --- |
| **STT** | **Cổng dữ liệu của nước** | **Công nghệ sử dụng** |
|  | Australia | CKAN |
|  | Canada | CKAN |
|  | Mỹ | CKAN |
|  | Singapore | CKAN |
|  | Châu Phi | CKAN |
|  | Brasil | CKAN |
|  | Liên minh châu Âu | CKAN |
|  | Anh | CKAN |
|  | Hà Lan | CKAN |
|  | Agentina | CKAN |
|  | Phần lan | CKAN |
|  | Hy lạp | CKAN |
|  | Estonia | JKAN |
|  | Ai len | CKAN |
|  | New Zealand | CKAN[[1]](#footnote-1) |
|  | Hàn Quốc | Tự xây dựng |
|  | Hà Lan | CKAN/DCAT |
|  | Pháp | CKAN[[2]](#footnote-2) |
|  | Indonesia | CKAN |

Như vậy, đa phần các cổng dữ liệu quốc gia cung cấp dữ liệu mở sử dụng công nghệ CKAN để làm nền tảng xây dựng cổng dữ liệu của mình. Điều này chứng tỏ rằng CKAN là nền tảng mạnh mẽ nhất và phù hợp nhất để sử dụng. Nguyên nhân có thể bao gồm:

* CKAN là mã nguồn mở, người sử dụng có thể sử dụng miễn phí và tự do phát triển bổ sung thêm chức năng
* CKAN được tổ chức Open Knowledge Foundation duy trì. Đây là tổ chức mạnh và nổi tiếng về kiến thức mở
* CKAN có một kiến trúc mở trong đó cho phép người sử dụng tự tạo các plugin để bổ sung tính năng tuỳ biến theo nhu cầu
* CKAN có cộng đồng phát triển các bộ thư viện plugin phong phú tuỳ chọn cho ngươi sử dụng.
* CKAN có đủ các tính năng cơ bản nhất và đặc biệt là khi nhiều người sử dụng thì khả năng kết nối, tích hợp rất dễ dàng.

Theo tổng hợp thì có hơn 50 quốc gia đang sử dụng CKAN làm nền tảng xây dựng dữ liệu mở.

Một số kinh nghiệm rút ra về kinh nghiệm thế giới

- Việc xây dựng cổng dữ liệu cung cấp dữ liệu mở trên thế giới diễn ra rất nhanh trong thời gian qua. Điều này thể hiện sự ưu tiên của các quốc gia trong việc mở dữ liệu của cơ quan nhà nước, tổ chức giúp thúc đẩy chuyển đổi số trong kinh tế và xã hội, xây dựng các cộng đồng chung về chia sẻ tri thức.

- Hầu như các nước không tự xây dựng các nền tảng riêng mà phát triển cổng dữ liệu cung cấp dữ liệu mở trên các nền tảng công nghệ chung.

- Nền tảng sử dụng chủ yếu là CKAN. Đây là một nền tảng mở, sử dụng miễn phí và hỗ trợ sự tuỳ biến đa dạng.

- Một số các cổng tích hợp thêm các công nghệ phụ trợ hoặc tích hợp nhiều công nghệ, nền tảng vào một cổng để tận dụng thế mạnh của mỗi nền tảng khi phát triển các cấu thành có chức năng khác nhau.

## Tình hình phát triển cổng dữ liệu cung cấp dữ liệu mở tại Việt Nam

*a) Cổng dữ liệu quốc gia*

Cổng dữ liệu quốc gia được xác định vị trí pháp lý tại Nghị định số 47/2020/NĐ-CP ngày 09/4/2022 của Chính phủ về quản lý, kết nối và chia sẻ dữ liệu số. Cổng dữ liệu quốc gia là điểm đầu mối truy cập thông tin, dữ liệu có địa chỉ data.gov.vn trên mạng Internet phục vụ việc công bố dữ liệu mở, cung cấp thông tin về chia sẻ dữ liệu của cơ quan nhà nước; cung cấp các tài liệu, dịch vụ, công cụ, ứng dụng xử lý, khai thác dữ liệu được công bố bởi các cơ quan nhà nước.

Chức năng của Cổng dữ liệu quốc gia:

* Là đầu mối cung cấp dữ liệu mở của các cơ quan nhà nước trên môi trường mạng nhằm tăng cường tính minh bạch trong hoạt động của Chính phủ và thúc đẩy sáng tạo, phát triển kinh tế, xã hội.
* Là đầu mối cung cấp hỗ trợ kỹ thuật, giám sát tình hình kết nối, chia sẻ dữ liệu giữa các hệ thống thông tin trong cơ quan nhà nước.
* Cơ quan nhà nước có trách nhiệm theo dõi, đăng ký và cung cấp thông tin về dữ liệu, dữ liệu mở; quản lý dữ liệu mở; khả năng chia sẻ dữ liệu và nhu cầu khai thác dữ liệu trên Cổng dữ liệu quốc gia.
* Ngoài cung cấp dữ liệu mở, Cổng dữ liệu quốc gia bao gồm cả Hệ thống quản lý dịch vụ chia sẻ dữ liệu để quản lý các dịch vụ chia sẻ dữ liệu, giao dịch dữ liệu trong cơ quan nhà nước.
* Đáp ứng các yêu cầu khác liên quan đến quản trị dữ liệu của cơ quan nhà nước.

Hiện nay, Cổng dữ liệu quốc gia mới được thiết lập ban đầu tại <http://data.gov.vn>; thành phần cổng dữ liệu cung cấp dữ liệu mở tại <http://open.data.gov.vn>.

Công nghệ sử dụng cho cổng dữ liệu quốc gia tại thời điểm thử nghiệm là CKAN; đã kết nối liên thông với Cổng dữ liệu iTrithuc để trao đổi dữ liệu đặc tả.

Cổng dữ liệu đã có hơn 10.600 tập dữ liệu mở đã được cung cấp.

*b) Cổng dữ liệu Thành phố Hồ Chí Minh*

Cổng dữ liệu của Thành phố Hồ Chí Minh được công bố trên mạng Internet tại địa chỉ <https://opendata.hochiminhcity.gov.vn/>. Cổng dữ liệu cung cấp dữ liệu mở cũng là một thành phần của kho dữ liệu dùng chung của Thành phố Hồ Chí Minh.

Đây là là nền tảng quan trọng trong việc triển khai xây dựng đô thị thông minh; phục vụ triển khai các hệ thống thông tin, ứng dụng phục vụ người dân, doanh nghiệp và công tác quản lý, chỉ đạo điều hành của thành phố, đồng thời đây là nguồn cung cấp dữ liệu làm cơ sở phát triển Hệ sinh thái dữ liệu mở cho Thành phố Hồ Chí Minh.

Cổng dữ liệu cung cấp dữ liệu mở của Thành phố Hồ Chí Minh được phát triển để tạo ra phương thức chia sẻ tài nguyên dữ liệu từ Kho dữ liệu dùng chung với người dân, doanh nghiệp và tổ chức.

Thông qua Cổng dữ liệu cung cấp dữ liệu mở, Thành phố thúc đẩy người dân, doanh nghiệp và tổ chức:

* Tham gia giám sát, quản lý các mặt hoạt động của chính quyền, xã hội;
* Sử dụng dữ liệu mở để tạo ra sản phẩm giá trị mới đóng góp cho hệ sinh thái ứng dụng, tiện ích của Thành phố để phục vụ người dân tốt hơn;
* Góp phần thúc đẩy phát triển hệ sinh thái khởi nghiệp sáng tạo, phát triển nền kinh tế số của Thành phố;
* Đóng góp dữ liệu được hình thành trong quá trình khai thác để phát triển Hệ sinh thái dữ liệu mở của Thành phố.

Cổng dữ liệu cung cấp dữ liệu mở Thành phố Hồ Chí Minh được phân thành 6 chủ đề bao gồm: Đầu tư, Y tế; Thị trường, Môi trường; Khí hậu, Giáo dục.

Hiện tại, Cổng dữ liệu cung cấp dữ liệu mở Thành phố Hồ Chí Minh cung cấp dữ liệu còn khiêm tốn, mới chỉ có 8 tập dữ liệu được công bố.

Khảo sát sơ bộ cho thấy, Cổng dữ liệu cung cấp dữ liệu mở Thành phố Hồ Chí Minh sử dụng nền tảng CKAN để phát triển. Do đó, đặc tả dữ liệu mở được sử dụng theo sự hỗ trợ của CKAN.

*c) Cổng dữ liệu Đà Nẵng*

Cổng dịch vụ dữ liệu Thành phố Đà Nẵng bao gồm chức năng của Cổng dữ liệu cung cấp dữ liệu mở là một dự án Công nghệ thông tin của Sở Thông tin và Truyền thông thành phố Đà Nẵng. Mục tiêu xây dựng Cổng dịch vụ dữ liệu là xây dựng một cơ sở dữ liệu phục vụ cho tra cứu của người dân, làm cơ sở cho việc khai thác và phát triển các ứng dụng khác dựa trên cơ sở dữ liệu nền tảng.

Tháng 10/2019 Cổng dịch vụ dữ liệu được chính chức đi vào hoạt động với kỳ vọng sẽ góp phần cùng với chính quyền thành phố phục vụ người dân và doanh nghiệp được tốt hơn, thúc đẩy phát triển, ứng dụng công nghệ thông tin, xây dựng thành phố thông minh, hiện đại.

Cổng dịch vụ dữ liệu cung cấp 3 dịch vụ sau: Dịch vụ tra cứu dữ liệu trên website, Dịch vụ tra cứu dữ liệu qua tin nhắn di động SMS và Dịch vụ tra cứu dữ liệu qua ứng dụng di động Zalo.

*\*Đối với người trực tiếp khai thác dữ liệu:*

- Trực tiếp tra cứu dữ liệu trên Cổng, hoặc qua các hình thức SMS, Zalo.

- Tải dữ liệu về để sử dụng.

- Sử dụng API được cung cấp để phát triển ứng dụng.

*\*Đối với người sở hữu dữ liệu:*

- Có thể tạo các dịch vụ tra cứu dữ liệu qua ba hình thức như trên.

- Kiếm lợi nhuận từ việc tra cứu dữ liệu.

- Tạo ra API tra cứu để sử dụng cho bên thứ 3 phát triển ứng dụng.

Cổng dữ liệu Đà Nẵng được tự xây dựng không sử dụng nền tảng CKAN.

*d) Cổng dữ liệu Itrithuc*

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 nhằm xây dựng nền tảng hạ tầng dữ liệu và tri thức trong các lĩnh vực, đặc biệt 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.

Các dữ liệu và tri thức được kết nối và chia sẻ với cấp số nhân, được phổ biến nhanh chóng và rộng rãi sẽ xoá bỏ khoảng cách số, tạo cơ hội để mọi người tiếp cận sử dụng, tạo ra giá trị gia tăng cho cả cộng đồng. Đây là cốt lõi để triển khai phát triển hệ thống đổi mới sáng tạo quốc gia, thúc đẩy khởi nghiệp sáng tạo trong mọi ngành, lĩnh vực, khơi dậy niềm đam mê và khát vọng sáng tạo, nhất là thế hệ trẻ.

Cổng dữ liệu cung cấp dữ liệu mở Itrithuc với hơn 10.000 tập dữ liệu mở đa dạng về chủng loại các lĩnh vực, được công bố tại <https://dulieu.itrithuc.vn/>.

Công nghệ của cổng dữ liệu cung cấp dữ liệu mở là sử dụng CKAN hỗ trợ đồng thời tiêu chuẩn mô tả tập dữ liệu mở của CKAN và DCAT.

b) Cổng dữ liệu của các bộ, ngành, địa phương khác

Tới cuối năm 2022, toàn quốc đã có 15 tỉnh, thành phố xây dựng Cổng dữ liệu cung cấp dữ liệu mở với các mức độ khác nhau. Nhiều cổng đang trong giai đoạn thử nghiệm.

|  |  |
| --- | --- |
| Bắc Giang | [http://data.bacgiang.gov.vn](http://data.bacgiang.gov.vn/) |
| Lạng Sơn | <http://data.langson.gov.vn/> |
| Long An | [http://data.longan.gov.vn](http://data.longan.gov.vn/) |
| Đà Nẵng | <https://congdulieu.vn/> |
| Cần Thơ | [https://data.cantho.gov.vn](https://data.cantho.gov.vn/) |
| Hậu Giang | [https://data.haugiang.gov.vn](https://data.haugiang.gov.vn/) |
| Bộ Kế hoạch và Đầu tư | <https://data.mpi.gov.vn/> |
| Ninh Bình | [https://data.ninhbinh.gov.vn](https://data.ninhbinh.gov.vn/) |
| Thái Bình | [https://data.thaibinh.gov.vn](https://data.thaibinh.gov.vn/) |
| Thừa Thiên Huế | <https://data.thuathienhue.gov.vn/> |
| Bình Định | <https://opendata.binhdinh.gov.vn/> |
| TP Hồ Chí Minh | <https://opendata.hochiminhcity.gov.vn/> |
| Bộ Tài nguyên và Môi trường | <https://opendata.monre.gov.vn/> |
| Quảng Ngãi | <https://opendata.quangngai.gov.vn/> |
| Trà Vinh | [https://opendata.travinh.gov.vn](https://opendata.travinh.gov.vn/) |

Qua khảo sát ban đầu bên ngoài, các công nghệ sử dụng rất đa dạng và chưa có sự thống nhất về các tập dữ liệu, đặc tả tập dữ liệu. Một số cổng dữ liệu cung cấp dữ liệu mở chưa có dữ liệu hoặc mới có giao diện cơ bản.

Nhiều cổng dữ liệu cung cấp dữ liệu mở chưa sử dụng bất kỳ tiêu chuẩn về tập dữ liệu, đặc tả tập dữ liệu nào mà chỉ thiết kế để cung cấp các tệp dữ liệu hoặc bảng dữ liệu cố định.

## Các quy định về phát triển dữ liệu mở

*a) Quy định pháp lý về dữ liệu mở*

Dữ liệu mở tại Việt nam là phạm trù mới, mới chỉ được quy định tại Nghị định số 47/2020/NĐ-CP.

Chính phủ mở là một nấc phát triển của Chính phủ điện tử khi Chính phủ cung cấp dữ liệu cho cộng đồng để thực hiện chủ trương “Nhà nước kiến tạo phát triển”. Lần đầu tiên, một văn bản pháp lý đưa nội dung “Dữ liệu mở của cơ quan nhà nước” đánh dấu một mốc quan trọng để thực thi chủ trương này đồng thời cũng thể hiện sự tích cực của Việt Nam khi sẵn sàng cung cấp dữ liệu mở cho cộng đồng, người dân, doanh nghiệp. Quy định pháp lý về dữ liệu mở cũng là một nội dung khá mới không chỉ đối với Việt Nam mà còn đối với nhiều nước trên thế giới khi công bố dữ liệu mở của các nước chủ yếu được triển khai dưới dạng sáng kiến.

Nội dung quy định về dữ liệu mở được xây dựng trên cơ sở tương thích với các quy định thông dụng phổ biến trên thế giới như: dữ liệu mở phải toàn vẹn, phản ánh đầy đủ thông tin cần cung cấp, cập nhật, máy có thể đọc được, ở định dạng mở, miễn phí, tự do sử dụng...

Nghị định số 47/2020/NĐ-CP cũng quy định các cơ quan nhà nước phải xây dựng một kế hoạch và triển khai cung cấp dữ liệu mở theo kế hoạch đã xây dựng. Kế hoạch phải đảm bảo có yêu cầu tối thiểu và phù hợp với nhu cầu của người dân, doanh nghiệp. Cơ chế triển khai dữ liệu mở cũng tạo cơ hội cho người dân, doanh nghiệp trong xã hội, cộng đồng tham gia ý kiến phản hồi, đóng góp mở rộng dữ liệu mở.

Dữ liệu mở nói chung chưa được quy định dưới dạng pháp lý mà chỉ có quy định về dữ liệu mở của cơ quan nhà nước. Trong đó, Dữ liệu mở của cơ quan nhà nước: là dữ liệu được cơ quan nhà nước có thẩm quyền công bố rộng rãi cho cơ quan, tổ chức, cá nhân tự do sử dụng, tái sử dụng, chia sẻ. Trong Nghị định số 47/2020/NĐ-CP, dữ liệu mở được hiểu là dữ liệu mở của cơ quan nhà nước.

Theo Nghị định số 47/2020/NĐ-CP, hoạt động cung cấp dữ liệu mở của cơ quan nhà nước cho tổ chức, cá nhân thực hiện theo các nguyên tắc sau:

- Dữ liệu mở phải toàn vẹn và phản ánh đầy đủ nội dung thông tin cơ quan nhà nước cung cấp;

- Dữ liệu mở được cung cấp là dữ liệu được cập nhật mới nhất;

- Dữ liệu mở phải có khả năng truy cập sử dụng được trên môi trường mạng;

- Dữ liệu mở phải đảm bảo khả năng thiết bị số có thể gửi, nhận, lưu trữ và xử lý được;

- Tổ chức, cá nhân được tự do truy cập sử dụng dữ liệu mở, không yêu cầu khai báo định danh khi khai thác, sử dụng dữ liệu mở;

- Dữ liệu mở ở định dạng mở;

- Sử dụng dữ liệu mở là miễn phí;

- Ưu tiên cung cấp dữ liệu mở đối với dữ liệu mà người dân, doanh nghiệp có nhu cầu sử dụng cao.

Quy định pháp lý tại Nghị định này quy định về điều kiện khai thác dữ liệu mở. Đây có thể coi như là các điều khoản trong **Giấy phép cung cấp dữ liệu mở** của các cơ quan nhà nước

- Cơ quan, tổ chức, cá nhân được phép tự do sao chép, chia sẻ, trao đổi, sử dụng dữ liệu mở hoặc kết hợp dữ liệu mở với dữ liệu khác; sử dụng dữ liệu mở vào sản phẩm, dịch vụ thương mại hoặc phi thương mại của mình.

- Cơ quan, tổ chức, cá nhân phải trích dẫn, ghi nhận thông tin sử dụng dữ liệu mở trong các sản phẩm, tài liệu liên quan có sử dụng dữ liệu mở.

- Dữ liệu mở của cơ quan nhà nước là dữ liệu nguyên trạng như được công bố; không bao gồm các hình thức trình bày và các thông tin phát sinh từ dữ liệu mở đã được cung cấp.

- Cơ quan, tổ chức, cá nhân không được bán dữ liệu mở đã được khai thác nguyên trạng từ cơ quan nhà nước cho tổ chức, cá nhân khác. Khi sử dụng dữ liệu mở trong sản phẩm, dịch vụ thương mại của mình phải cung cấp miễn phí dữ liệu mở kèm theo sản phẩm, dịch vụ thương mại đó.

- Cơ quan nhà nước không chịu trách nhiệm cho bất kỳ tổn thất hoặc thiệt hại phát sinh của cơ quan, tổ chức, cá nhân do việc sử dụng dữ liệu mở gây ra.

Ngoài ra, các cơ quan nhà nước phải xây dựng kế hoạch cung cấp dữ liệu mở: Trong phạm vi quản lý của mình, bộ, cơ quan ngang bộ, cơ quan thuộc Chính phủ, Ủy ban nhân dân các tỉnh, thành phố trực thuộc trung ương có trách nhiệm triển khai các nội dung sau:

- Xây dựng và triển khai kế hoạch về dữ liệu mở, bao gồm kế hoạch công bố dữ liệu mở, mức độ tối thiểu cần đạt được trong từng giai đoạn của kế hoạch.

- Xác định cơ chế thu thập, phân tích thông tin phản hồi của cá nhân, tổ chức về việc sử dụng dữ liệu mở; chỉ định đầu mối liên hệ tiếp nhận thông tin phản hồi của tổ chức, cá nhân về các vấn đề liên quan đến chất lượng, tính khả dụng, định dạng, sự tuân thủ các quy định pháp luật liên quan đến dữ liệu mở.

- Có các biện pháp phòng ngừa cần thiết để duy trì việc xây dựng và công bố dữ liệu mở có liên quan trực tiếp đến hoạt động bảo vệ an toàn cho con người và tài sản bị ảnh hưởng bởi kế hoạch triển khai dữ liệu mở trong phạm vi cơ quan mình.

- Thực hiện các giải pháp khuyến khích tổ chức, cá nhân sử dụng dữ liệu mở:

+ Tiếp nhận ý kiến phản hồi của tổ chức, cá nhân để xác định mức độ ưu tiên và thực hiện công bố dữ liệu mở phù hợp với nhu cầu;

+ Hỗ trợ các tổ chức, cá nhân trong xã hội, cộng đồng thực hiện các hoạt động xây dựng, đóng góp để mở rộng dữ liệu mở.

*b) Định hướng phát triển về dữ liệu mở*

Định hướng phát triển dữ liệu mở tại Việt Nam được quy định trong các văn bản sau:

- Quyết định của Thủ tướng Chính phủ số 749/QĐ-TTg ngày 3/6/2020 phê duyệt “Chương trình Chuyển đổi số quốc gia đến năm 2025, định hướng đến năm 2030” với những nội dung chủ yếu sau:

+ Tại phần một số nhiệm vụ và giải pháp phát triển Chính phủ số có quy định: Chuyển đổi số trong hoạt động của các cơ quan nhà nước, phát triển Chính phủ điện tử, hướng đến Chính phủ số trong đó tập trung phát triển hạ tầng số phục vụ các cơ quan nhà nước một cách tập trung, thông suốt; tạo lập dữ liệu mở dễ dàng truy cập, sử dụng, tăng cường công khai, minh bạch, phòng, chống tham nhũng, thúc đẩy phát triển các dịch vụ số trong nền kinh tế; cung cấp dịch vụ công trực tuyến mức độ 4 trên thiết bị di động thông minh để người dân, doanh nghiệp có trải nghiệm tốt nhất về dịch vụ, nhanh chóng, chính xác, không giấy tờ, giảm chi phí; cải thiện chỉ số xếp hạng quốc gia về Chính phủ điện tử.

+ Xây dựng Cổng dữ liệu quốc gia (data.gov.vn), tích hợp với thành phần dữ liệu mở của Hệ tri thức Việt số hóa, cung cấp dữ liệu mở, cung cấp hiện trạng và lộ trình xây dựng dữ liệu trong các cơ quan nhà nước, mức độ chia sẻ và sử dụng dữ liệu và các thông tin cần thiết để kết nối; đánh giá, xếp hạng công khai mức độ phát triển dữ liệu của các bộ, ngành, địa phương.

- Tại Quyết định số 942/QĐ-TTg ngày 15 tháng 6 năm 2021 Phê duyệt Chiến lược phát triển Chính phủ điện tử hướng tới Chính phủ số giai đoạn 2021 - 2025, định hướng đến năm 2030, với những nội dung sau:

+ Dữ liệu mở đã đưa vào quan điểm: *Dữ liệu là tài nguyên mới. Cơ quan nhà nước mở dữ liệu và cung cấp dữ liệu mở phục vụ phát triển Chính phủ số, kinh tế số, xã hội số*. Các cơ quan nhà nước kết nối, chia sẻ dữ liệu để người dân chỉ phải khai báo, cung cấp dữ liệu một lần cho các cơ quan nhà nước và các đơn vị cung ứng dịch vụ công thiết yếu.

+ Mục tiêu đến năm 2025: 100% cơ quan nhà nước cấp bộ, tỉnh tham gia mở dữ liệu và cung cấp dữ liệu mở phục vụ phát triển Chính phủ số, kinh tế số, xã hội số; Việt Nam thuộc nhóm 50 nước dẫn đầu về chỉ số dữ liệu mở.

+ Trong phát triển dữ liệu số quốc gia: Phát triển dữ liệu số quốc gia tạo nền tảng cho triển khai Chính phủ số, bảo đảm cung cấp dữ liệu số cho các dịch vụ công trực tuyến, chia sẻ dữ liệu thông suốt giữa các cơ quan nhà nước, cung cấp các bộ dữ liệu mở có chất lượng và giá trị khai thác cao, mở dữ liệu theo quy định của pháp luật để phát triển Chính phủ số, kinh tế số, xã hội số

+ Trong mục phát triển ứng dụng, các dịch vụ quốc gia: Xây dựng Cổng dữ liệu quốc gia là đầu mối cung cấp dữ liệu mở của cơ quan nhà nước trên môi trường mạng nhằm minh bạch, tăng cường chia sẻ dữ liệu, thúc đẩy sáng tạo, phát triển kinh tế số, xã hội số, bảo đảm an toàn thông tin của tổ chức, cá nhân theo quy định của pháp luật.

- Tại Quyết định số 411/QĐ-TTg ngày 31 tháng 3 năm hê duyệt Chiến lược quốc gia phát triển kinh tế số và xã hội số đến năm 2025, định hướng đến năm 2030, với những nội dung sau:

+ Tại phần phát triển dữ liệu số: Điểm đột phá là phát triển dữ liệu chủ trong các cơ sở dữ liệu quốc gia có độ chính xác cao làm dữ liệu gốc và *dữ liệu mở đặc thù của Việt Nam* để phân tích, xử lý, dán nhãn phục vụ phát triển trí tuệ nhân tạo.

+ Dữ liệu ngành nông nghiệp nông thôn: Xây dựng và tổ chức triển khai nền tảng dữ liệu số nông nghiệp trong đó cơ quan nhà nước đóng vai trò dẫn dắt, hợp tác xã và doanh nghiệp nông nghiệp đóng vai trò nòng cốt và người nông dân tham gia tích cực trong việc thu thập, chuẩn hóa, *mở dữ liệu và cung cấp dữ liệu mở phục vụ sản xuất nông nghiệp, tiêu thụ nông sản*; kịp thời cung cấp thông tin, tư vấn, dự báo tình hình cho nông dân

+ Ngành lao động, việc làm và an sinh xã hội: Xây dựng và tổ chức triển khai nền tảng dữ liệu số về lao động, việc làm và an sinh xã hội trong đó cơ quan nhà nước đóng vai trò dẫn dắt, tổ chức sử dụng lao động đóng vai trò nòng cốt và người lao động, người dân tham gia tích cực trong thu thập, chuẩn hóa, mở dữ liệu và *cung cấp dữ liệu mở phục vụ phân tích, dự báo, công bố thông tin thị trường lao động và thực hiện các chính sách an sinh xã hội*. Hình thành nền tảng hợp đồng lao động điện tử và thúc đẩy người lao động và tổ chức sử dụng lao động thực hiện ký kết hợp đồng lao động điện tử. Xây dựng và tổ chức triển khai sổ lao động điện tử cho người lao động.

+ Ngành du lịch: Xây dựng và tổ chức triển khai Nền tảng dữ liệu số du lịch trong đó cơ quan nhà nước đóng vai trò dẫn dắt, doanh nghiệp du lịch đóng vai trò nòng cốt và khách du lịch tham gia tích cực trong việc thu thập, chuẩn hóa, *mở dữ liệu và cung cấp dữ liệu mở phục vụ hoạch định chính sách, lập kế hoạch kinh doanh, tối ưu hóa, cải thiện chất lượng dịch vụ và đưa ra các gói ưu đãi cho khách du lịch*. Điều tra, đánh giá, phân loại và số hóa tài nguyên du lịch, cập nhật thường xuyên thông tin, dữ liệu số về khu du lịch, điểm du lịch, cơ sở dịch vụ du lịch, thị trường du lịch Việt Nam.

+ Các lĩnh vực khác: Xây dựng và tổ chức triển khai nền tảng dữ liệu số ngành, lĩnh vực trong đó cơ quan nhà nước đóng vai trò dẫn dắt, doanh nghiệp và cơ sở sản xuất kinh doanh đóng vai trò nòng cốt và người dân tham gia tích cực trong việc thu thập, chuẩn hóa, *mở dữ liệu và cung cấp dữ liệu mở phục vụ sản xuất, kinh doanh trong ngành, lĩnh vực*; kịp thời cung cấp thông tin, tư vấn, dự báo tình hình cho doanh nghiệp và người dân.

Như vậy, pháp lý và chính sách về cung cấp dữ liệu mở đã có định hướng rõ ràng để triển khai thúc đẩy mạnh mẽ trong thời gian tới.

## Sự cần thiết và vai trò của tiêu chuẩn về mô tả tập dữ liệu mở

Cung cấp dữ liệu mở của cơ quan nhà nước luôn được đánh giá là một chỉ tiêu thể hiện sự minh bạch của mỗi quốc gia. Vì vậy, các nước trên thế giới luôn quan tâm đến việc mở dữ liệu cho người dân, doanh nghiệp sử dụng. Việc cung cấp dữ liệu mở còn là hoạt động để thúc đẩy toàn xã hội tham gia nghiên cứu, đóng góp cho các hoạt động của cơ quan nhà nước trong các hoạt động xây dựng chính sách và xử lý các vấn đề xã hội. Cung cấp dữ liệu mở còn là phương tiện để thúc đẩy nghiên cứu khoa học công nghệ, thúc đẩy nghiên cứu trong lĩnh vực giáo dục, đào tạo.

Trước nhu cầu cấp thiết của xã hội về dữ liệu mở, Ngày 09/4/2020, Chính phủ đã ban hành Nghị định 47/2020/NĐ-CP về quản lý, kết nối và chia sẻ dữ liệu số. Đây có thể nói là văn bản đầu tiên của Việt Nam đề cập đến cung cấp dữ liệu mở của cơ quan nhà nước cho người dân, doanh nghiệp, tạo tiền đề cho các cơ quan nhà nước chia sẻ dữ liệu mở rộng rãi. Nghị định này đã quy định các cơ quan nhà nước phải cung cấp dữ liệu mở cho người dân, doanh nghiệp phục vụ phát triển kinh tế, xã hội tại Mục 3, Chương II. Dữ liệu mở phải có khả năng truy cập qua môi trường mạng, phải có khả năng đọc được bằng máy và cung cấp trên Cổng dữ liệu quốc gia, các cổng dữ liệu khác.

Về triển khai cung cấp dữ liệu mở, Đà Nẵng và Thành phố HCM đã triển khai xây dựng cổng dữ liệu cung cấp dữ liệu mở của mình để cung cấp dữ liệu mở. Trong năm 2020 vừa qua, Bộ TTTT cũng đã khởi động xây dựng cổng dữ liệu quốc gia cung cấp dữ liệu mở để làm đầu mối cung cấp dữ liệu mở của Chính phủ. Cổng dữ liệu quốc gia và cổng dữ liệu của các bộ, ngành, địa phương (nếu có) sẽ tạo thành hệ sinh thái dữ liệu mở của cơ quan nhà nước.

Như vậy, chính sách cung cấp dữ liệu đã có, các cổng dữ liệu từng bước hình thành. Tuy nhiên, do không có chuẩn nên các cổng dữ liệu hình thành sẽ rời rạc, không thể chia sẻ dữ liệu với nhau. Để dữ liệu của các bộ, ngành, địa phương có thể liên kết, tích hợp với nhau và tích hợp với Cổng dữ liệu quốc gia đòi hỏi phải có tiêu chuẩn chung về đặc tả các tập dữ liệu được cung cấp. Tiêu chuẩn chung này cũng làm cơ sở để người dân, doanh nghiệp dễ dàng tìm kiếm và khai thác các tập dữ liệu này.

*Vai trò của tiêu chuẩn về đặc tả tập dữ liệu mở*

Về bản chất, siêu dữ liệu hay mô tả về tập dữ liệu mở là thông tin có cấu trúc giúp truy xuất, sử dụng hoặc quản lý tài nguyên thông tin dễ dàng. Trên thực tế, siêu dữ liệu mô tả một tập dữ liệu và cấu trúc của nó, đồng thời giúp người dùng khám phá nó. Thông tin thường bao gồm các yếu tố cơ bản như: tiêu đề, ai đã xuất bản tập dữ liệu, khi nó được xuất bản, tần suất cập nhật và giấy phép nào được liên kết với tập dữ liệu. Chúng được phân loại là 'siêu dữ liệu mô tả' thay vì 'siêu dữ liệu cấu trúc', mô tả thông tin ví dụ về bố cục hoặc thành phần của đối tượng và các mối quan hệ của chúng (giống như các chương hoặc bảng trong sách).

Thông tin này được tổng hợp từ siêu dữ liệu của trang web; tuy nhiên, có vẻ như mặc dù siêu dữ liệu có mặt nhưng chỉ một phần tư có các trường được chú thích. Bản thân điều này có thể xuất phát từ nhiều vấn đề khác nhau, nhưng đặc biệt nổi bật là: dữ liệu này không đầy đủ vì các cổng dữ liệu được liệt kê không cung cấp siêu dữ liệu toàn diện mô tả nền tảng của riêng chúng. Theo thống kê của dataportal.org trong số 385 cổng có siêu dữ liệu được kiên kết chỉ 8 cung cấp liên kết đến tải xuống siêu dữ liệu đầy đủ và chỉ 12 cung cấp điểm API (giao diện lập trình ứng dụng) đang hoạt động.

Vì vậy, vấn đề cần quan tâm là với số lượng cổng dữ liệu cung cấp dữ liệu mở ngày càng tăng thì làm thế nào để các cổng dữ liệu có thể tương tác, chia sẻ dữ liệu với nhau, lập chỉ mục tìm kiếm nguồn tài nguyên dữ liệu mở thống nhất. Để giải quyết vấn đề này thì tiêu chuẩn về mô tả dữ liệu/siêu dữ liệu cần phải thống nhất sử dụng. Tuy nhiên tiêu chuẩn siêu dữ liệu nào được sử dụng là một vấn đề cần phải xác định.

Thông thường, các nền tảng cổng dữ liệu cung cấp dữ liệu mở được xây dựng tự xây dựng trên các tiêu chuẩn riêng của chúng hoặc sử dụng một tiêu chuẩn đã có sẵn. Hai nền tảng chủ yếu - CKAN và Socrata - đều được phát triển trên các tiêu chuẩn mô tả dữ liệu của riêng mình dựa trên các tiêu chuẩn chính như Dublin Corevà [RDF.](https://www.w3.org/TR/2004/REC-rdf-schema-20040210/) Các nền tảng có xu hướng sử dụng một tiêu chuẩn để tạo siêu dữ liệu hoặc kết hợp một vài tiêu chuẩn với nhau (ví dụ như CKAN vừa có tiêu chuẩn của riêng mình vừa hỗ trợ tiêu chuẩn chung là DCAT).

# Sở cứ xây dựng tiêu chuẩn

## Một số tiêu chuẩn, quy định trên thế giới

Các tiêu chuẩn thông dụng thường được sử dụng để mô tả siêu dữ liệu/dữ liệu mô tả trên thế giới gồm:

### Dublin Core

Dublin Core , còn được gọi là bộ phần tử siêu dữ liệu Dublin Core (DCMES) , là một tập hợp mười lăm (15) phần tử thuộc tính để mô tả tài nguyên. Lõi Dublin gồm mười lăm thuộc tính này đã được chính thức tiêu chuẩn hóa thành ISO 15836 vào tháng 2 năm 2009 và IETF RFC 5013. Sáng kiến ​​siêu dữ liệu Dublin Core (DCMI) hình thành nên Dublin Core. DCMI là một dự án của [Hiệp hội Khoa học và Công nghệ Thông tin](https://en.wikipedia.org/wiki/Association_for_Information_Science_and_Technology). Đây là một tổ chức phi lợi nhuận.  Dublin Core cũng được sử dụng như một thuộc tính mô tả siêu dữ liệu Dublin Core - một kiểu [siêu dữ liệu](https://en.wikipedia.org/wiki/Metadata) dựa trên nhiều Từ vựng của Khung mô tả tài nguyên (RDF), được đóng gói và giới hạn trong hồ sơ ứng dụng Dublin Core.

Các tài nguyên được mô tả bằng cách sử dụng Dublin Core có thể là tài nguyên kỹ thuật số (video, hình ảnh, trang web, v.v.) cũng như tài nguyên vật lý như sách hoặc tác phẩm nghệ thuật. Siêu dữ liệu Dublin Core có thể được sử dụng cho nhiều mục đích, từ mô tả tài nguyên đơn giản đến kết hợp các từ vựng siêu dữ liệu của các [tiêu chuẩn siêu dữ liệu](https://en.wikipedia.org/wiki/Metadata#Standards) khác nhau, cung cấp khả năng tương tác cho các từ vựng siêu dữ liệu trong đám mây [dữ liệu được liên kết](https://en.wikipedia.org/wiki/Linked_data) và triển khai web ngữ nghĩa ([Semantic Web](https://en.wikipedia.org/wiki/Semantic_Web)).

- Dublin Core Metadata bao gồm 15 thuộc tính sau:

1. Nhan đề: Nhan đề của tài liệu.

2. Tác giả: Tác giả của tài liệu, bao gồm cả tác giả cá nhân và tác giả tập thể.

3. Chủ đề: Chủ đề tài liệu đề cập dùng để phân loại tài liệu. Có thể thể hiện bằng từ, cụm từ/(Khung chủ đề), hoặc chỉ số phân loại/ (Khung phân loại).

4. Tóm tắt: Tóm tắt, mô tả nội dung tài liệu. Có thể bao gồm tóm tắt, chú thích, mục lục, đoạn văn bản để làm rõ nội dung...

5. Nhà xuất bản: Nhà xuất bản, nơi ban hành tài liệu có thể là tên cá nhân, tên cơ quan, tổ chức, dịch vụ...

6. Tác giả phụ: Tên những người cùng tham gia cộng tác đóng góp vào nội dung tài liệu, có thể là cá nhân, tổ chức...

7. Ngày tháng: Ngày, tháng ban hành tài liệu.

8. Loại (kiểu): Mô tả bản chất của tài liệu. Dùng các thuật ngữ mô tả phạm trù kiểu: trang chủ, bài báo, báo cáo, từ điển...

9. Khổ mẫu: Mô tả sự trình bày vật lý của tài liệu, có thể bao gồm; vật mang tin, kích cỡ độ dài, kiểu dữ liệu (.doc, .html, .jpg, xls, phần mềm....).

10. Định danh: Các thông tin về định danh tài liệu, các nguồn tham chiếu đến, hoặc chuỗi ký tự để định vị tài nguyên: URL (Uniform Resource Locators) (bắt đầu bằng http://), URN (Uniform Resource Name), ISBN (International Standard Book Number), ISSN (International Standard Serial Number), SICI (Serial Item & Contribution Identifier), ...

11.Nguồn: Các thông tin về xuất xứ của tài liệu, tham chiếu đến nguồn mà tài liệu hiện mô tả được trích ra/tạo ra, nguồn cũng có thể là: đường dẫn (URL), URN, ISBN, ISSN...

12. Ngôn ngữ: Các thông tin về ngôn ngữ, mô tả ngôn ngữ chính của tài liệu tham khảo.

13. Liên kết: Mô tả các thông tin liên quan đến tài liệu khác. có thể dùng đường dẫn (URL), URN, ISBN, ISSN...

14. Diện bao quát: Các thông tin liên quan đến phạm vi, quy mô hoặc mức độ bao quát của tài liệu. Phạm vi đó có thể là địa điểm, không gian hoặc thời gian, tọa độ...

15. Bản quyền: Các thông tin liên quan đến bản quyền của tài liệu.

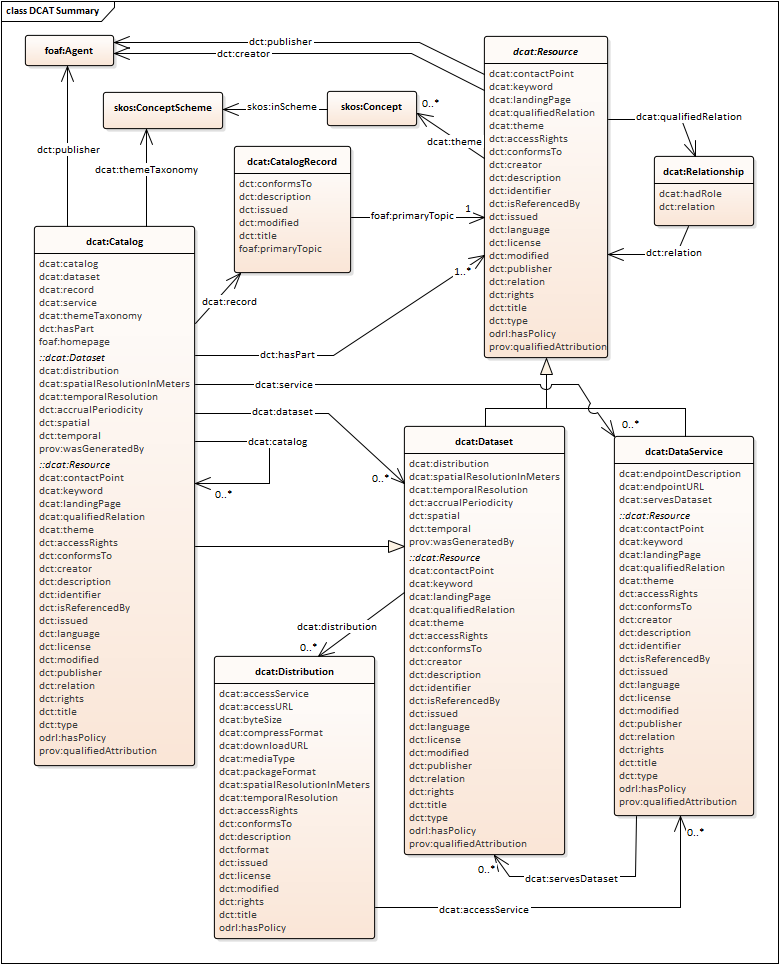
Dublin Core thường được sử dụng để đặc tả nguồn tài nguyên nói chung đặc biệt là tạo đặc tả cho các trang web. Nhiều tiêu chuẩn đặc tả dữ liệu được phát triển kế thừa trên Dublin Core.

### DCAT

DCAT là tiêu chuẩn siêu dữ liệu được sử dụng phổ biến nhất. DCAT phổ biến do tính linh hoạt và thiết kế phù hợp. Mục tiêu chính của DCAT là cải thiện khả năng tương tác của danh mục dữ liệu để các ứng dụng/cổng dữ liệu cung cấp dữ liệu mở có thể dễ dàng sử dụng siêu dữ liệu từ nhiều danh mục. Một phiên bản biến thể của về DCAT là [DCAT-AP,](https://joinup.ec.europa.eu/asset/dcat_application_profile/description) được thiết kế để đáp ứng nhu cầu xuất bản siêu dữ liệu trong liên minh Châu Âu.

Các biến thể khác của DCAT bao gồm: [Lược đồ siêu dữ liệu cốt lõi chung](https://project-open-data.cio.gov/schema/)(CCMS), cung cấp từ vựng mà lược đồ khác có thể ánh xạ tới [Lược đồ siêu dữ liệu](https://www.w3.org/TR/vocab-adms/)mô tả nội dung (ADMS) để mô tả Nội dung ngữ nghĩa trong danh mục hoặc gần đây nhất là tiêu chuẩn siêu dữ liệu cho phép chia sẻ siêu dữ liệu trên các danh mục dữ liệu khác nhau được gọi là [Giao thức tương tác danh mục dữ liệu](http://spec.dataportals.org/) (DCIP).

Mô hình dữ liệu DCAT rất phức tạp và ở mức khái quát cao. Để sử dụng, thông thường các như Mỹ, Singapore, Canada, EU phải đơn giản hóa và biến đổi để phù hợp theo mục đích của mình.



**Hình 3 - Cấu trúc dữ liệu của DCAT**

Các lớp cơ bản của DCAT bao gồm:

* Lớp Resource: Đây là lớp ảo/ trừu tượng để định nghĩa các thuộc tính chung cho các lớp mà kế thừa từ Resource. Lớp này định nghĩa các thuộc tính của một nguồn tài nguyên dữ liệu. Các thể hiện của tài nguyên dữ liệu bao gồm một tập dữ liệu mở (Dataset) hoặc một dịch vụ cung cấp dữ liệu mở (DataService)
* Lớp Catalog: Mô tả danh mục các tập dữ liệu được trao đổi có các thuộc tính chung.
* Lớp Dataset: đây là lớp lõi cơ bản nhất của DCAT mô tả thuộc tính của tập dữ liệu mở.
* Lớp DataService: đây là lớp mô tả tập dữ liệu mở được cung cấp dưới dạng dịch vụ dữ liệu (API).
* Lớp Distribution: mô tả một bản phân phối của tập dữ liệu mở (bao gồm cả tập dữ liệu mở hoặc dịch vụ API). Một Dataset hoặc DataService có thể có nhiều bản phân phối khác nhau.
* Lớp Agent: mô tả tổ chức cung cấp dữ liệu mở. Lớp này không được định nghĩa trong gói DCAT mà tham chiếu và được định nghĩa trong FOAF[[3]](#footnote-3) (là một gói định nghĩa các thực thể như cá nhân, tổ chức).

DCAT phát triển trên cơ sở sử dụng DublinCore.

### DCAT-AP

Danh mục từ vựng/mô tả tập dữ liệu cho cổng dữ liệu ở Châu Âu  (DCAT-AP) là một đặc tả dựa trên từ vựng Danh mục dữ liệu ([DCAT](http://www.w3.org/TR/vocab-dcat/)) của W3C để mô tả các bộ dữ liệu khu vực công ở Châu Âu. Trường hợp sử dụng cơ bản của nó là cho phép tìm kiếm trên cổng dữ liệu chéo cho các tập dữ liệu và làm cho dữ liệu khu vực công có thể tìm kiếm tốt hơn qua các biên giới và các lĩnh vực. Điều này có thể đạt được bằng cách trao đổi mô tả các tập dữ liệu giữa các cổng dữ liệu.

Đặc điểm kỹ thuật của DCAT-AP là sáng kiến ​​chung của DG CONNECT, Văn phòng Xuất bản EU và Chương trình ISA. Thông số kỹ thuật được xây dựng bởi một Nhóm công tác đa ngành với đại diện từ 16 Quốc gia Thành viên Châu Âu, một số Tổ chức Châu Âu và Hoa Kỳ.

Kiến trúc của DCAT-AP về cơ bản cũng gần tương tự như DCAT với một số thay đổi nhỏ và cụ thể hóa ngữ cảnh đặc thù của các các trường dữ liệu.

Một trong những khác biệt cơ bản giữa DCAT-AP và DCAT là DCAT-AP bổ sung thông tin về sự bắt buộc, tùy chọn, khuyến nghị của các lớp, các trường dữ liệu khi sử dụng.

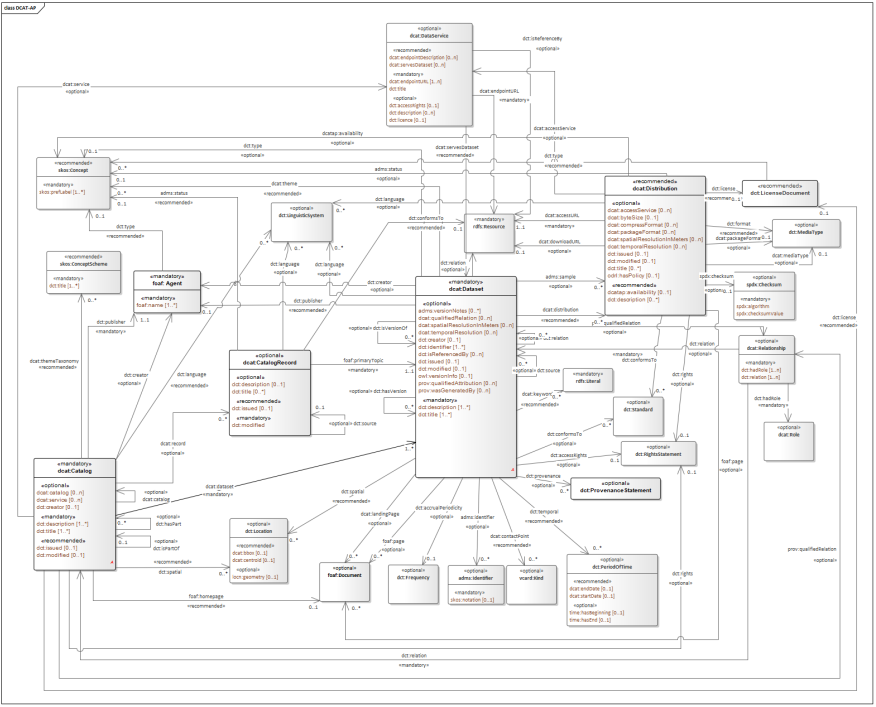
• Lớp bắt buộc: người nhận dữ liệu PHẢI có khả năng xử lý thông tin về các thể hiện của lớp; người gửi dữ liệu PHẢI cung cấp thông tin về các thể hiện của lớp.

• Lớp khuyến nghị: người gửi dữ liệu NÊN cung cấp thông tin về các thể hiện của lớp; người gửi dữ liệu PHẢI cung cấp thông tin về các thể hiện của lớp, nếu thông tin đó có sẵn; người nhận dữ liệu PHẢI có thể xử lý thông tin về các thể hiện của lớp.

• Lớp tùy chọn: người nhận PHẢI có thể xử lý thông tin về các thể hiện của lớp; người gửi CÓ THỂ cung cấp thông tin nhưng không có nghĩa vụ phải làm như vậy.

• Thuộc tính bắt buộc: người nhận PHẢI có thể xử lý thông tin cho thuộc tính đó; người gửi PHẢI cung cấp thông tin cho tài sản đó.

• Thuộc tính đề xuất: người nhận PHẢI có thể xử lý thông tin cho thuộc tính đó; người gửi NÊN cung cấp thông tin cho thuộc tính đó nếu nó có sẵn.



**Hình 4 - Cấu trúc dữ liệu của DCAT-AP**

• Thuộc tính tùy chọn: người nhận PHẢI có thể xử lý thông tin cho thuộc tính đó; người gửi CÓ THỂ cung cấp thông tin cho tài sản đó nhưng không có nghĩa vụ phải làm như vậy.

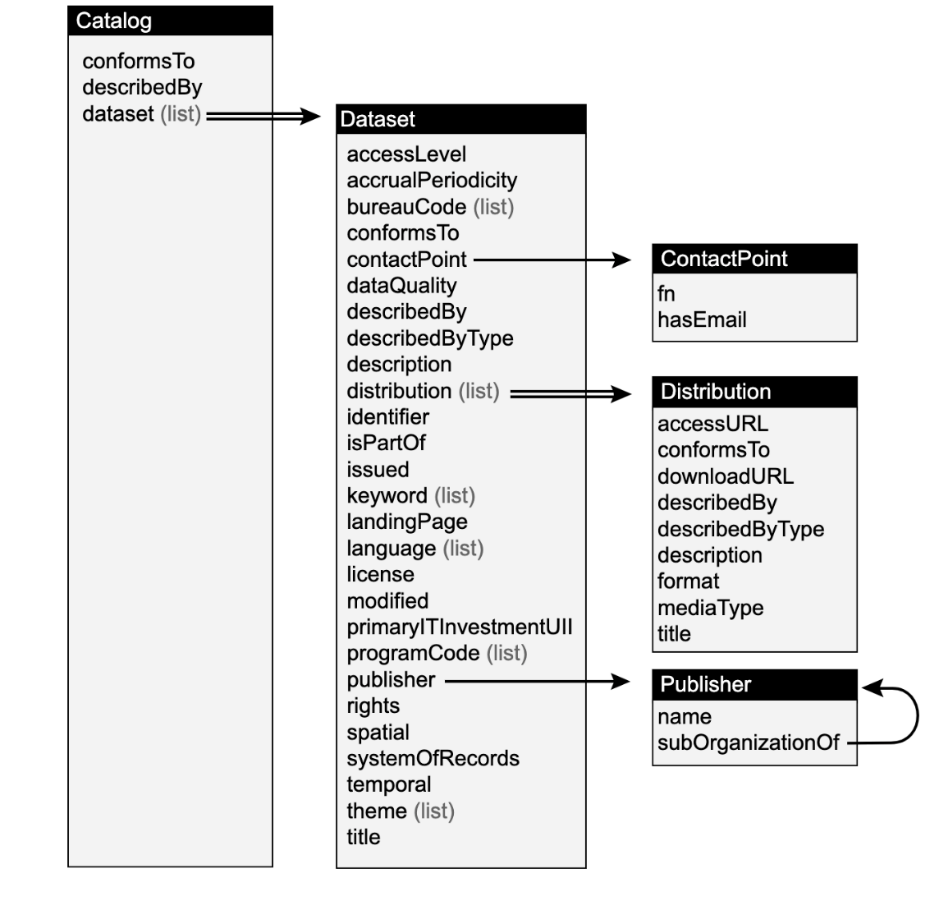
DCAT-AP kế thừa gần như toàn bộ DCAT.

### DCAT-US

DCAT-US là một lược đồ siêu dữ liệu Thiết lập một vốn từ vựng chung là chìa khóa để giao tiếp, chia sẻ dữ liệu mở trong cổng dữ liệu liên bang Mỹ. Lược đồ siêu dữ liệu dựa trên [DCAT](http://www.w3.org/TR/vocab-dcat/), một từ vựng phân cấp cụ thể cho các tập dữ liệu. Đặc tả DCAT-US xác định ba loại phần tử siêu dữ liệu: Bắt buộc, bắt buộc có điều kiện và các trường được mở rộng. Những yếu tố này được chọn để đại diện cho thông tin thường được tìm kiếm trên web. Để hỗ trợ người dùng các tiêu chuẩn siêu dữ liệu khác, [ánh xạ trường](https://resources.data.gov/resources/podm-field-mapping/#field-mappings) tới các phần tử tương đương trong các tiêu chuẩn khác được cung cấp.

DCAT-US là bản thu gọn của DCAT được chọn lọc các trường quan trọng và cần thiết nhất. Thay vì sử dụng XML, DCAT-US sử dụng ngôn ngữ JSON để mã hoá dữ liệu. [JSON](http://www.json.org/) là một định dạng trao đổi dữ liệu nhẹ, rất dễ đọc, phân tích cú pháp. Dựa trên một tập hợp con của ngôn ngữ lập trình JavaScript, JSON là một định dạng văn bản được tối ưu hóa cho việc trao đổi dữ liệu. JSON được xây dựng trên hai cấu trúc: (1) tập hợp các cặp tên / giá trị và (2) danh sách các giá trị có thứ tự.

Các lớp cơ bản của DCAT-US bao gồm:



**Hình 5 - Cấu trúc dữ liệu của DCAT-US**

Lớp Catalog: mô tả toàn bộ tệp danh mục Danh sách các tập dữ liệu mở bao gồm các trường để chỉ bối cảnh của lược đồ, chỉ dẫn lược đồ và là lớp bao chứa các tập dữ liệu trong nó.

Lớp tập trữ liệu: chứa các trường mô tả tập dữ liệu

Lớp phân phối: chứa các thông tin cụ thể của một bản phân phối dữ liệu.

Bên cạnh việc chọn lọc những trường thông tin quan trọng của DCAT, DCAT-US còn bổ sung thêm một số trường thông tin để lưu trữ thông tin theo yêu cầu đặc thù của Mỹ:

- Thêm trường **accessLevel** để giúp dễ dàng sắp xếp các tập dữ liệu thành ba danh mục hiện có: công khai, công khai bị hạn chế và không công khai. Trường này có nghĩa là một cơ quan cung cấp dữ liệu có thể chạy một bộ lọc cơ bản dựa trên danh mục dữ liệu của mình để tạo danh sách tập dữ liệu công khai hoặc một ngày nào đó có thể được cung cấp công khai (hoặc trong trường hợp dữ liệu bị hạn chế, có sẵn trong các điều kiện nhất định ). Trường này cũng giúp mọi người dễ dàng tạo danh sách các tập dữ liệu có thể sẵn có nhưng chưa được phát hành bằng cách lọc **accessLevel** thành public và **accessURL** để trống .

- Thêm trường **rights** cho người quản lý dữ liệu để giải thích cách truy cập các tập dữ liệu công khai bị hạn chế và để các cơ quan có nơi để ghi lại (ngay cả khi chỉ trong nội bộ) lý do không phát hành tập dữ liệu không công khai.

Thêm trường **systemOfRecords** cho người quản lý dữ liệu để tùy chọn liên kết đến URL Thông báo Hệ thống Hồ sơ có liên quan. Hệ thống Hồ sơ là một nhóm bất kỳ hồ sơ nào dưới sự kiểm soát của bất kỳ cơ quan nào mà từ đó thông tin được truy xuất bằng tên của cá nhân hoặc bằng một số nhận dạng, ký hiệu hoặc số nhận dạng khác được chỉ định cho cá nhân đó.

Thêm trường **bureauCode** để đảm bảo mọi tập dữ liệu được cung cấp bởi một cơ quan với mã của cơ quan đó gắn kèm

Thêm trường **programCode** để đảm bảo rằng khi có thể, mọi tập dữ liệu được kết nối theo cách tiêu chuẩn với một chương trình/dự án cụ thể của đơn vị cung cấp dữ liệu

Thêm **dataQuality** để cho biết liệu dữ liệu có đáp ứng Nguyên tắc chất lượng thông tin của một cơ quan hay không.

DCAT-US lại không giống như DCAT mà có sự khác biệt khá lớn cả về công nghệ và cấu trúc dữ liệu. DCAT-US kế thừa các trường dữ liệu từ DCAT.

### ISO 19115

Đây là một tiêu chuẩn của tổ chức tiêu chuẩn quốc tế mô tả lược đồ thông tin địa lý và các dịch vụ. Tiêu chuẩn này cung cấp thông tin về việc xác định, mức độ, chất lượng, lược đồ không gian và thời gian, tham chiếu không gian và phân phối dữ liệu địa lý kỹ thuật số.

Chính phủ sử dụng Chuẩn ISO 19115 xác định việc mô tả dữ liệu GIS của quốc gia dưới dịnh dạng XML nhằm có thể dễ dàng đưa vào CSDL để quản lý, dễ dàng bảo trì và có khả năng chia sẻ thông tin về dữ liệu giữa các hệ thống thông tin địa lý.

Việc xây dựng lược đồ ứng dụng dữ liệu thông tin địa lý cần thiết trong việc lưu trữ, sử dụng, chia sẻ dữ liệu, giúp các hệ thống GIS có thể làm việc cùng nhau một cách hiệu quả.

ISO 19115 do Ban kỹ thuật ISO / TC 211, Thông tin địa lý/Geomatics chuẩn bị.

 Được tài trợ bởi ISO (Tổ chức Tiêu chuẩn hóa Quốc tế), ấn bản đầu tiên của [ISO 19115](http://www.iso.org/iso/catalogue_detail.htm?csnumber=26020) được xuất bản vào năm 2003. Kể từ đó, nó được chia thành các phần: ISO 19115-1: 2014 bao gồm các nguyên tắc cơ bản của tiêu chuẩn; ISO 19115-2: 2009 chứa các phần mở rộng cho hình ảnh và dữ liệu dạng lưới; và ISO / TS 19115-3: 2016 cung cấp triển khai lược đồ XML cho các khái niệm cơ bản tương thích với ISO / TS 19138: 2007 (Siêu dữ liệu thông tin địa lý XML, hoặc GMD).

Mục đích của ISO 19115:

* Cho phép các nhà cung cấp tài nguyên thông tin mô tả một cách hiệu quả và đầy đủ các tài nguyên của họ.
* Tạo điều kiện thuận lợi cho việc tổ chức và quản lý siêu dữ liệu cho các nguồn thông tin.
* Cho phép sử dụng thích hợp các nguồn thông tin thông qua sự hiểu biết chính xác về các đặc điểm của chúng.
* Tạo điều kiện thuận lợi cho việc khám phá, truy cập, truy xuất và tái sử dụng tài nguyên.
* Cho phép người dùng xác định liệu một nguồn thông tin có được sử dụng cho họ hay không.

ISO 19115 có những đặc điểm sau:

* Tạo mô hình để mô tả thông tin hoặc tài nguyên có thể có phạm vi địa lý, xác định các phần tử siêu dữ liệu, thuộc tính của chúng và mối quan hệ giữa các phần tử, đồng thời thiết lập một bộ thuật ngữ, định nghĩa và quy trình mở rộng siêu dữ liệu chung.
* Mô hình siêu dữ liệu được mô tả cho phép triển khai các tiện ích mở rộng người dùng dành riêng cho miền dựa trên một mẫu chung để tạo điều kiện triển khai phần mềm sử dụng các tiện ích mở rộng đó.
* Khi tiêu chuẩn ISO được nâng cấp, sửa đổi, khả năng tương thích ngược được cung cấp bằng cách sử dụng dịch vụ chuyển đổi. Các phiên bản siêu dữ liệu trong quá khứ có thể tiếp tục tham chiếu / sử dụng phiên bản trước đó.
* ISO 19115 thường được sử dụng để đặc tả các tập dữ liệu trong cổng dữ liệu địa lý.

### CKAN-API

Nền tảng CKAN sử dụng một định dạng riêng để phục vụ mục đích quản lý đặc tả các tập dữ liệu mở trong nền tảng. Các thông điệp dữ liệu chia sẻ từ các dịch vụ của CKAN có định dạng cấu trúc gọi là CKAN-API.

Do định dạng CKAN-API của riêng nền tảng nên nó phụ thuộc lớn vào thiết kế của nền tảng CKAN và theo các phiên bản CKAN phát hành.

Ưu điển của định dạng CKAN-API là nếu các cổng dữ liệu cùng sử dụng CKAN thì sẽ dễ dàng tương tác, kết nối đồng bộ dữ liệu, việc chuyển đổi dữ liệu không cần thiết trong trường hợp này do tương tích hoàn toàn.

Tuy nhiên, do là chuẩn của chính nền tảng CKAN nên nó không có sự độc lập tương đối. Trường hợp các cổng dữ liệu sử dụng các công nghệ khác nhau thì CKAN-API không sử dụng được mà bắt buộc phải sử dụng qua tiêu chuẩn trung gian là DCAT.

Chính vì thế, cộng đồng phát triển nền tảng CKAN đã xây dựng thành phần chuyển đổi từ sang sử dụng DCAT nằm trong nền tảng CKAN để dễ dàng tương thích với các công nghệ khác nhau.

Sự khác biệt giữa CKAN-API và DCAT như trong bảng so sánh sau:

Đối với lớp tập dữ liệu: Dataset

| **Tên** | **CKAN API** | **DCAT** |
| --- | --- | --- |
| Tiêu đề | *title* | [dct:title](http://www.w3.org/TR/vocab-dcat/#Property:dataset_title) |
| Mô tả | *notes* | [dct:description](http://www.w3.org/TR/vocab-dcat/#Property:dataset_description) |
| Nhãn | *tags* | [dcat:keyword](http://www.w3.org/TR/vocab-dcat/#Property:dataset_keyword) |
| Thời gian cập nhật | *n/a* | [dct:modified](http://www.w3.org/TR/vocab-dcat/#Property:dataset_update_date) |
| Cơ quan cung cấp | *organization* → *title* | [dct:publisher](http://www.w3.org/TR/vocab-dcat/#Property:dataset_publisher) → [foaf:name](http://xmlns.com/foaf/spec/#term_name) |
| Cơ quan chủ quản | *n/a* | [dct:publisher](http://www.w3.org/TR/vocab-dcat/#Property:dataset_publisher) → [org:subOrganizationOf](http://www.w3.org/TR/vocab-org/#org:subOrganizationOf) |
| Tên liên hệ | *maintainer* | [dcat:contactPoint](http://www.w3.org/TR/vocab-dcat/#Property:dataset_contactPoint) → [vcard:fn](http://www.w3.org/TR/vcard-rdf/#d4e199) |
| Thư điện tử liên hệ | *maintainer\_email* | [dcat:contactPoint](http://www.w3.org/TR/vocab-dcat/#Property:dataset_contactPoint) → [vcard:hasEmail](http://www.w3.org/TR/vcard-rdf/#d4e183) |
| Mã định danh | *id* | [dct:identifier](http://www.w3.org/TR/vocab-dcat/#Property:dataset_identifier) |
| Các bản phân phối | *resources* | [dcat:distribution](http://www.w3.org/TR/vocab-dcat/#Property:dataset_distribution) |
| Giấy phép | *license\_title* | [dct:license](http://www.w3.org/TR/vocab-dcat/#Property:distribution_license) |
| Quyền | *n/a* | [dct:rights](http://www.w3.org/TR/vocab-dcat/#Property:distribution_rights) |
| Địa chỉ dịch vụ | *resources* → *url* | [dcat:accessURL](http://www.w3.org/TR/vocab-dcat/#Property:distribution_accessurl) |
| Vị trí không gian | *n/a* | [dct:spatial](http://www.w3.org/TR/vocab-dcat/#Property:dataset_spatial) |
| Khoảng thời gian thu thập | *n/a* | [dct:temporal](http://www.w3.org/TR/vocab-dcat/#Property:dataset_temporal) |
| Ngày cung cấp | *n/a* | [dct:issued](http://www.w3.org/TR/vocab-dcat/#Property:dataset_release_date) |
| Tần suất cập nhật dữ liệu | *n/a* | [dct:accrualPeriodicity](http://www.w3.org/TR/vocab-dcat/#Property:dataset_frequency) |
| Ngôn ngữ | *n/a* | [dct:language](http://www.w3.org/TR/vocab-dcat/#Property:dataset_language) |
| Chủ đề | *groups* | [dcat:theme](http://www.w3.org/TR/vocab-dcat/#Property:dataset_theme) |
| Các tài liệu liên quan | *n/a* | [dct:references](http://dublincore.org/documents/dcmi-terms/#terms-references) |
| Địa chỉ trang chủ | *n/a* | [dcat:landingPage](http://www.w3.org/ns/dcat#Property:dataset_landingpage) |

Đối với Lớp Phân phối (Distribution):

| **Tên trường** | **CKAN API** | **DCAT** |
| --- | --- | --- |
| Địa chỉ tải về | resources → url | [dcat:downloadURL](http://www.w3.org/TR/vocab-dcat/#Property:distribution_downloadurl) |
| Địa chỉ truy cập | resources → url | [dcat:accessURL](http://www.w3.org/TR/vocab-dcat/#Property:distribution_accessurl) |
| Loại dữ liệu | resources → mimetype | [dcat:mediaType](http://www.w3.org/TR/vocab-dcat/#Property:distribution_media_type) |
| Định dạng | resources → format | [dct:format](http://www.w3.org/TR/vocab-dcat/#Property:distribution_format) |
| Tiêu đề | resources → name | [dct:title](http://www.w3.org/TR/vocab-dcat/#Property:distribution_title) |
| Mô tả | resources → description | [dct:description](http://www.w3.org/TR/vocab-dcat/#Property:distribution_description) |

Như vậy, có sự khác biệt nhất định giữa CKAN API và DCAT. Một số trường dữ liệu có trong DCAT nhưng không có trong CKAN-API. Điều này cũng nảy sinh một vấn đề khi CKAN tiếp nhận dữ liệu theo DCAT nhưng có thể bỏ qua một số trường dữ liệu. Điều đó cũng có nghĩa một số trường trong DCAT không thực sự cần thiết khi triển khai bởi CKAN.

### Chuẩn dữ liệu của cổng dữ liệu Singapore

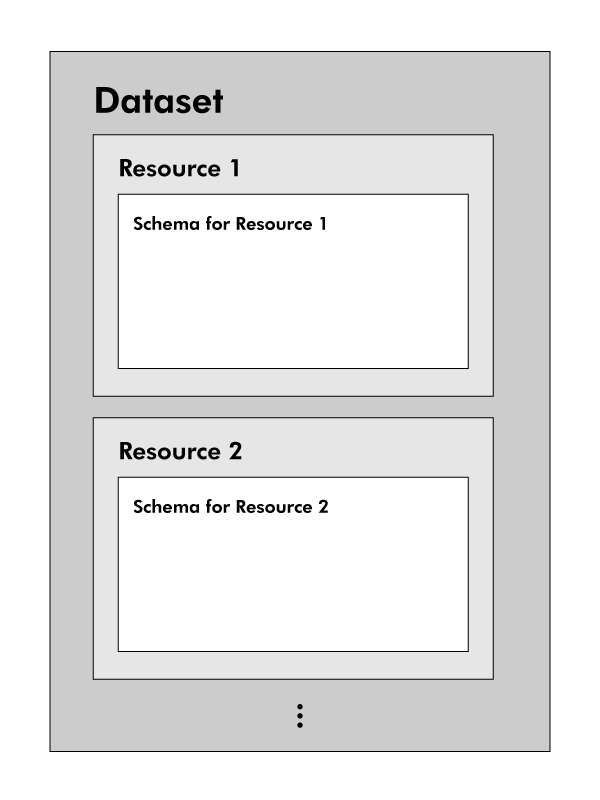
Chuẩn lữ liệu của cổng Singore hơi đặc thù, khác biệt với DCAT nhưng lại có nét tương tự như DCAT-US.

Có 3 cấp siêu dữ liệu được yêu cầu cho dữ liệu.

*- Siêu dữ liệu tập dữ liệu:* Đây là siêu dữ liệu chung cho một tập dữ liệu, bao gồm thông tin như danh mục, cơ quan xuất bản và tần suất cập nhật.

*- Siêu dữ liệu tài nguyên:* Một tập dữ liệu có thể có nhiều tài nguyên. Mỗi tài nguyên có một tiêu đề duy nhất, ngày cập nhật lần cuối và các chú thích dành riêng cho tài nguyên đó.

*- Lược đồ:* Điều này xác định các đặc điểm của mỗi cột trong tệp tài nguyên, chẳng hạn như loại biến, định dạng và chú thích cuối trang (nếu có).



**Hình 6 - Cấu trúc dữ liệu của Singapore**

Do Singapore sử dụng CKAN nên có thể nói: Singapore hỗ trợ cả chuẩn CKAN và DCAT và chuẩn riêng của mình.

## Đánh giá các tiêu chuẩn

Qua việc rà soát các tiêu chuẩn quốc tế, có thể rút ra một số kết luận sau:

* Các tiêu chuẩn đều là các khuyến nghị áp dụng (Recommendation), là hình thức quy chuẩn không thành văn và mang tính quy định kỹ thuật;
* Các tiêu chuẩn mang tính quốc gia và của tổ chức tiêu chuẩn cho cổng dữ liệu cung cấp dữ liệu mở là: DCAT, DCAT-AP, DCAT-US, Singapore;
* Các tiêu chuẩn mô tả dữ liệu nhưng không sử dụng trực tiếp mà gián tiếp qua tiêu chuẩn khác là: DublinCore, ISO 19115;
* Các tiêu chuẩn riêng của nền tảng: CKAN-API, Socrate, DKAN, OpenDataSoft;
* Gần như các tiêu chuẩn đều lấy DCAT làm nền tảng phát triển hoặc tương thích với DCAT;
* DCAT là một tiêu chuẩn phức tạp và do tham chiếu đến hơn 10 tiêu chuẩn khác;
* DCAT-AP gần như tương đương với DCAT;
* DCAT-US sử dụng nền tảng công nghệ JSON thay vì RFD như DCAT nhưng tương thích về lớp và trường dữ liệu. DCAT-US mang tính thực dụng và dễ dàng áp dụng;
* Chuẩn của Singapore cũng đơn giản như DCAT-AP dễ dàng áp dụng do tính đơn giản của chuẩn này.

## Thông tin tham chiếu của Asean và đánh giá chuẩn mô tả dữ liệu mở tại Việt Nam

Theo đánh giá về mô tả dữ liệu mở của nhóm khảo sát ASEAN về các cổng dữ liệu quốc gia của các nước Đông Nam Á, hiện nay các cổng dữ liệu quốc gia của các nước đang cung cấp các trường thông tin mô tả dữ liệu mở như sau:

Đối với các trường thông tin của tập dữ liệu (Dataset).

|  |  |
| --- | --- |
| **Tên trường** | **Có xuất hiện tại cổng dữ liệu của Việt Nam** |
| **Tiêu đề** | Tiêu đề |
| **Sự mô tả** | Sự mô tả |
| **Định dạng** | Định dạng |
| **Định danh** |  |
| **Cơ quan cung cấp** | Cơ quan cung cấp |
| **Thời gian cập nhật** | Cập nhật mới nhất |
| **Thời gian cung cấp** | Ngày thành lập |
| **Thời gian thu thập** |  |
| **Tần suất cập nhật** |  |
| **Nguồn** | Nguồn dữ liệu |
| **Địa chỉ nguồn (trang chủ)** |  |
| **Giấy phép** |  |
| **Từ khóa** |  |
| **Chủ đề** |  |
| **Cấp độ truy cập** |  |
| **Khác** |  |

Đối với Mô tả nguồn dữ liệu hay bản phân phối dữ liệu tùy theo tên gọi của mỗi quốc gia tiêu chuẩn khác nhau (Resource/Distribution).

|  |  |
| --- | --- |
| **Tên trường** | **Có xuất hiện tại cổng dữ liệu của Việt Nam** |
| **Tên** | Tên |
| **Sự mô tả** |  |
| **Định dạng** | Định dạng |
| **Định danh** |  |
| **Cập nhật mới nhất** | Dữ liệu cập nhật lần cuối |
| **Ngày tạo** | Ngày thành lập |
| **Giấy phép** | Giấy phép |
| **Khác** |  |

*Một số nhận xét:*

* Các cổng dữ liệu quốc gia của các nước cũng đã định vị được đặc tả tập dữ liệu của mình.
* Mô tả các tập dữ liệu của các quốc gia có những trường chung nhất định. Đó là các trường đã được đưa vào cột đầu tiên của bảng so sánh. Đây là những trường quan trọng nhất mà đặc tả tập dữ liệu phải có.
* Tùy theo điều kiện của mỗi quốc gia mà có thể có hoặc không các trường dữ liệu bổ sung để phục vụ mục đích của mình.

- Các trường mô tả tập dữ liệu này đều nhỏ hơn rất nhiều so với tiêu chuẩn DCAT. Do vậy, ta có thể thấy rằng DCAT là tiêu chuẩn phổ quát làm nền tảng chung, nhiều trường dữ liệu không cần thiết có thể được giản lược đi để sử dụng cho mỗi quốc gia khác nhau.

  Tuy nhiên, các thông tin này được thu thập trực tiếp từ giao diện bên ngoài của tập dữ liệu, chưa phản ánh quy định của các nước này đối với tiêu chuẩn về đặc tả tập dữ liệu mở được ban hành.

# Phương án xây dựng và giải thích nội dung tiêu chuẩn

Cổng dữ liệu cung cấp dữ liệu mở tại Việt Nam đã triển khai bao gồm: Cổng dữ liệu quốc gia; Cổng dữ liệu Thành phố Hồ Chí Minh, Cổng dữ liệu Đà Nẵng; Cổng dữ liệu ItriThuc. Trong 4 cổng dữ liệu đã được xác lập tại Việt Nam có 3 cổng dữ liệu cung cấp dữ liệu mở sử dụng nền tảng công nghệ CKAN, một cổng dữ liệu tự xây dựng.

Đối với chuẩn về đặc tả tập dữ liệu, hiện tại Việt Nam chưa ban hành, các cổng dữ liệu này được triển khai chủ yếu dựa theo nền tảng sử dụng mà chưa có quy định cụ thể về đặc tả dữ liệu.

Đánh giá của ASEAN cũng đã chỉ ra các trường dữ liệu cơ bản nhất và so sánh với hiện trạng dữ liệu Việt Nam.

Đặc tả dữ liệu tập trung vào hai đối tượng chính là Dataset và Distribution

Chỉ sử dụng một số trường quan trọng nhất để đưa vào quy định.

Hiện trạng pháp lý đã có quy định về dữ liệu mở tại Việt Nam, các định hướng triển khai dữ liệu mở trong các Chiến lược, Chương trình triển khai trong thời gian tới. Do vậy, việc chuẩn hóa, ban hành tiêu chuẩn về đặc tả tập dữ liệu mở là hết sức cần thiết.

## Đánh giá về việc khả năng chấp thuận các tiêu chuẩn nước ngoài

Xu hướng phát triển dữ liệu mở đang phát triển mạnh mẽ và quy định pháp lý của Việt Nam đã sẵn sàng, điều này đòi hỏi phải xây dựng và ban hành các quy định về mô tả tập dữ liệu mở. Vấn đề đặt ra là có chấp nhận được nguyên trạng các tiêu chuẩn nước ngoài hay không? Qua đánh giá và phân tích chi tiết, không thể chấp nhận nguyên trạng một trong số các tiêu chuẩn nước ngoài được. Lý do như sau:

|  |  |  |
| --- | --- | --- |
| **Tiêu chuẩn/ quy định** | **Ưu điểm** | **Khó khăn khi chấp nhận nguyên trạng** |
| DCAT | Chuẩn gốc  Sử dụng rộng rãi  Trung lập công nghệ | Quá phức tạp.  Có tham chiếu đến rất nhiều chuẩn khác ( như trong bảng phụ lục) mà Việt Nam chưa có.  Chuẩn không thành văn mà ban hành dưới dạng khuyến nghị kỹ thuật. |
| DCAT-AP | Tương tự như DCAT | Tương tự như DCAT |
| DCAT-US | Đơn giản  Sử dụng JSON nhỏ, nhẹ | Phụ thuộc vào đặc thù riêng của cổng DataGov của Mỹ, một số trường dữ liệu chỉ phù hợp với Mỹ |
| Singapore | Đơn giản  Sử dụng JSON nhỏ nhẹ | Tương tự như DCAT-US phụ thuộc vào Cổng của Singapore đã được sửa đổi cho phù hợp riêng với Singapore |
| CKAN-API | Có sẵn trong nền tảng CKAN thông dụng nhất | Phụ thuộc vào một nền tảng cụ thể, không trung lập về công nghệ, khó khăn khi triển khai với nền tảng khá |
| Dublin Core  ISO 19115 | Chuẩn quốc tế | Đây là chuẩn gốc, sử dụng để tham chiếu chứ không sử dụng trực tiếp mô tả cho tập dữ liệu mở |
| Cocrata  DKAN  OpenDataSoft |  | Chỉ đáp ứng đúng một nền tảng, các nền tảng này đều ít được sử dụng rộng rãi. |

Như vậy, không thể sử dụng chấp nhận nguyên vẹn một tiêu chuẩn hiện có mà phải có sự kế thừa và phát triển một tiêu chuẩn cho Việt Nam.

## Yêu cầu cho phương án đề xuất

Để xây dựng một tiêu chuẩn cho Việt Nam, cần phải đặt ra các yêu cầu thực tế để phát triển;

- Phải tương thích nhất định với tiêu chuẩn quốc tế. Cụ thể ở đây là DCAT là tiêu chuẩn gốc về mô tả tập dữ liệu mở.

- Phải đủ đơn giản để dễ dàng triển khai tại Việt Nam. Điều này mới có thể thúc đẩy sự tuân thủ. Đây cũng là kinh nghiệm của Mỹ và Singapore

- Không quá lệ thuộc vào nhiều tiêu chuẩn chưa được định nghĩa như DCAT vì điều kiện hiện tại Việt Nam chưa có.

- Có thể sử dụng với nhiều ngôn ngữ mã hoá dữ liệu khác nhau, cụ thể ở đây là JSON và XML.

- Linh động có thể mở rộng các thuộc tính dữ liệu để đáp ứng yêu cầu chuyên sâu. Ví dụ như dữ liệu không gian.

## Phương án xây dựng chuẩn về mô tả tập dữ liệu mở

Với tất cả các yêu cầu trên, đề xuất phương pháp luận xây dựng như sau:

* Tên của tiêu chuẩn mô tả tập dữ liệu mở DCAT-VN;
* Kế thừa và tương thích với DCAT. Cụ thể ở đây là mô hình các lớp và trường thông tin sử dụng đồng bộ với DCAT để đảm bảo có thể chuyển dữ liệu theo chuẩn DCAT sang DCAT-VN;
* Tham khảo DCAT-US, Singapre để chọn lọc các trường cơ bản nhất mà các nước quan tâm sử dụng để áp dung cho Việt Nam;
* Tương thích và sử dụng cả ngôn ngữ JSON và XML;
* Đảm bảo sự tương thích với CKAN do khả năng CKAN là nền tảng mở thông dụng nhất, rất nhiều cơ quan sử dụng;



**Hình 7 - Mô hình quan hệ giữa chuẩn DCAT-VN đề xuất với các chuẩn thông dụng trên thế giới**

* Đối với trường hợp mở rộng khi cần thiết thì tương thích với ISO 19115 về dữ liệu không gian;
* Đối sánh và sử dụng các trường thông tin thông dụng nhất theo báo cáo khảo sát của nhóm Phát triển thư mục dữ liệu mở ASEAN;
* Có sự điều chỉnh để phù hợp với Việt Nam. Ví dụ như sử dụng mã định danh của cơ quan nhà nước;
* Không sử dụng các lớp, trường dữ liệu trong DCAT mà tham chiếu hoặc phụ thuộc đến các tiêu chuẩn khác Việt Nam chưa có hoặc không cần thiết khi so sánh với danh mục các các tiêu chuẩn DCAT-US, Singapore và khảo sát của ASEAN về dữ liệu mở tại các nước Đông Nam Á.

## Nội dung chuẩn mô tả tập dữ liệu mở DCAT-VN



**Hình 8 - Mô hình dữ liệu đặc tả tập dữ liệu mở DCAT-VN đề xuất**

Nội dung chi tiết:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Lớp** |  | | **Trường dữ liệu** |  | **Tên** | **Tham khảo, tương đồng với** |
| **Catalog** | | |  |  |  | DCAT |
|  | * title | | |  | Tiêu đề | DCAT |
|  | * description | | |  | Mô tả | DCAT |
|  | * dataset | | |  | Tập dữ liệu | DCAT |
| **Dataset** | | * title | |  | Tiêu đề | DCAT |
|  |  | * description | | | Mô tả | DCAT |
|  |  | * keyword | |  | Từ khóa | DCAT |
|  |  | * contactPoint | | | Liên hệ | DCAT |
|  |  |  | * fn | | Tên liên hệ | DCAT-US  Đơn giản khi ứng dụng |
|  |  |  | * email | | Địa chỉ email liên hệ | DCAT-US  Đơn giản ứng dụng |
|  |  | * distribution | | | Bản phân phối | DCAT |
| **Distribution** | | | * title | | Tiêu đề | DCAT |
|  |  |  | * description | | Mô tả | DCAT |
|  |  |  | * accessURL | | Đường dẫn dịch vụ | DCAT |
|  |  |  | * downloadURL | | Đường dẫn tải về | DCAT |
|  |  |  | * format | | Kiểu dữ liệu | DCAT |
|  |  |  | * mediaType | | Định dạng | DCAT |
|  |  |  | * modified | | Ngày cập nhật | DCAT |
|  |  | * publisher | | | Cơ quan, tổ chức cung cấp | DCAT |
| **Agent** |  |  | * name | | Tên cơ quan, tổ chức | DCAT |
|  |  |  | * type | | Loại cơ quan, tổ chức | Đặc thù Việt Nam để phân biệt giữa CQNN và tổ chức cung cấp |
|  |  |  | * code | | Mã cơ quan, tổ chức | Đặc thù Việt Nam để lưu mã cơ quan nhà nước |
|  |  | * spatial | |  | Phạm vi địa lý | DCAT-US  Đảm bảo sự linh động khi sử dụng |
|  |  | * temporal | |  | Thời gian thu thập | DCAT |
|  |  | * accrualPeriodicity | | | Tần xuất cung cấp | DCAT |
|  |  | * identifier | |  | Định danh tập dữ liệu | DCAT |
|  |  | * theme | |  | Chủ đề | DCAT |
|  |  | * landingPage | | | Trang chủ của Dataset | DCAT |
|  |  | * Issued | |  | Ngày cung cấp | DCAT |
|  |  | * modified | |  | Ngày cập nhật | DCAT |
|  |  | * licence | |  | Giấy phép | DCAT |
|  | * homePage | | |  | Trang chủ Catalog | DCAT |

Các lớp Catalog, Dataset, Distribution, Agent tương thích với DCAT, DCAT-AP, DCAT-US. Không sử dụng lớp DataService do học tập kinh nghiệm của DCAT-US là hợp nhất với Distribution vì có đặc tính tương tự.

# Bảng đối thiếu nội dung tiêu chuẩn với tài liệu tham khảo

|  |  |  |  |
| --- | --- | --- | --- |
| **Điều, mục** | **Nội dung** | **Tài liệu tham khảo** | **Biên soạn/chấp nhận** |
| 1 | Phạm vi áp dụng |  | Tự biên soạn |
| 2 | Tài liệu viện dẫn |  | Tự biên soạn |
| 3 | Thuật ngữ, định nghĩa và chữ viết tắt |  | Tự biên soạn |
| 4 | Mô hình cấu trúc dữ liệu đặc tả tập dữ liệu mở |  |  |
| 4.1 | Lược đồ mô hình dữ liệu |  | Tự biên soạn |
| 4.2 | Mô tả các lớp dữ liệu |  |  |
| 4.2.1 | Lớp Catalog | DCAT | Chấp nhận một phần và có điều chỉnh |
| 4.2.2 | Lớp Dataset | DCAT | Chấp nhận một phần và có điều chỉnh |
| 4.2.3 | Lớp Distribution | DCAT | Chấp nhận một phần và có điều chỉnh |
| 4.2.4 | Lớp Agent | DCAT | Chấp nhận một phần và có điều chỉnh |
| 4.3 | Cấu trúc các lớp dữ liệu đặc tả |  |  |
| 4.3.1 | Catalog | DCAT | Tự biên soạn, tiếp thu một phần (số trường) |
| 4.3.2 | Dataset | DCAT | Tự biên soạn, tiếp thu một phần (số trường) |
| 4.3.3 | Distribution | DCAT | Tự biên soạn, tiếp thu một phần (số trường) |
| 4.3.4 | Agent | DCAT | Tự biên soạn, tiếp thu một phần (số trường) |
| Phụ lục A | Lược đồ dữ liệu mô tả tập dữ liệu mở được chia sẻ |  | Tự biên soạn |
| Phụ lục B | Ví dụ đặc tả tập dữ liệu mở bằng ngôn ngữ XML |  | Tự biên soạn |
| Phụ lục C | Ví dụ đặc tả tập dữ liệu mở bằng ngôn ngữ JSON |  | Tự biên soạn |
| Phụ lục D | Bảng mã tần suất cập nhật | ISO 8601 | Chấp nhận nội dung chính |

# Khuyến nghị áp dụng tiêu chuẩn

Tiêu chuẩn này quy định cấu trúc dữ liệu đặc tả tập dữ liệu mở được công bố trên cổng dữ liệu cung cấp dữ liệu mở của cơ quan nhà nước, tổ chức, cá nhân để phục vụ mục đích công bố tập dữ liệu mở, kết nối các cổng dữ liệu cung cấp dữ liệu mở với nhau để trao đổi thông tin về tập dữ liệu mở. Do vậy, tiêu chuẩn này khuyến nghị áp dụng trong các hoạt động sau:

- Xây dựng cổng dữ liệu cung cấp dữ liệu mở;

- Xây dựng và công bố tập dữ liệu mở;

- Kết nối, chia sẻ thông tin về dữ liệu mở giữa các cổng dữ liệu.

- Tra cứu, khai thác các tập dữ liệu mở được công bố.

# Kết luận

Hiện nay, nhiều nước trên thế giới đều xác định sự cần thiết của việc xây dựng cổng dữ liệu quốc gia để cung cấp dữ liệu cho người dân, tổ chức, doanh nghiệp để thể hiện sự minh bạch kích thích sáng tạo xã hội. Chuẩn hóa dữ liệu cung cấp là một hạng mục công việc rất quan trọng để giúp các các cá nhân, tổ chức khai thác dữ liệu hiệu quả và dễ dàng.

Đặc tả/mô tả các tập dữ liệu là tiêu chuẩn hết sức cần thiết để giúp con người hay máy hiểu về dữ liệu đã được cung cấp. Ngoài ra, các cổng dữ liệu phải có sự kết nối, trao đổi tập dữ liệu với nhau thì đặc tả tập dữ liệu là yêu cầu bắt buộc.

Các nước trên thế giới đều quan tâm xây dựng các tiêu chuẩn đặc tả dữ liệu mở này, tùy theo từng nước mà mức độ quy định có sự phức tạp hay đơn giản khác nhau nhưng đều căn cứ trên tiêu chuẩn gốc DCAT.

Việt Nam đang tích cực cung cấp dữ liệu mở nên việc nghiên cứu và ban hành tiêu chuẩn là hết sức cần thiết. Việc ban hành sớm sẽ giúp dễ dàng cho việc triển khai, tránh phải điều chỉnh sau này.

Tiêu chuẩn đề xuất cho Việt Nam (DCAT-VN) đã được phát triển trên tiêu chuẩn gốc, tiếp thu được các kinh nghiệm các quốc gia và phù hợp với tình hình phát triển của Việt Nam, giúp cho các cơ quan nhà nước, tổ chức dễ dàng áp dụng, thuận tiện cho các cơ quan, tổ chức cung cấp dữ liệu mở, từ đó thúc đẩy sự sáng tạo trong cộng đồng, xã hội; là mục tiêu trong chương trình chuyển đổi số của nước ta hiện nay./.

1. https://github.com/data-govt-nz [↑](#footnote-ref-1)
2. https://doc.data.gouv.fr/ [↑](#footnote-ref-2)
3. http://xmlns.com/foaf/0.1/ [↑](#footnote-ref-3)