

PHÁT TRIỂN DỮ LIỆU MỞ THEO MÔ HÌNH DỮ LIỆU RDF VÀ LINKED DATA

Hoàng Hữu Hạnh^{1*}, Phạm Vũ Minh Tú¹, Hoàng Bảo Hùng²,
Nguyễn Trọng Khánh³, Huỳnh Ngọc Mỹ Vân⁴

¹Học viện Công nghệ Bưu chính Viễn thông

²Trung tâm Công nghệ Thông tin Thừa Thiên Huế

³Cục Chuyển đổi số Quốc gia

⁴Trường Trung học cơ sở Nhựt Tảo, Long An

*Email: hhhanh@ptit.edu.vn

Ngày nhận bài: 25/5/2022; ngày hoàn thành phản biện: 5/6/2022; ngày duyệt đăng: 22/6/2022

TÓM TẮT

Dữ liệu Mở (Open data) ngày càng trở nên quan trọng trong các dịch vụ dữ liệu số nhằm tạo ra thông tin số và nội dung số dựa trên các nguồn dữ liệu mở tin cậy và tạo ra các giá trị vô cùng lớn trong phát triển kinh tế xã hội và cộng đồng-người dân. Việc sử dụng các công cụ công nghệ thông tin, truyền thông trong việc thu thập, xử lý và tạo ra các dịch vụ dữ liệu số, thông tin số và nội dung số phục vụ công tác thông tin, cung cấp dịch vụ dựa trên dữ liệu và thúc đẩy chuyển đổi số trong phát triển kinh tế - xã hội đóng vai trò thiết yếu trong cơ cấu kinh tế xã hội của mỗi quốc gia và Việt Nam không nằm ngoài xu thế đó; hơn nữa nhu cầu ngữ nghĩa, liên tác và tích hợp dữ liệu cho cổng dữ liệu quốc gia với chất lượng dữ liệu tốt hơn. Trong bài báo này, chúng tôi trình bày tiếp cận xây dựng dữ liệu mở chất lượng cao đáp ứng các tiêu chí của “Open Data” theo W3C đó là định dạng dữ liệu RDF và Linked Data cho dữ liệu mở của các cổng dữ liệu.

Từ khóa: Open Data, Data Portal, Linked Data, RDF.

1. MỞ ĐẦU

Dữ liệu là tài nguyên quốc gia quý giá và là tài sản chiến lược của Chính phủ, Quốc gia. Quản lý dữ liệu này như một tài sản và làm cho nó sẵn sàng, có thể khai phá và có thể sử dụng, một cách dễ hiểu; nó thúc đẩy hiệu quả trong hoạt động chính phủ, và còn có tiềm năng tạo ra cơ hội kinh tế và cải thiện chất lượng của người dân. Một số thuật ngữ và các khái niệm quan trọng nhất được sử dụng trong bài báo này được trình bày ở phần này để làm điểm mở đầu.

Dữ liệu Mở (Open Data) - Dữ liệu Mở (Open Data) được định nghĩa là dữ liệu số với đặc tính sẵn có và quyền truy cập mở; có thể tái sử dụng và phân phối lại; có thể truy cập quy mô toàn cầu với các giấy phép dữ liệu cụ thể [1].

Định nghĩa 'Mở'¹ đặt ra 11 yêu cầu mà Dữ liệu mở phải tuân theo. Các yêu cầu sau về cơ bản chỉ ra cách cho phép sử dụng miễn phí, tái sử dụng và phân phối lại dữ liệu. Hơn nữa, dữ liệu mở không được phân biệt đối xử với bất kỳ người nào và không được hạn chế việc sử dụng dữ liệu cho một lĩnh vực hoặc liên ngành cụ thể. Do đó, dữ liệu được xuất bản ở định dạng dữ liệu mở sẽ là "độc lập với nền tảng, có thể đọc được bằng máy và được cung cấp cho công chúng mà không có hạn chế cản trở việc sử dụng lại thông tin đó"². Do vậy, dữ liệu mở chỉ đề cập đến dữ liệu được cung cấp miễn phí cho công chúng mà không có bất kỳ giới hạn nào. Dữ liệu mở được coi là yếu tố quan trọng của Chính phủ mở.

Dữ liệu công khai (Public Data) - Điều quan trọng cần lưu ý là sự phân biệt giữa dữ liệu công khai và dữ liệu mở. Mặc dù dữ liệu công khai được cung cấp miễn phí cho công chúng, nhưng nó không nhất thiết phải mở. Một ví dụ điển hình về dữ liệu công khai không được mở là kho lưu trữ các tài liệu pháp lý. Trong khi chúng có thể truy cập tự do, hãy tưởng tượng nỗ lực cần thiết để xác định và định vị một tài liệu cụ thể. Mặt khác, nếu dữ liệu đó được số hóa và cung cấp trực tuyến ở định dạng chuẩn hóa (cũng được lập chỉ mục), thì dữ liệu công khai này cũng được mở.

Dữ liệu Chính phủ Mở (Open Government Data) - Dữ liệu Chính phủ Mở là một tập hợp con của Dữ liệu Mở, và chỉ đơn giản là dữ liệu liên quan đến chính phủ được công khai. Dữ liệu của chính phủ có thể chứa nhiều bộ dữ liệu, bao gồm ngân sách và chi tiêu, dân số, điều tra dân số, địa lý, biên bản quốc hội, v.v. Nó cũng bao gồm dữ liệu thuộc 'sở hữu' gián tiếp của các cơ quan hành chính nhà nước (ví dụ: thông qua các cục hoặc công ty nhà nước), chẳng hạn như dữ liệu liên quan đến khí hậu / ô nhiễm, giao thông công cộng, tắc nghẽn / giao thông, chăm sóc / giáo dục trẻ em.

Nghị định 47/2020/NĐ-CP quy định, **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".

Dữ liệu Liên kết (Linked Data) - Liên kết dữ liệu là quá trình tuân theo một tập hợp các phương pháp hay nhất để xuất bản và kết nối dữ liệu có cấu trúc trên Web. Đây là bước cuối cùng trong sơ đồ triển khai năm sao³ cho dữ liệu mở. Do đó, thuật ngữ 'dữ

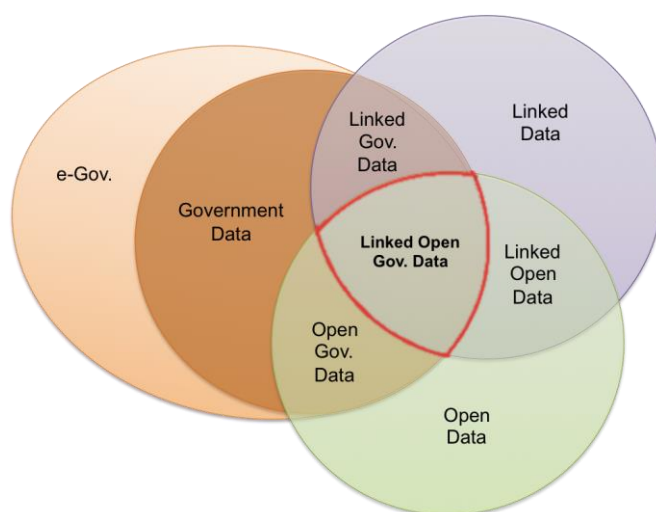
¹ <http://opendefinition.org/od/>

² <http://www.whitehouse.gov/open/documents/open-government-directive>

³ <http://5stardata.info/>

liệu liên kết' (linked data) dùng để chỉ dữ liệu được xuất bản trên Web và ngoài việc máy có thể đọc được, nó còn được liên kết với các bộ dữ liệu bên ngoài khác. Tỷ lệ áp dụng các phương pháp hay nhất về dữ liệu được liên kết ngày càng tăng đã khiến Web phát triển thành một không gian thông tin toàn cầu chứa hàng tỷ xác nhận, nơi cả tài liệu và dữ liệu đều được liên kết với nhau. Sự phát triển của Web cho phép khám phá các mối quan hệ mới giữa dữ liệu và sự phát triển tiếp theo của các ứng dụng mới.

Hình 1 thể hiện một cách trực quan các mối quan hệ giữa Dữ liệu mở, Dữ liệu chính phủ và Dữ liệu liên kết.



Hình 1. Mối quan hệ giữa các loại Dữ liệu mở

Lợi ích dữ liệu mở là thúc đẩy sự phát triển của xã hội dựa trên dữ liệu; cung cấp các dịch vụ thông tin dựa trên dữ liệu mở hướng đến một xã hội số minh bạch dựa trên dữ liệu cho cộng đồng, đồng thời tăng hiệu quả hoạt động cho các cơ quan, tổ chức, doanh nghiệp.

Với các nội dung như trên, cấu trúc bài báo được chia thành các phần như sau: mô hình dữ liệu RDF được giới thiệu ở phần 2. Phần 3 đề xuất hệ thống chuyển đổi và xử lý dữ liệu MXH nói chung. Và cuối cùng, phát triển hệ thống và kết quả thực nghiệm sẽ được trình bày ở Phần 4.

2. MÔ HÌNH DỮ LIỆU RDF

2.1. Mô hình dữ liệu RDF

Mô hình cơ bản của RDF gồm ba đối tượng sau:

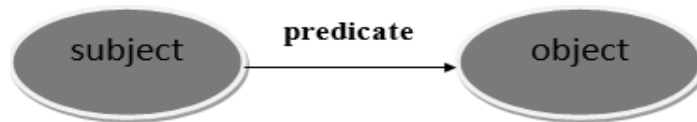
- *Tài nguyên (Resources)*: là tất cả những gì được mô tả bằng biểu thức RDF. Nó có thể là một trang Web, một phần của trang web hay một tập các trang Web.

- *Thuộc tính (Properties)*: thuộc tính, đặc tính, hoặc quan hệ dùng để mô tả tính chất của tài nguyên.
- *Giá trị (Value)*: là giá trị được gán cho thuộc tính.

Mô hình RDF mô tả các tài nguyên thông qua các phát biểu (Statements), mỗi phát biểu là một bộ ba $\langle S, P, O \rangle$, gồm ba thành phần sau :

- *S (Subject - Chủ thể)*: địa chỉ hay vị trí tài nguyên muốn mô tả, được xác định bởi URI.
- *P (Predicate - Vị từ/thuộc tính)*: xác định tính chất của tài nguyên, cũng được xác định bởi URI.
- *O (Object - Đối tượng)*: nội dung gán cho thuộc tính, có thể là tài nguyên hoặc giá trị literal.

Đồ thị RDF: Tập các *triple* hợp lại tạo thành đồ thị RDF (RDF Graph). Các node trong đồ thị có thể là các *subject* và *object* trong *triple* và các cung trong đồ thị là các *predicate*.



Hình 2. Mối quan hệ giữa các thành phần trong bộ ba RDF

2.2. Truy vấn dữ liệu RDF với SPARQL

SPARQL – Simple Protocol And RDF Query Language là ngôn ngữ truy vấn và giao thức trao đổi dữ liệu RDF của W3C. SPARQL được phát triển dựa trên các ngôn ngữ RDF trước đây là rdfDB, RDQL và SeRQL, và bổ sung một số đặc tính mới [4]. SPARQL có thể được sử dụng để tạo các truy vấn trên các nguồn dữ liệu đa dạng, cho dù dữ liệu được lưu giữ một cách đơn thuần như RDF hay được xem như là RDF thông qua phần mềm trung gian.

SPARQL có thể được sử dụng để tạo các truy vấn trên các nguồn dữ liệu đa dạng, cho dù dữ liệu được lưu giữ một cách đơn thuần như RDF hay được xem như là RDF thông qua phần mềm trung gian. SPARQL được phát triển bởi nhóm RDF Data Access Working Group-một phần trong hoạt động của Semantic Web và đã được W3C khuyến nghị sử dụng vào ngày 15 tháng 01 năm 2008. Ngôn ngữ truy vấn SPARQL lấy thông tin từ các đồ thị RDF , nó cung cấp các tính năng sau:

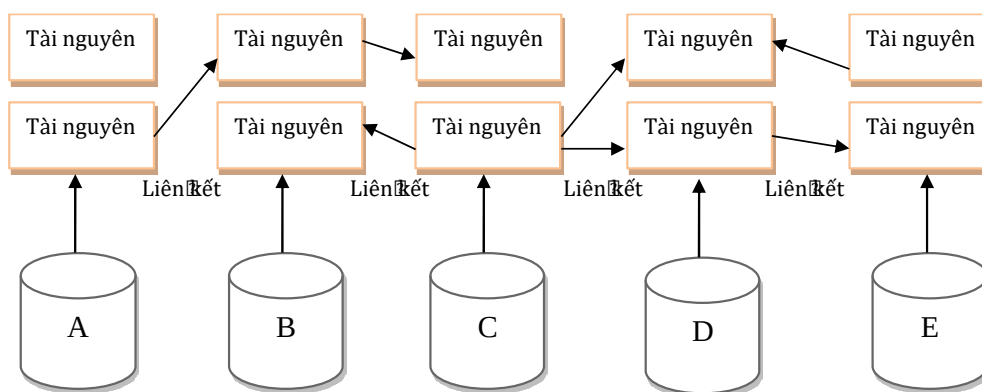
- Chiết xuất thông tin dưới dạng các URI, các node trắng (*blank node*) ,các *plain literal* và *typed literal*.
- Chiết xuất các đồ thị con RDF.

- Xây dựng các đồ thị RDF mới dựa trên thông tin của các đồ thị truy vấn.

Kết quả của câu truy vấn là tất cả các giải pháp mà một câu truy vấn có thể phù hợp với đồ thị được truy vấn. Một câu truy vấn có thể có không, một hay nhiều kết quả.

2.3. Linked Data

Thuật ngữ Linked Data được Tim Berner-Lee đưa ra trong các ghi chép về kiến trúc “Linked Data Web” của mình [6]. Thuật ngữ này chỉ cách thức để xuất bản và liên kết các dữ liệu có cấu trúc trên Web. Giả thuyết cơ bản của Linked Data là lợi ích và giá trị của dữ liệu tăng lên khi nó được liên kết với các dữ liệu khác. Nói cách khác, Linked Data đơn giản là sử dụng Web để tạo ra các liên kết định kiểu (typed link) giữa các dữ liệu từ nhiều nguồn tài nguyên khác nhau. Điều này giống như hai CSDL của hai tổ chức ở các vùng địa lý khác nhau hay đơn giản là một hệ thống hỗn tạp của cùng một tổ chức không dễ dàng để trao đổi, liên thông ở mức dữ liệu.



Hình 3. Mô hình liên kết dữ liệu trong Web ngữ nghĩa

Tim Berner-Lee đã đưa ra bốn quy tắc, hay còn gọi là bốn công thức cốt lõi và cơ bản nhất của Linked Data như sau [6]:

1. Dùng URI để đặt tên cho mọi thứ, không chỉ là các tài liệu
2. Sử dụng HTTP URI để tìm kiếm tên tài nguyên
3. Khi tìm kiếm phải cung cấp thêm các thông tin hữu ích, sử dụng các kỹ thuật đã được chuẩn hóa (RDF, SPARQL)
4. Có những liên kết đến URI khác cho phép khám phá thêm những tài nguyên (tri thức) khác

Quy tắc đầu tiên dùng định danh tài nguyên toàn cầu URI để đặt tên cho mọi thứ như con người, nơi chốn, tài liệu,... Ví dụ như <http://www.hueuni.edu.vn/hhhanh> là URI được đặt cho Hoàng Hữu Hạnh hay <http://www.w3.org/People/Berners-Lee/card> là URI cho Tim Berner Lee. Có thể có nhiều URI cùng được đặt tên cho cùng một thứ trong thế giới thực.

Ví dụ: DBpedia⁴ dùng URI là <http://dbpedia.org/resource/Berlin> để định danh cho địa danh Berlin trong khi Geonames⁵ lại sử dụng URI là <http://sws.geonames.org/2950159/>.

Điều này hoàn toàn tự nhiên như trong thế giới thực, một sự vật có thể được đặt bởi nhiều tên khác nhau cũng như được thể hiện ở nhiều ngôn ngữ khác nhau.

Từ việc định danh cho các tài nguyên như trên, quy tắc thứ hai xác định cách thức tìm kiếm các tài nguyên đó bằng cách sử dụng HTTP URI. HTTP là giao thức nền tảng và phổ biến trên Web dùng để tham chiếu đến các địa chỉ URI trên Web.

Quy tắc thứ ba nói về các thông tin liên quan đến tài nguyên tìm kiếm cũng được cung cấp cho người dùng khi duyệt web. Chẳng hạn như khi tìm kiếm thông tin về một cá nhân thì các thông tin khác liên quan đến cá nhân đó cũng được cung cấp hỗ trợ người dùng như ngoài thông tin về họ tên, giới tính, ngày sinh thì có thể thêm các thông tin về các bài báo hay công trình đã thực hiện, sở thích cá nhân,...

Quy tắc cuối cùng mô tả việc kết nối các dữ liệu trong phạm vi toàn cầu sử dụng các liên kết định kiểu và đó chính là các liên kết RDF. Người dùng có thể duyệt các thông tin từ các nguồn tài nguyên khác thông qua các liên kết dạng móc xích. Từ đó, người dùng có thể khám phá thêm nhiều tài nguyên hữu ích khác trên Web.

3. TỪ OPEN DATA ĐẾN LINKED OPEN DATA

Phần này giới thiệu các nguyên tắc liên kết dữ liệu; định nghĩa các thuật ngữ quan trọng như Chính phủ mở, Dữ liệu mở (Chính phủ) – Open (Government) Data và Dữ liệu Liên kết mở (Chính phủ) – Linked Open (Government) Data; và giải thích các cơ chế liên quan để đảm bảo nền tảng vững chắc trước khi đi sâu hơn.

3.1. Open Government và Open (Government) Data

Open Government Data (OGD) – *Dữ liệu Chính phủ Mở* -- là một nỗ lực trên toàn thế giới nhằm mở dữ liệu, thông tin và nội dung của chính phủ sang các định dạng không độc quyền có thể đọc được của con người và máy móc để xã hội dân sự, kinh tế, truyền thông và học thuật cũng như các chính trị gia và quản trị viên công cộng sử dụng lại. Điều này sẽ chỉ áp dụng cho dữ liệu và thông tin được sản xuất hoặc ủy quyền bởi chính phủ hoặc các tổ chức do chính phủ kiểm soát và không liên quan đến dữ liệu về các cá nhân.

Đáp ứng với các quy tắc này về Dữ liệu của Chính phủ sẽ được coi là "mở" nếu dữ liệu được công khai theo cách tuân thủ các nguyên tắc dưới đây quy định bởi Nghị

⁴ dbpedia.org

⁵ geonames.org

định 47/2020/NĐ-CP (Khoản 1, Điều 17, Mục 3 - **Dữ liệu mở của Cơ quan Nhà nước**) nêu rõ các nguyên tắc cung cấp dữ liệu mở của cơ quan nhà nước.

Các quốc gia hàng đầu hiện nay trong các hoạt động và sáng kiến Dữ liệu mở quốc gia là chính phủ Liên minh Châu Âu, Hoa Kỳ, Úc, các quốc gia Scandinavia và chính phủ Vương quốc Anh. Tất cả các quốc gia này đều có cam kết chính trị cao đối với cả Dữ liệu mở và các cổng Dữ liệu mở trung tâm, và tất cả đều có một cộng đồng Dữ liệu mở mạnh mẽ. Những quốc gia đổi mới này và những người đứng sau họ có thể được coi là những người tiên phong của OGD. Việt Nam một chương trình Chuyển đổi số đầy tham vọng, cũng đặt ra các mục tiêu về phát triển dữ liệu số và dữ liệu mở của chính phủ trong những năm gần đây, đặc biệt là sau khi Chương trình Chuyển đổi số Quốc gia được Thủ tướng CP ban hành năm 2020.

3.2. Từ Open Data đến Linked Open Data

Như đã đề cập ở trên, OGD là tất cả về việc mở thông tin và dữ liệu, cũng như làm cho nó có thể sử dụng và tái sử dụng nó. Một phân tích các yêu cầu của OGD đã được thực hiện vào tháng 6 năm 2011 tại Áo và nêu rõ mười một lĩnh vực sau đây cần xem xét khi nghĩ về OGD:

1. Cần định nghĩa
2. Chính phủ mở: minh bạch, dân chủ, tham gia và cộng tác
3. Các vấn đề pháp lý
4. Tác động đến xã hội
5. Đổi mới và xã hội tri thức
6. Ảnh hưởng đến kinh tế và công nghiệp
7. Giấy phép, mô hình khai thác, điều khoản sử dụng
8. Các khía cạnh liên quan đến dữ liệu
9. Quản trị dữ liệu
10. Các ứng dụng và trường hợp sử dụng
11. Các khía cạnh công nghệ

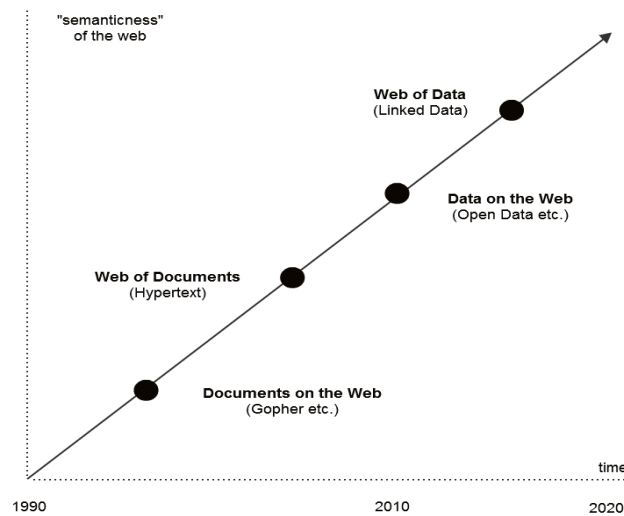
Khi xem xét cách thức để hưởng lợi đầy đủ từ OGD trong các trường hợp cụ thể, rõ ràng là khả năng tương tác và các tiêu chuẩn là chìa khóa. Đây là lúc các nguyên tắc LOD phát huy tác dụng.

Để được hưởng lợi đầy đủ từ Dữ liệu mở, điều quan trọng là phải đặt thông tin và dữ liệu vào bối cảnh tạo ra kiến thức mới và kích hoạt các dịch vụ và ứng dụng mạnh

mẽ. Vì LOD tạo điều kiện cho sự đổi mới và sáng tạo tri thức từ dữ liệu được liên kết với nhau, nên nó là một cơ chế quan trọng để quản lý và tích hợp thông tin.

3.3. Điểm mạnh của Linked Open Data

Ngày nay, ý tưởng liên kết các trang web bằng cách sử dụng siêu liên kết là hiển nhiên, nhưng đây là một khái niệm đột phá cách đây 20 năm. Ngày nay chúng ta cũng đang ở trong tình huống tương tự vì nhiều tổ chức không hiểu ý tưởng xuất bản dữ liệu trên web, chưa nói đến lý do tại sao dữ liệu trên web phải được liên kết. Sự phát triển của Web có thể được nhìn thấy như sau:

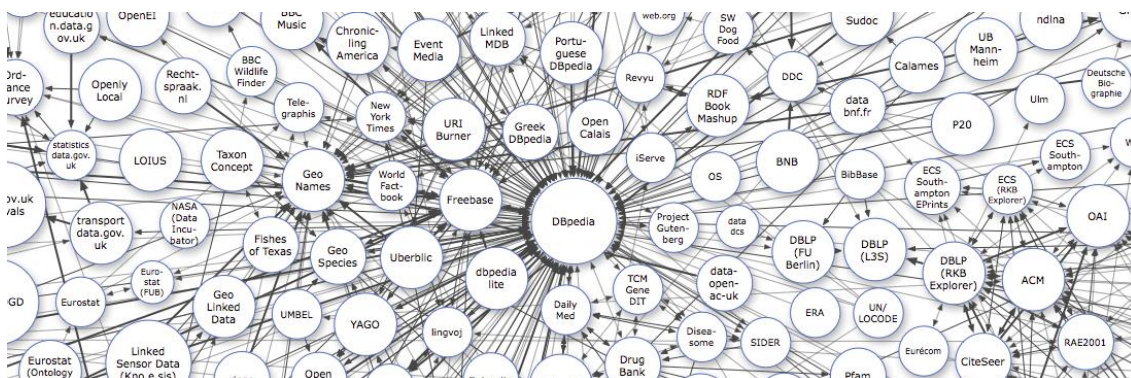


Hình 4. Sự phát triển của tài liệu và dữ liệu Web

Mặc dù ý tưởng về Dữ liệu mở Liên kết (LOD) vẫn chưa được công nhận là chủ đạo (như Web mà chúng ta biết ngày nay), nhưng đã có rất nhiều LOD đã có sẵn. Cái gọi là ‘LOD cloud’⁶ bao gồm hơn 50 tỷ dữ liệu ước tính từ nhiều lĩnh vực khác nhau như địa lý, truyền thông, sinh học, hóa học, kinh tế, năng lượng, v.v. Dữ liệu có chất lượng khác nhau và hầu hết dữ liệu cũng có thể được sử dụng lại cho mục đích thương mại.

Hầu hết các hệ thống ngày nay đều xử lý lượng thông tin khổng lồ. Tất cả thông tin được tạo ra hoặc trong ranh giới hệ thống (và một phần được công bố cho các hệ thống khác) hoặc nó được tiêu thụ “từ bên ngoài”, “trộn lẫn” (mash) và “tiêu hóa” (digest) trong ranh giới. Sự phức tạp ngày càng tăng có nguyên nhân từ trình độ học vấn cao hơn và những cải tiến kỹ thuật do lĩnh vực CNTT-TT thực hiện trong 30 năm qua. Nói một cách đơn giản, nhân loại hiện có thể xử lý nhiều thông tin hơn bao giờ hết với chi phí có lẽ là thấp nhất từ trước đến nay (hãy nghĩ đến băng thông cao hơn và chi phí lưu trữ dữ liệu thấp hơn).

⁶ <https://lod-cloud.net>



Với trọng tâm rõ ràng là quản lý siêu dữ liệu chất lượng cao, Dữ liệu được liên kết là chìa khóa để khắc phục vấn đề này. Giá trị của dữ liệu tăng lên mỗi khi nó được sử dụng lại và được liên kết với một tài nguyên khác. Việc sử dụng lại chỉ có thể được kích hoạt bằng cách cung cấp thông tin về thông tin có sẵn. Để thực hiện nhiệm vụ này một cách bền vững, thông tin phải được công nhận là một nguồn tài nguyên quan trọng cần được quản lý giống như bất kỳ nguồn nào khác.

Dữ liệu chính phủ mở được liên kết (LOGD), được tiên phong bởi Data.gov và Data.gov.uk, đang nổi lên trên Web dữ liệu liên kết như một cách tạo điều kiện thuận lợi cho việc mở, liên kết và tái sử dụng OGD. Dữ liệu được liên kết mang lại sự đồng thuận tối thiểu về biểu diễn dữ liệu (sử dụng URI và Khung mô tả tài nguyên) và truy cập dữ liệu (qua HTTP) và nó cho phép xuất bản OGD gia tăng theo “5 sao của dữ liệu mở được liên kết” của Tim Berners-Lee hay W3C. LOGD được công nhận là một hệ sinh thái mở dựa trên Web kết nối hữu cơ với nhau giữa chủ sở hữu dữ liệu gốc (chẳng hạn như cơ quan chính phủ), nhà cung cấp dịch vụ xử lý dữ liệu (chẳng hạn như dịch vụ phân giải thực thể) và người tiêu dùng dữ liệu (doanh nghiệp và công dân).

67

Nam tới các mã tiêu chuẩn xử lý thông tin liên bang tương ứng hoặc một dịch vụ Web tìm các thực thể DBpedia có liên quan cho một tên.

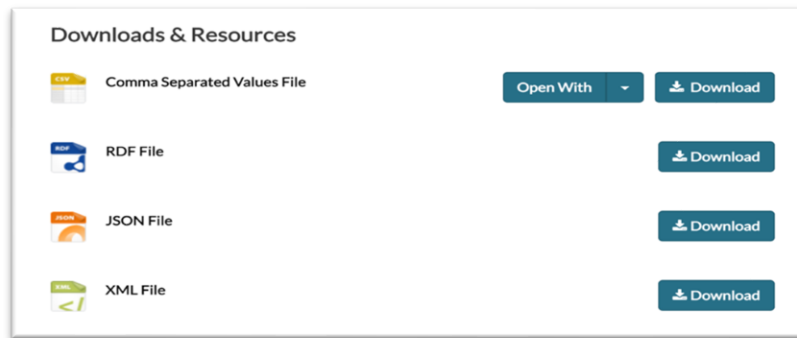
Mặc dù LOGD có lẽ là phần phát triển nhanh của Web dữ liệu được liên kết, nhưng hầu hết các kết quả nghiên cứu đều thừa nhận một rào cản trong việc xuất bản LOGD. Công cụ phần mềm nguồn mở đã được phát triển và sử dụng lại để tạo điều kiện thuận lợi cho việc biên mục và tạo bộ dữ liệu LOGD. Đặc biệt, các cổng dữ liệu như Mạng Lưu trữ Tri thức Toàn diện (Comprehensive Knowledge Archive Network - <http://ckan.org>) ở Vương quốc Anh và sự hợp tác Nền tảng Chính phủ Mở Hoa Kỳ-Ấn Độ (www.data.gov/opengovplatform) giúp phát hành nhiều bộ dữ liệu OGD hơn.

4. CÁC CỔNG DỮ LIỆU QUỐC GIA

4.1. Các quốc gia tiên tiến về Open Government Data

- Cổng dữ liệu mở cấp quốc gia như “data.gov” và “data.gov.uk”:

Data.gov.uk là Cổng Dữ liệu Mở đầu tiên ở Châu Âu và hiện nay nó đã được mở rộng lên hơn 20.000 bộ dữ liệu khác nhau bao gồm các lĩnh vực công cộng khác nhau ở Vương quốc Anh. Mặc dù chế độ và Danh mục dữ liệu khác nhau, các nước EU lớn đã công bố dữ liệu của họ thông qua Danh mục dữ liệu cấp quốc gia. Các cổng đó thường dựa trên “CKAN”, nhưng các chức năng của cổng được tùy chỉnh theo sở thích và yêu cầu khác nhau của họ.



Hình 6. Tính đa dạng trong tiêu chuẩn dữ liệu mở cung cấp (Nguồn: data.gov)

Cổng dữ liệu mở của Mỹ, công bố khá sớm, đa dạng và tiêu chí khá chặt chẽ: <https://www.data.gov/>. Được đánh giá là một cổng dữ liệu mở tốt nhất, với các dữ liệu mở công bố thường xuyên, cập nhật và tiêu chuẩn dữ liệu cao.

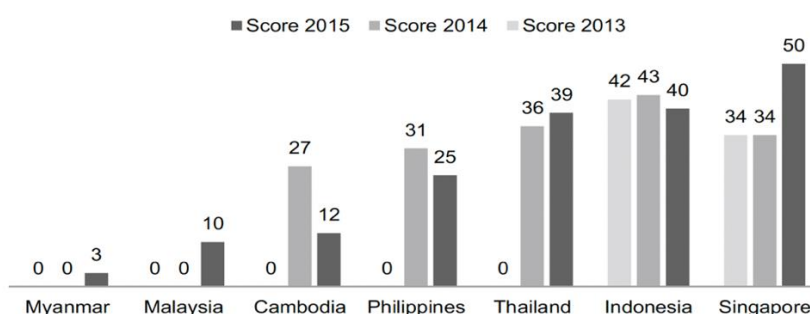
- Cổng dữ liệu mở Châu Âu

Cổng Dữ liệu Châu Âu (EDP) là điểm truy cập trung tâm cho siêu dữ liệu của Dữ liệu mở do các cơ quan công quyền ở Châu Âu xuất bản và thu thập dữ liệu từ hơn 70 nhà cung cấp dữ liệu quốc gia. Nền tảng này là điểm khởi đầu trong việc áp dụng đặc

tả Dữ liệu liên kết DCAT-AP, nhằm mục đích tăng khả năng tương tác và khả năng truy cập của Dữ liệu mở. Thành phần cốt lõi dựa trên CKAN, được mở rộng nhờ hỗ trợ sao chép Dữ liệu được Liên kết gốc đến một bộ ba để đảm bảo khả năng tương thích kế thừa và hỗ trợ DCAT-AP. EDP đang trên đà trở thành nền tảng cốt lõi cho Dữ liệu mở của Châu Âu và thúc đẩy việc chấp nhận DCAT-AP. Cổng thông tin Liên Châu Âu được phát triển bởi "dự án LOD2" và cổng thông tin của họ xuất bản các tập dữ liệu thông qua cách tiếp cận Dữ liệu được liên kết và tất cả siêu dữ liệu của tập dữ liệu có thể được truy cập thông qua các điểm cuối "SPARQL".

- **Cổng dữ liệu mở một số nước ASEAN**

Việc so sánh độ mở của dữ liệu ở các quốc gia khác nhau là rất khó và có thể dễ dẫn đến kết luận sai. Do đó, phân tích so sánh này không có ý định thông qua đánh giá về các sáng kiến dữ liệu mở trong ASEAN, mà có mục tiêu định lượng các yếu tố của các sáng kiến dữ liệu mở ở các nước thành viên ASEAN, cung cấp vốn từ vựng chung trong các cuộc thảo luận về chủ đề này.



Hình 7. Chỉ số ODI của các quốc gia ASEAN được thống kê 2016

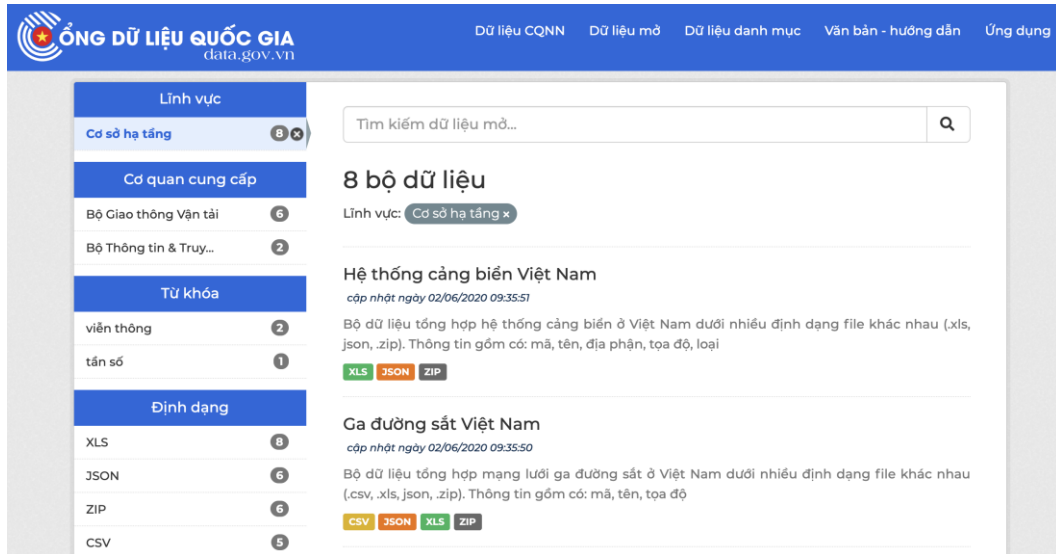
Kể từ năm 2013, Quỹ tri thức mở tổ chức một cuộc khảo sát hàng năm để tạo ra ODI (Open Data Index) trong nỗ lực cung cấp một cuộc kiểm toán xã hội dân sự về cách các chính phủ trên thế giới công bố dữ liệu chính thức. ODI là một đánh giá độc lập từ góc độ công dân và nó cung cấp tài khoản về cách dữ liệu mở có thể truy cập được xuất hiện đối với người dùng dữ liệu. Chỉ số thống kê 2016 bao gồm 149 quốc gia, bao gồm hầu hết các quốc gia trong ASEAN, ngoại trừ Brunei, Lào và Việt Nam (chưa có số liệu theo thống kê). ODI tổng hợp không chỉ xem xét dữ liệu mở từ chính phủ mà còn bao gồm các sáng kiến tư nhân - công bố dữ liệu từ các nguồn chính thức - ở Campuchia và Myanmar, nơi không có cổng dữ liệu mở chính thức nào tồn tại.

Người dùng thường gọi chỉ số này như một công cụ đo điểm chuẩn cho các sáng kiến của các chính phủ trong việc mở dữ liệu của họ, nhưng những diễn giải như vậy cũng phải tính đến hệ sinh thái lớn hơn xung quanh dữ liệu mở để tránh kết luận sai lầm. ODI cũng đóng một vai trò trong việc xây dựng một cộng đồng xung quanh dữ liệu mở chính phủ trên toàn cầu. Chi tiết có ở trang web ODI: <https://index.okfn.org/place/>

4.2. Các cổng dữ liệu trong nước

- **Cổng dữ liệu quốc gia – data.gov.vn**

Cổng dữ liệu quốc gia có rất nhiều thông tin về CSDL Quốc gia ở mức độ khái quát nhưng vẫn thiếu chi tiết cụ thể. Các dịch vụ về dữ liệu mới chỉ có tiêu mục, nhiều nơi không có dữ liệu, hoặc nếu có dữ liệu thì thường là định dạng Excel như vậy có thể gây khó cho người dùng bình thường khi muốn khai thác dữ liệu.



Hình 7. Các bộ dữ liệu và định dạng (openness) được công bố trên Cổng dữ liệu quốc gia

Việc công bố dữ liệu mở thể các bộ dữ liệu với các tiêu chuẩn hoá và được cập nhật sẽ tạo ra một kho dữ liệu mở có chất lượng trong việc thúc đẩy tính minh bạch, cung cấp các kênh thông tin đến người dân, doanh nghiệp và các cộng đồng khởi nghiệp, vốn muốn có các nguồn dữ liệu mở cho phát triển các dịch vụ dựa trên dữ liệu công khai.

- **Một số cổng dữ liệu công khai và mở ở một số địa Phương**

- *Thành phố Hồ Chí Minh - opendata.hochiminhcity.gov.vn*

Kho dữ liệu dùng chung của thành phố (TP) xây dựng dựa trên các nền tảng là cơ sở dữ liệu người dân, cơ sở dữ liệu doanh nghiệp, cơ sở dữ liệu bản đồ số và cơ sở dữ liệu danh mục dùng chung trên cơ sở tích hợp dữ liệu, chia sẻ, dùng chung các cơ sở dữ liệu tại các sở, ban, ngành, quận, huyện.

Việc tích hợp, kết nối, chia sẻ, sử dụng, khai thác Kho dữ liệu dùng chung đều được thực hiện thông qua nền tảng tích hợp và chia sẻ dữ liệu của TP với cơ chế định danh, xác thực, phân quyền người sử dụng.

Việc truy cập, khai thác Kho dữ liệu dùng chung của TP sẽ được thực hiện qua các dịch vụ truy cập trực tuyến, di động, các giao diện lập trình ứng dụng, các phương

thức, giải pháp kỹ thuật khai thác dữ liệu khác. Một phần của Kho dữ liệu dùng chung của TP sẽ được chia sẻ cho người dân, tổ chức phi chính phủ, doanh nghiệp thông qua các dịch vụ dữ liệu mở để khai thác sử dụng để tạo giá trị gia tăng, góp phần phát triển kinh tế - xã hội cho TP.

- *Thành phố Đà Nẵng - congdlieu.vn*

Sau khi xây dựng khung định hướng phát triển Đà Nẵng theo hướng “đô thị thông minh”, Đà Nẵng sớm bắt tay vào xây dựng hệ thống tài nguyên phục vụ chiến lược này và hình thành kho dữ liệu số mở trực tuyến là trụ cột kiến trúc cho mô hình TP thông minh được Đà Nẵng đặt ra.

Sau hơn một năm xây dựng, tháng 10/2019, Sở TT&TT TP Đà Nẵng đã đưa Cổng dịch vụ dữ liệu mở Đà Nẵng chính thức đưa vào vận hành, cung cấp các thông tin nhà nước, chính quyền được phép trên địa bàn TP Đà Nẵng. Và sau khi đưa vào sử dụng, Cổng dữ liệu được giao cho Trung tâm Thông tin dịch vụ công Đà Nẵng vận hành kỹ thuật, thu thập, xử lý, cập nhật dữ liệu và cung cấp đến cộng đồng.

- *Tỉnh Thừa Thiên Huế - data.thuathienhue.gov.vn*

Đối với dữ liệu mở của chính quyền địa phương (CQDP), Thừa Thiên Huế cũng là một trong những đơn vị tiên phong trong việc xây dựng và công bố dữ liệu của chính quyền tỉnh. Năm 2018, UBND tỉnh Thừa Thiên Huế đã xây dựng và cho công bố Hệ thống dữ liệu mở của tỉnh trên cổng DLM tại: <https://data.thuathienhue.gov.vn>. Hệ thống dữ liệu mở của tỉnh được xây dựng và vận hành trước thời điểm Chính phủ ban hành Nghị định số 47/2020/NĐ-CP ngày 09/4/2020 của Chính phủ về quản lý, kết nối và chia sẻ dữ liệu của cơ quan nhà nước.

Việc phát triển các cổng dữ liệu mở là một xu hướng tốt trong việc minh bạch hoá hoạt động công quyền, và thúc đẩy đổi mới sáng tạo xã hội với các dữ liệu và thông tin được cung cấp từ các nguồn tin cậy, mở và định dạng máy tính có thể xử lý. Việc tuân theo các tiêu chuẩn DLM là một điều quan trọng, và tổ chức dữ liệu không làm mất mát ý nghĩa của thông tin cũng như tăng tính kết nối các datasets của các hệ thống dữ liệu mở là một điều rất quan trọng trong lộ trình kỹ thuật phát triển dữ liệu mở của các cơ quan nhà nước.

4.3. Một số khuyến nghị cho các Cổng dữ liệu tại Việt Nam

Đánh giá một cách tổng quan, việc các tỉnh thành và cấp độ chính phủ có các bước phát triển theo định hướng dữ liệu và công bố dữ liệu mở là một bước đi đúng hướng và quan trọng trong xây dựng Chính phủ số. Tuy nhiên, hiện chưa có một hướng dẫn cụ thể dưới Nghị định trong việc quy định các tiêu chí, tiêu chuẩn về việc xuất bản các dữ liệu mở này, cũng như việc đánh giá sự trưởng thành của dữ liệu mở ở các cấp độ khác nhau.

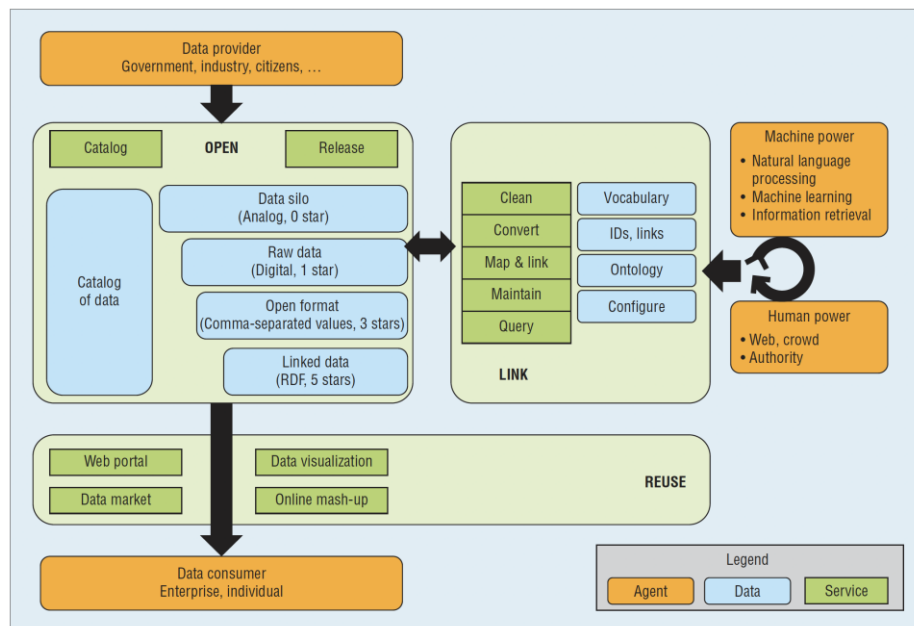
- Xuất bản theo các bộ dữ liệu, không công bố cấu trúc dữ liệu để đảm bảo an toàn thông tin;
- Cung cấp các bộ dữ liệu theo các tiêu chuẩn dữ liệu tốt hơn (bộ tiêu chuẩn 5* với định hướng LOD);
- Định nghĩa danh mục theo tiêu chuẩn DCAT;
- Quản lý tốt các metadata;
- Cung cấp các API lấy dữ liệu và các ứng dụng mẫu cho việc sử dụng các tập dữ liệu mở.

5. TRIỂN KHAI XUẤT BẢN DỮ LIỆU MỞ LIÊN KẾT

5.1. Hệ thống Cổng dữ liệu mở liên kết (Linked Open Data Portol - LODP)

Cấu trúc của một Cổng dữ liệu mở liên kết (Linked Open Data Portal - LODP) tổng quan bao gồm các thành phần chính, các bên liên quan tiêu thụ dữ liệu và các nhóm đối tượng dữ liệu được nhóm thành 3 nhóm OPEN, LINK, RESUE như sau:

- Các bên liên quan (Agent): Data provider, Data Consumer và các thành phần liên quan khác về học máy, và tính toán
- Các hình thức dữ liệu (Data): bao gồm các thành phần liên quan đến dữ liệu, các hình thức dữ liệu khác nhau trong cổng dữ liệu
- Dịch vụ cổng dữ liệu (Serivce): các dịch vụ của Cổng dữ liệu.



Hình 8. Cổng dữ liệu chính phủ mở liên kết (LOGD) cho Kiến thức chính phủ mở.

Do đó phát triển hệ thống phân tích dữ liệu mạng xã hội nhằm cung cấp một giải pháp tổng thể về thu thập và phân tích từ nhiều nguồn dữ liệu khác nhau như Mạng xã hội, báo chí,... Cho phép theo dõi các nhân vật chủ đề sự kiện cần lưu ý, tự động cảnh báo khi phát hiện vấn đề hỗ trợ phân tích cảnh báo sớm khủng hoảng truyền thông.

5.2. Xuất bản dữ liệu LOD lên LODP

Để có cách tiếp cận tương tự, chúng tôi đề xuất "Thành phần cho Dữ liệu được Liên kết (Mở) chất lượng cao" của W3C Linked Data Cookbook. Các bước thiết yếu để xuất bản LOD của riêng bạn là:

1. Lập mô hình và liên kết dữ liệu
2. Đặt tên mọi thứ bằng URI
3. Sử dụng lại từ vựng bất cứ khi nào có thể
4. Xuất bản các mô tả do con người và máy móc có thể đọc được
5. Chuyển đổi dữ liệu sang RDF
6. Chỉ định một giấy phép thích hợp
7. Thông báo (các) Tập dữ liệu được liên kết mới

5.3. Các tính năng của hệ thống

Hệ thống LODP được thử nghiệm tại <https://lod.neosol.vn>, cung cấp các tính năng sau:

- Hệ thống chuyển đổi dữ liệu RDF
- Hệ thống truy vấn bằng SPARQL
- Hệ thống tích hợp API cho truy vấn
- Portal dữ liệu mở theo tiêu chuẩn W3C 5*

6. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Bài báo trình bày kết quả nghiên cứu ứng dụng trong mô hình hoá dữ liệu với thông tin ngữ nghĩa. Mô hình RDF/Linked Data đáp ứng các yêu cầu cho việc biểu diễn dữ liệu mở trên các cổng dữ liệu mở sử dụng tiêu chuẩn dữ liệu W3C 5* với định hướng mô hình dữ liệu đồ thị RDF và Linked Data. Việc triển khai các cổng dữ liệu LODP cần thực hiện một quy trình chuẩn với nhân sự được đào tạo để có thể xuất bản dữ liệu với chất lượng cao, phục vụ chuyển đổi số trong chính phủ và chính quyền số. Trong tương lai, nhóm tác giả sẽ tập trung đẩy mạnh nghiên cứu các phương pháp phân tích dữ liệu cải tiến để tận dụng lợi thế của việc biểu diễn theo ngữ nghĩa này.

TÀI LIỆU THAM KHẢO

- [1]. Paolillo, J. C., & Wright, E. (2006). Social network analysis on the semantic web: Techniques and challenges for visualizing FOAF. In *Visualizing the semantic web* (pp. 229-241).
- [2]. Resource Description Framework (RDF) (2014). Website: <https://www.w3.org/RDF/>
- [3]. SPARQL 1.1 Query Language (2013). Website: <https://www.w3.org/TR/sparql11-query/>
- [4]. M. San Martín and C. Gutierrez (2009). Representing, Querying and Transforming Social Networks with RDF/SPARQL, *Proc. ESWC*. Boston, LNCS 5554, pp. 293-307.
- [5]. S. Elbassuoni, M. Ramanath, R. Schenkel, G. Weikum (2010). Searching RDF Graphs with SPARQL and Keywords, *IEEE Data Eng. Bull.*, 33(1), pp.16-24.
- [6]. Linked Data (2019). Website: <https://www.w3.org/standards/semanticweb/data>

OPEN DATA DEVELOPMENT USING RDF DATA MODEL AND LINKED DATA

Hoang Huu Hanh^{1*}, Pham Vu Minh Tu¹, Hoang Bao Hung²,
Nguyen Trong Khanh³, Huynh Ngoc My Van

¹ Posts and Telecommunications Institute of Technology

² Hue Center for Information Technology

³ National Digital Transformation Authority

⁴ Nhut Tao Secondary School, Long An

*Email: hhhanh@ptit.edu.vn

ABSTRACT

Open data is becoming more important in digital data-driven services to generate digital information and contents based on reliable open data sources and significantly contribute to the socio-economic development and community. The use of information and communication technology tools for the collection, processing and creation of digital data information and digital content services for information work, provision of services based on data and promotion of digital transformation in socio-economic development play essential roles in the socio-economic structure of each country and Vietnam is no exception to that trend. In addition, semantics and interoperability for data integration for data portals with better data quality are currently being worked on. In this paper, we present a method for creating high-quality open data that satisfies W3C definition's "Open Data" which calls for RDF and Linked Data data formats for open data portals.

Keywords: Open Data, Data Portal, Linked Data, RDF.



Hoàng Hữu Hạnh sinh ngày 13/04/1974 tại Huế. Năm 1996, ông tốt nghiệp Cử nhân ngành Toán-Tin học tại Trường Đại học Sư phạm Huế, Thạc sĩ khoa học ngành Công nghệ thông tin tại Trường Đại học Bách khoa Hà Nội. Năm 2007, ông nhận học vị Tiến sĩ chuyên ngành Hệ thống thông tin tại Trường Đại học Công nghệ Vienna, Cộng hoà Áo. Năm 2012, ông nhận Chức danh Phó giáo sư tại Việt Nam. Hiện nay, ông là Giảng viên cao cấp tại Học viện Công nghệ Bưu chính Viễn thông (PTIT). Hiện nay là Giám đốc Trung tâm Đào tạo Quốc tế của PTIT.

Lĩnh vực nghiên cứu: Biểu diễn tri thức, Web ngữ nghĩa, Linked Data, Ontology, Công nghệ phần mềm, Công nghệ dữ liệu, Quản lý quy trình nghiệp vụ.



Phạm Vũ Minh Tú sinh ngày 30/09/1990 tại thành phố Hà Nội. Năm 2013, ông tốt nghiệp kỹ sư ngành Kỹ thuật Điện – Điện tử tại Học Viện Công Nghệ Bưu Chính Viễn Thông. Năm 2015, ông tốt nghiệp thạc sĩ chuyên ngành Kỹ thuật Viễn Thông tại Học Viện Công Nghệ Bưu Chính Viễn Thông. Từ năm 2015 đến nay, ông giảng dạy tại Học viện Công nghệ Bưu chính Viễn thông.

Lĩnh vực nghiên cứu: Truyền thông quang không dây, Xử lý tín hiệu số, Dữ liệu đa phương tiện



Hoàng Bảo Hùng sinh ngày 23/10/1971 tại Huế. Ông tốt nghiệp Đại học Sư phạm Huế năm 1993, chuyên ngành Toán học; tốt nghiệp kỹ sư CNTT tại Đại học Bách khoa Hà Nội năm 1995; tốt nghiệp Thạc sĩ ngành CNTT năm 2002 tại Đại học Bách khoa Hà Nội; bảo vệ thành công luận án Tiến sĩ CNTT năm 2007 tại Viện CNTT, Viện Hàn lâm Khoa học công nghệ Việt Nam. Hiện đang công tác tại Trung tâm Công nghệ thông tin thuộc UBND tỉnh Thừa Thiên Huế.

Lĩnh vực nghiên cứu: Cơ sở dữ liệu hướng đối tượng, GIS, khoa học dữ liệu.



Huỳnh Ngọc Mỹ Vân sinh ngày 01/7/1983 tại Thành Phố Hồ Chí Minh. Năm 2010 Bà Tốt nghiệp Cử nhân ngành Tin học tại Trường Đại học Mở Thành Phố Hồ Chí Minh. Năm 2014, bà tốt nghiệp Thạc sĩ Khoa học Máy tính tại trường Đại học Khoa học, ĐH Huế. Năm 2023 bà tốt nghiệp cử nhân ngành Ngôn ngữ Anh tại trường Đại học Kinh tế Công nghiệp Long An. Hiện nay, bà giảng dạy tại trường Trung học cơ sở Nhựt Tảo, Thành phố Tân An, tỉnh Long An.

Lĩnh vực nghiên cứu: Web ngữ nghĩa, Linked Data, Hệ thống thông tin.

