Chủ đề cuộc thi phần mềm nguồn mở - OLP'2025

Ứng dụng dữ liệu mở liên kết phục vụ chuyển đổi số

1. Tổng quan dữ liệu mở

1.1 Giấy phép dữ liệu mở

Dữ liệu là một dạng tài nguyên số và được bảo vệ sở hữu trí tuệ giống như các sản phẩm sáng tạo khác (ví dụ các tác phẩm văn học, nghệ thuật và khoa học trong đó có chương trình máy tính và bộ sưu tập dữ liệu). Bên cạnh dữ liệu, tài nguyên số còn bao gồm cả các loại tư liệu số hóa khác là tài liệu điện tử, âm thanh, hình ảnh. Việc truy cập và khai thác sử dụng các tài nguyên số phải tuân thủ luật bản quyền tác giả theo hai phương thức chính là: được tự do sử dụng miễn phí và phân phối lại dưới một giấy phép truy cập mở; hoặc đóng hoàn toàn cần có sự đồng ý cho phép sử dụng của chủ sở hữu quyền tác giả. Tài nguyên số cấp phép mở được quản lý lưu trữ trong các kho truy cập mở để có thể khai thác sử dụng trên không gian mạng. Căn cứ vào tính chất và mục đích sử dụng của nội dung, chúng ta có thể phân loại các kho lưu trữ gồm có kho dữ liệu mở, kho xuất bản truy cập mở, kho tài nguyên giáo dục mở, kho di sản số hóa mở,…

Tương tự như phần mềm nguồn mở, các tài nguyên số cũng có thể được cấp phép truy cập mở. Giấy phép mở đầu tiên cho nội dung ra đời từ năm 1998 có tên là GFDL (GNU Free Documentation License). Đây là loại giấy phép có tính mở chặt chẽ nhất. Nó không cho phép tạo mới các sản phẩm phái sinh để phục vụ cho mục đích lợi nhuận. Một nhánh giấy phép thứ hai ra đời sau đó là OPL (Open Publication License). Nó yêu cầu chỉ cần ghi công tác giả và cho phép được phân phối các sản phẩm phái sinh bằng một loại giấy phép khác có thể thu lợi nhuận. Cả hai nhánh này sau đó được kế thừa để hòa nhập tạo chung một dòng giấy phép truy cập mở được dùng phổ biến nhất hiện nay là Creative Commons (CC).

Giấy phép CC quy định các quyền tự do sao chép, xuất bản đi kèm với các điều kiện ràng buộc có thể được tùy chọn bao gồm: BY — phải ghi công tác giả; SA — không được thay đổi giấy phép cho các sản phẩm phái sinh; NC — không được phép thương mại hóa; ND — không được phép tùy biến sửa đổi, tạo các sản phẩm phái sinh. Tổ hợp các điều kiện lựa chọn khác nhau, ta có tương ứng các loại giấy phép CC khác nhau.

Một dòng giấy phép mở được sử dụng chuyên dùng cho dữ liệu là Open Data Commons (ODC). Các loại giấy phép có thể lựa chọn cho dữ liệu mở gồm có: PDDL (Public Domain Dedication and License) tương đương với giấy phép công cộng CC0; ODC-BY tương đương với giấy phép truy cập mở CC-BY (ghi công tác giả); và ODbL (Open Database License) tương đương với giấy phép truy cập mở CC BY-SA (ghi công và chia sẻ tương tự). Cả 3 loại giấy phép đều cho phép người dùng được tự do chia sẻ,

tạo dữ liệu mới hoặc sửa đổi cơ sở dữ liệu gốc. Trong trường hợp sử dụng ODbL thì các dữ liệu phái sinh phải được tiếp tục công bố với giấy phép tương tự.

1.2 Nguyên tắc dữ liệu mở

Dữ liệu mở phải được xây dựng trên cơ sở áp dụng bộ nguyên tắc FAIR quy định các yêu cầu cần được bảo đảm để dữ liệu có thể dễ dàng tìm thấy, truy cập, tương hợp và tái sử dụng bởi cả con người và máy tính, cụ thể như sau:

- *Khả năng tìm thấy (Findable)*: F1 — sử dụng định danh toàn cầu và vĩnh viễn cho dữ liệu và siêu dữ liệu; F2 — dữ liệu phải được mô tả đầy đủ với các thuộc tính siêu dữ liệu; F3 — siêu dữ liệu phải chứa tham chiếu tường minh tới định danh duy nhất của dữ liệu mà nó mô tả; F4 — dữ liệu và siêu dữ liệu được đăng kí và đánh chỉ mục trong một kho tìm kiếm.
- *Khả năng truy cập (Accessible)*: A1 — có thể truy xuất dữ liệu và siêu dữ liệu thông qua một giao thức tiêu chuẩn; A2 — siêu dữ liệu vẫn phải có khả năng truy cập được ngay cả khi dữ liệu không còn tồn tại nữa.
- *Khả năng tương hợp (Interoperable)*: I1 — sử dụng ngôn ngữ máy hiểu để biểu diễn dữ liệu và siêu dữ liệu; I2 — khai thác các từ điển thuật ngữ dùng chung tuân thủ bộ nguyên tắc FAIR; I3 — có thể chứa tham chiếu tới các bộ dữ liệu khác.
- *Khả năng tái sử dụng (Reusable)*: R1 — xuất bản dữ liệu và siêu dữ liệu đi kèm với giấy phép truy cập mở; R2 — có mô tả chi tiết về nguồn cung cấp dữ liệu; R3 — thỏa mãn các tiêu chuẩn ngành của lĩnh vực áp dụng.

1.3 Chuẩn cấp độ dữ liệu mở

Theo Tim Berners-Lee, các công nghệ của web ngữ nghĩa sẽ được dùng để quản lí chia sẻ dữ liệu trên mạng Internet trong tương lai. Công nghệ sử dụng cho hạ tầng chia sẻ dữ liệu có thể phân chia theo 5 mức độ tăng dần để phù hợp với nguyên tắc FAIR như sau.

- *Mức độ 1 — Cấp phép mở (Open License)*: đưa dữ liệu chia sẻ truy cập trên Internet và cấp giấy phép truy cập mở.
- *Mức độ 2 — Máy đọc được (Machine Readable)*: đã đạt mức độ 1 và dữ liệu phải được cung cấp dưới định dạng mà máy có thể đọc được.
- *Mức độ 3 — Định dạng mở (Open Format)*: đã đạt mức độ 2, cộng thêm yêu cầu phải sử dụng các định dạng dữ liệu tiêu chuẩn mở (không bị khống chế bởi một nhà cung cấp duy nhất).
- *Mức độ 4 — Định danh URI (Uniform Resource Identifier)*: đã đạt mức độ 3, cộng thêm yêu cầu phải sử dụng các mã định danh URI (thông qua biểu diễn XML) để mô tả (siêu) dữ liệu và các thuật ngữ dùng chung.
- *Mức độ 5 — Dữ liệu liên kết (Linked Data)*: đạt mức độ cao nhất thỏa mãn đủ các nguyên tắc tiêu chuẩn FAIR, cho phép tham chiếu tới các bộ dữ liệu khác trên toàn cầu (thông qua biểu diễn RDF).

2. Dữ liệu mở liên kết (LOD)

2.1 Mô hình dữ liệu liên kết

Mô hình dữ liệu liên kết là cách tổ chức và xuất bản dữ liệu sao cho mỗi thực thể (người, địa điểm, sự kiện, khái niệm…) được định danh bằng URI, có thể được truy cập qua HTTP và liên kết với các thực thể khác thông qua các quan hệ ngữ nghĩa chuẩn hóa. Nền tảng cốt lõi của dữ liệu liên kết là dựa trên mô hình dữ liệu RDF/S. Có 4 nguyên tắc do Tim Berners-Lee đề ra với dữ liệu liên kết:

1. Sử dụng URI để định danh đối tượng.
2. URI phải có thể truy cập qua HTTP.
3. Khi truy cập URI, cung cấp dữ liệu chuẩn (RDF, JSON-LD…).
4. Liên kết đến dữ liệu khác để mở rộng thông tin.

Một số dự án tiêu biểu

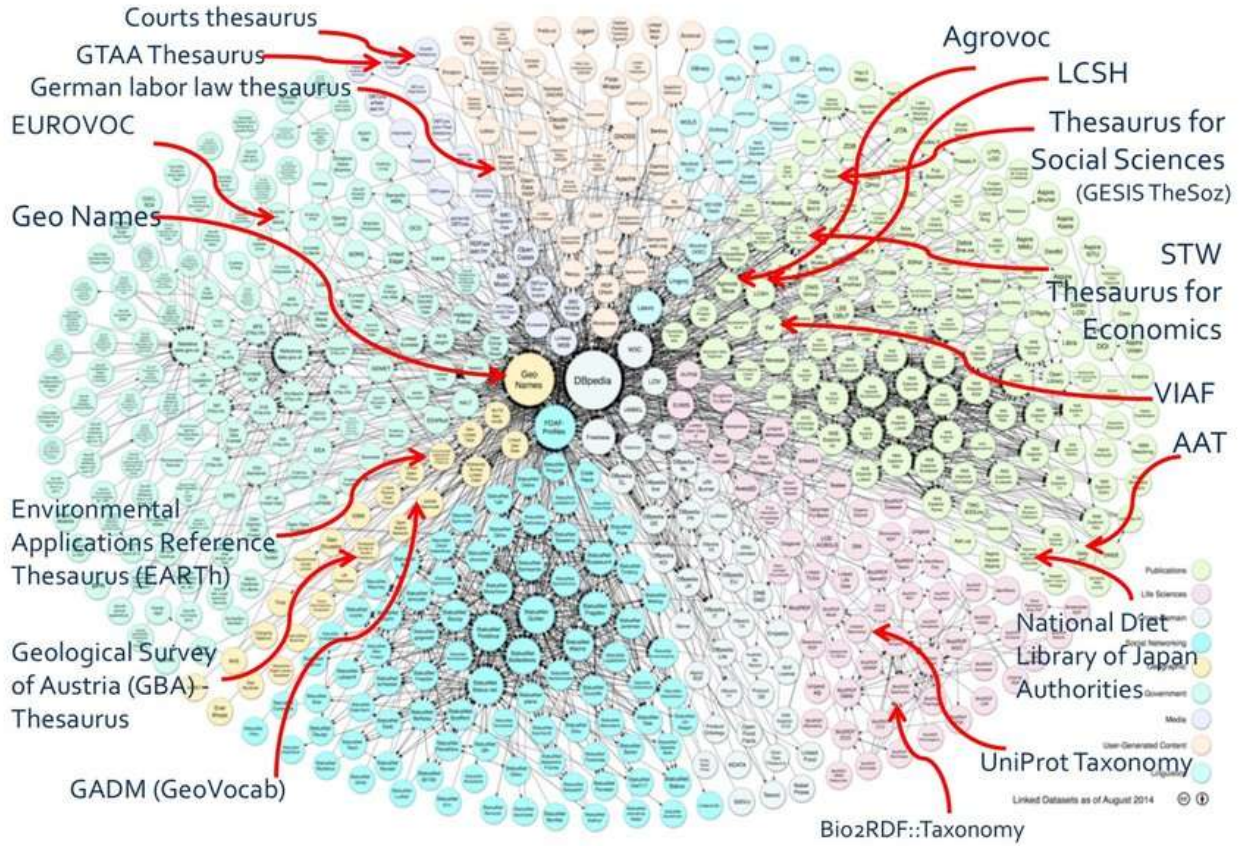
1. Dự án schema.org
   - Mục tiêu: Chuẩn hóa từ vựng mô tả dữ liệu trên web để các công cụ tìm kiếm (Google, Bing, Yahoo, Yandex…) hiểu được nội dung trang web.
   - Nguyên lý: Dùng JSON-LD hoặc RDFa để nhúng metadata vào HTML. Qua đó bổ sung thông tin nổi bật vào kết quả tìm kiếm web (rich snippets); tối ưu SEO ngữ nghĩa (Semantic SEO).
2. Dự án DBpedia
   - Nguồn gốc: Trích xuất dữ liệu có cấu trúc từ Wikipedia (infobox, categories) và công bố dưới dạng RDF.
   - Quy mô: Chứa hơn 6 triệu thực thể, được liên kết với các bộ dữ liệu khác (GeoNames, MusicBrainz, Wikidata…).
   - Ứng dụng: Là hub dữ liệu trong LOD Cloud, cung cấp dữ liệu nền cho NLP, chatbot, phân tích tri thức.
3. Dự án Wikidata
   - Mục tiêu: Kho dữ liệu tri thức mở, cộng đồng đóng góp, được duy trì bởi Wikimedia Foundation.
   - Khác với DBpedia: Không chỉ trích xuất từ Wikipedia, mà cho phép nhập dữ liệu trực tiếp. Mỗi mục (item) có ID Q-number (VD: Hà Nội = Q1858). Dữ liệu được lưu trữ ngôn ngữ độc lập, đa ngôn ngữ.
   - Ứng dụng: Nguồn dữ liệu cho Wikipedia, Google Knowledge Graph, trợ lý ảo; phân tích dữ liệu tri thức đa lĩnh vực; liên kết mạnh mẽ với các nguồn khác qua URI.

2.2 Đám mây dữ liệu mở liên kết

Linked Open Data Cloud (LOD Cloud) là tập hợp các bộ dữ liệu mở được công bố theo nguyên tắc dữ liệu liên kết của Tim Berners-Lee, trong đó các bộ dữ liệu này liên kết với nhau bằng URI chuẩn để tạo thành một mạng lưới dữ liệu toàn cầu. Hình ảnh LOD Cloud diagram thường được cập nhật bởi cộng đồng (LOD Community) và hiển thị dưới dạng biểu đồ bong bóng — mỗi bong bóng là một bộ dữ liệu, kích thước bong bóng tỉ lệ với số lượng bộ ba RDF, và đường nối thể hiện các liên kết liên miền (cross-dataset links).

Một bộ dữ liệu muốn được đưa vào sơ đồ LOD Cloud phải:

1. Mở (Open): được cấp phép sử dụng công khai (CC-BY, ODbL, Public Domain…).
2. Truy cập được qua HTTP.
3. Cung cấp SPARQL endpoint hoặc dump RDF để tải dữ liệu.
4. Có liên kết (linkset) tới ít nhất một bộ dữ liệu khác trong LOD Cloud.



LOD Cloud hiện có hơn 1.300 bộ dữ liệu (phiên bản cập nhật 2023–2024), với hơn 16 tỷ bộ ba RDF. Các bộ dữ liệu được phân loại thành các miền chủ đề (domain) chính:

- Cross-domain: DBpedia, Wikidata

- Media: BBC Programmes, Europeana
- Government: data.gov.uk, EU Open Data Portal
- Life sciences: UniProt, Bio2RDF, DrugBank
- Geographic: GeoNames, LinkedGeoData
- Publications: OpenCitations, CrossRef
- Social networking: FOAF profiles, SemanticTweet

Vai trò LOD Cloud trong Hệ sinh thái Dữ liệu Liên kết

- Hub kết nối dữ liệu toàn cầu: Các node trung tâm như DBpedia, Wikidata, GeoNames đóng vai trò như trục (hub) giúp các bộ dữ liệu khác liên kết.
- Tái sử dụng và kết hợp dữ liệu: Các nhà nghiên cứu và doanh nghiệp có thể kết hợp dữ liệu từ nhiều nguồn để tạo ra dịch vụ mới (mashup).
- Chuẩn hóa và liên thông dữ liệu: Việc dùng RDF, URI, ontology chung giúp dữ liệu từ nhiều nguồn trở nên tương thích.
- Thúc đẩy Semantic Web: LOD Cloud là phần “cơ sở hạ tầng dữ liệu” cho Web ngữ nghĩa.

Vị trí của các dự án tiêu biểu trong LOD Cloud

- DBpedia: Hub lớn, liên kết với hầu hết các bộ dữ liệu khác (GeoNames, Freebase, YAGO…).
- Wikidata: Node kết nối mạnh, nhiều liên kết ra/vào, phục vụ cả cộng đồng và doanh nghiệp.
- Schema.org: Không phải bộ dữ liệu trong LOD Cloud, nhưng là từ điển (vocabulary) quan trọng mà nhiều bộ dữ liệu LOD sử dụng để mô tả.

3. Chuẩn bị cho cuộc thi PMNM - OLP 2025

Cuộc thi phần mềm nguồn mở do Hội Tin học Việt Nam tổ chức hàng năm cho sinh viên của các trường đại học, cao đẳng trên cả nước tham gia. Thể lệ của cuộc thi năm 2025 đã được công bố trên trang thông tin của Ban tổ chức. Hình thức của cuộc thi là các đội tuyển tham gia làm một dự án phát triển phần mềm nguồn mở theo yêu cầu của đề thi do BTC công bố.

Chủ đề của cuộc thi năm 2025 là “Ứng dụng dữ liệu mở liên kết phục vụ chuyển đổi số". Mục đích của chủ đề là giúp sinh viên nắm bắt được xu thế chuyển đổi số dựa trên các nguồn dữ liệu mở liên kết. Qua đó sinh viên làm chủ được công nghệ và các kỹ năng cần thiết để xây dựng hệ sinh thái dữ liệu mở đáp ứng yêu cầu hòa nhập với các tiêu chuẩn chung trên thế giới.

Các trường tham gia cuộc thi cần lựa chọn sinh viên để huấn luyện đội tuyển với các kĩ năng, kiến thức cần thiết sau đây:

- Hiểu biết về dự án xây dựng phần mềm nguồn mở, dữ liệu mở: Nắm vững và có kĩ năng trong tạo lập và triển khai một dự án mở đáp ứng được các tiêu chí cần thiết theo đúng thông lệ quốc tế.

- Nắm vững kiến thức, công nghệ web ngữ nghĩa với dữ liệu mở liên kết: Tư duy thiết kế từ điển dữ liệu, ứng dụng ontology trong xây dựng các hệ thống thông minh dựa trên cơ sở tri thức.
- Làm quen, thực hành sử dụng các nền tảng công cụ mã nguồn mở hỗ trợ quản lý lưu trữ, khai thác, truy vấn, trao đổi dữ liệu liên kết bằng các ngôn ngữ như SPARQL, RDF/XML, JSON-LD,...
- Sáng tạo trong giải quyết vấn đề: Tạo lập và khai thác các nguồn dữ liệu mở sẵn có để tạo ra các dịch vụ giá trị gia tăng mới phục vụ phát triển kinh tế - xã hội.