# CÀI ĐẶT THỬ NGHIỆM MÔ HÌNH THƯ VIỆN SỐ DỰA TRÊN LOGIC VỊ TỪ

Lý Anh Tuấn[1], Trần Thị Minh Hoàn[2]
[1]Trường Đại học Thủy lợi. Emai: tuanla@wru.edu.vn
[2]Emai: hoantm@wru.edu.vn

## 1. GIỚI THIỆU

Theo tiếp cận của chúng tôi, thư viện số là một hệ thống thông tin hỗ trợ việc lưu trữ các tài nguyên số và cung cấp các dịch vụ phục vụ cho việc truy cập và thao tác với các tài nguyên đó. Phạm vi của các tài nguyên số bao gồm từ các đối tượng thông tin đơn giản (văn bản, hình ảnh, video, và những định dạng tương tự) cho đến các đối tượng kết hợp từ những định dạng khác nhau.

Các dịch vụ cơ bản của một thư viện số cho phép người sử dụng: (a) định danh một tài nguyên; (b) truy cập một tài nguyên; (c) mô tả một tài nguyên bằng một bộ từ vựng; (d) tạo một tài nguyên số phức hợp bằng việc sử dụng lại các tài nguyên sẵn có; và (e) khám phá các tài nguyên dựa vào siêu dữ liệu của chúng.

Để thiết kế và thi hành một hệ thống thư viện số hỗ trợ các dịch vụ trên, nhóm nghiên cứu chúng tôi phát triển một mô hình dữ liệu và một ngôn ngữ truy vấn [2]. Mô hình hướng tới các ứng dụng Web và được trình bày như một lý thuyết cấp một. Thêm vào đó một phép chuyển đổi từ mô hình tới RDF [3] và từ ngôn ngữ truy vấn tới SPARQL [4] đã được đề xuất để minh hoạ tính khả thi của mô hình. Sự lựa chọn RDF là do trên thực tế nó là ngôn ngữ biểu diễn được chấp nhận rộng rãi trong ngữ cảnh của các thư viện số và Semantic Web.

Trong phần công việc bài báo này mô tả, chúng tôi đã thiết kế và phát triển một nguyên mẫu thử nghiệm nhằm chỉ ra tính khả thi của mô hình lý thuyết [1]. Nguyên mẫu được xây dựng dựa trên các công nghệ Jena và bộ công cụ Web của Google. Nguyên mẫu đã được xây dựng, được kiểm thử, được gỡ lỗi cục bộ và được triển khai thử nghiệm. Trong tương lai, nó sẽ được mở rộng để trở thành một thư viện số chính thức đầy đủ. Bài báo này sẽ trình bày tóm tắt về công việc thiết kế và cài đặt nguyên mẫu.

## 2. MỘT MÔ HÌNH DỮ LIỆU VÀ NGÔN NGỮ TRUY VẤN CHO CÁC THƯ VIỆN SỐ

### 2.1. Các khái niệm cơ bản của mô hình

- Tài nguyên

Trong một thư viện số tài nguyên được hiểu là bất cứ thứ gì có thể được định danh. Trong mô hình của nhóm nghiên cứu chúng tôi, các tài nguyên được chia làm hai loại là *các tài nguyên số* và *các tài nguyên phi số*.

- Một tài nguyên số là một mẫu dữ liệu ở dạng số chẳng hạn như một tài liệu PDF, một ảnh JPEG, một văn bản được số hóa, vân vân.
- Một tài nguyên phi số là bất cứ tài nguyên nào không ở dạng số. Các tài nguyên phi số có thể là các đối tượng vật lý hoặc các thực thể trừu tượng.

***- Bảng tham chiếu***

Chúng tôi gọi các định danh là các tài nguyên số được sử dụng làm phương tiện để tham chiếu đến các tài nguyên khác trong một thư viện số. *Bảng tham chiếu* của thư viện số lưu giữ tất cả các mối liên hệ giữa các định danh và các tài nguyên được tham chiếu nằm trong thư viện số.

***- Cơ sở siêu dữ liệu***

*Cơ sở siêu dữ liệu* của một thư viện số bao gồm hai kiểu tri thức mà người dùng có thể biểu diễn về một tài nguyên: *siêu dữ liệu* và *nội dung* của tài nguyên.

*Siêu dữ liệu:* Để thực hiện các chức năng, thư viện số cần phải biểu diễn, lưu trữ và xử lý một lượng nhất định thông tin về tài nguyên. Ví dụ, các thông tin về độ phân giải, tên tác giả, biểu đồ màu sắc,… Các thông tin này sẽ được mô hình hoá dưới dạng được gọi là *mô tả*. Chúng tôi mô hình hoá các mô tả bằng việc định nghĩa hai vị từ sau đây:

- DescCl(d, s, c) biểu diễn sự thật rằng lớp c trên toàn lược đồ s thuộc về mô tả d.
- DescPr(d, s, p, i) biểu diễn sự thật rằng cặp thuộc tính-giá trị (p, i), trong đó p trên toàn lược đồ s, thuộc về d.

Mối liên kết giữa các tài nguyên và các mô tả của chúng được mô hình hoá bằng vị từ:

- DescOf(d, i) biểu diễn sự thật rằng d định danh một mô tả của một tài nguyên được định danh bởi i.

*Siêu dữ liệu của tài nguyên* là tập tất cả các mô tả liên kết với tài nguyên đó.

Các khái niệm lớp, thuộc tính, giá trị và lược đồ ở trên là các khái niệm trừu tượng của một ngôn ngữ học, tức là các tài nguyên phi số được tham chiếu tới bởi các định danh. Tri thức trong một lược đồđược bắt giữ bằng tập các vị từ sau:

- SchCl(s, c) - bắt giữ mối liên hệ giữa các lớp và các lược đồ;
- SchPr(s, p) - bắt giữ mối liên hệ giữa các thuộc tính và các lược đồ;
- IsaCl($s, c_1, c_2$) - bắt giữ mối quan hệ is-a giữa các lớp;
- IsaPr($s, p_1, p_2$) - bắt giữ mối quan hệ is-a giữa các thuộc tính;
- Dom($s,p,c$) - bắt giữ miền của các thuộc tính;
- Ran($s,p,c$) - bắt giữ phạm vi của các thuộc tính.

*Nội dung:* Nội dung của một tài nguyên r là tập các tài nguyên khác cấu tạo nên r từ quan điểm ứng dụng; mỗi tài nguyên này được gọi là một thành phần của r. Nội dung có thể được biểu diễn bởi vị từ:

- PartOf(i, j) biểu diễn sự thật rằng i định danh một tài nguyên có thể hợp thành là một thành phần của tài nguyên có thể hợp thành được định danh bởi j.

### 2.2. Định nghĩa một thư viện số

Gọi Σ là tập tất cả các tên của các quan hệ đã được giới thiệu, tức là: Σ = {SchCl, SchPr, Dom, Ran, IsaCl, IsaPr, DescCl, DescPr, DescOf, PartOf}. Chúng ta có định nghĩa sau đây về thư viện số:

**Định nghĩa 1. (Thư viện số)** Một thư viện số D là một cặp D = (REF, I) trong đó:

- *REF, bảng tham chiếu của D, là một hàm hữu hạn trên toàn bộ các tài nguyên số.*
- *I, cơ sở siêu dữ liệu của D, là một hàm tổng liên kết tất cả các tên trong tập Σ với một quan hệ trên toàn bộ các tài nguyên số.*

Một thư viện số là *nhất quán* nếu tập các định danh của các kiểu tài nguyên khác nhau là tách rời nhau từng đôi một.

Một thư viện số là *hoàn chỉnh* nếu nó bao chứa cả tri thức được ghi lại bởi người dùng (tức là tri thức tường minh) và tri thức ẩn được suy diễn từ tri thức tường minh.

**Định nghĩa 2. (Ngôn ngữ truy vấn thư viện số)** Ngôn ngữ truy vấn thư viện số Q được định nghĩa quy nạp như sau:

- Một từ truy vấn *là một định danh hoặc một biến*.
- Một truy vấn nguyên tử *là một công thức* $P(t_1,\ldots, t_n)$, *P là một ký hiệu vị từ trong Σ sao cho* $t_1, \ldots, t_n$ *là các từ truy vấn bao gồm ít nhất một biến*.
- Một truy vấn hội *là một công thức* $\alpha_1 \wedge \ldots \wedge \alpha_k$, *trong đó* $k \geq 1$ *và* $\alpha_i$ *là một truy vấn nguyên tử, với mọi* $1 \leq i \leq k$.

## 3. CHUYỂN ĐỔI MÔ HÌNH TỚI RDF VÀ NGÔN NGỮ TRUY VẤN TỚI SPARQL

Phép chuyển đổi mô hình tới RDF bao gồm việc cài đặt cơ sở siêu dữ liệu là một đồ thị RDF, và việc sử dụng một máy suy diễn RDF để tính toán tính hoàn chỉnh của thư viện số. Để hoàn thành công việc thứ nhất chúng ta cần chuyển đổi các định danh của mô hình thành các tham chiếu URI và chuyển đổi các bộ siêu dữ liệu thành các bộ ba RDF. Để hoàn thành công việc thứ hai chúng ta cần định nghĩa một kỹ thuật suy diễn mới là sự mở rộng kỹ thuật suy diễn của lược đồ RDF.

Phép chuyển đổi ngôn ngữ truy vấn tới SPARQL được thực hiện thông qua việc trích ra từ truy vấn ban đầu một tập các biến tự do và dịch chúng thành các biến SPARQL, và việc ánh xạ truy vấn ban đầu thành mẫu đồ thị tương đương. Hai thành phần này cấu thành một truy vấn SPARQL hoàn chỉnh theo khuôn dạng *SELECT <danh sách biến> WHERE {<mẫu đồ thị>}*, tương ứng với truy vấn ban đầu.

## 4. THIẾT KẾ VÀ CÀI ĐẶT THỬ NGHIỆM

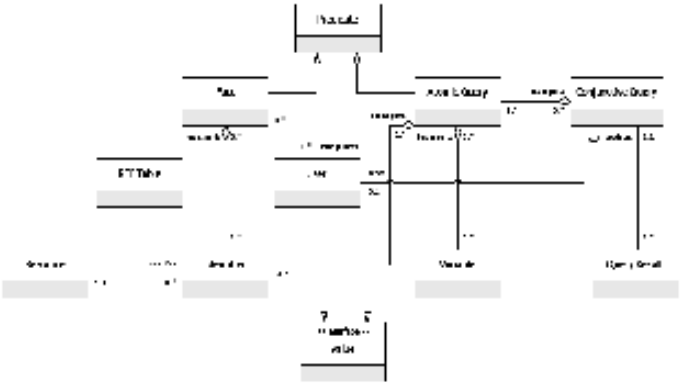
### 4.1 Các yêu cầu

- *Quản lý siêu dữ liệu.* Người cung cấp quản lý các siêu dữ liệu của thư viện số thông qua một giao diện (một trang Web). Tất cả các kiểu siêu dữ liệu khác nhau của mô hình phải được hỗ trợ.
- *Truy vấn thông tin từ thư viện số.* Thông qua giao diện, người dùng cuối tạo ra các truy vấn thông tin và gửi nó cho thư viện số. Các thông tin thoảmãn truy vấn sẽ được hiển thị ra cho người dùng.
- Nguyên mẫu phải hỗ trợ việc chuyển đổi từ các vị từ của mô hình thành các bộ ba RDF và hỗ trợ việc chuyển đổi ngược lại từ các bộ ba RDF thành các vị từ.

- Nguyên mẫu phải hỗ trợ việc chuyển đổi các truy vấn hội của ngôn ngữ truy vấn thành các truy vấn dạng SPARQL và thực thi chúng trên bộ lưu trữ RDF.

### 4.2 Mô hình khái niệm

Để thoả mãn các yêu cầu người sử dụng và các yêu cầu hệ thống, chúng tôi đã thiết kế nguyên mẫu dựa vào các khái niệm được trình bày trong Hình 1 dưới dạng một lược đồ lớp UML.



***Hình 1:*** *Lược đồ lớp UML của nguyên mẫu*

### 4.3 Kiến trúc và các thành phần

Trong phần này chúng ta thảo luận về kiến trúc tổng quan của nguyên mẫu thư viện số đã được đề xuất. Nó bao gồm các thành phần chính sau đây:

- *Một giao diện người dùng* cho phép người dùng tương tác với thư viện số dựa vào quyền họ được cấp.
- *Các thành phần chuyển đổi* bao gồm thành phần chuyển đổi truy vấn và thành phần chuyển đổi siêu dữ liệu.
- *Thành phần trả lời truy vấn* đảm nhiệm việc thực thi các truy vấn SPARQL trên một bộ lưu trữ RDF đóng vai trò là cơ sở siêu dữ liệu của thư viện số.
- *Thành phần quản lý bền vững* quản lý các thao tác đọc/ghi các bộ ba RDF đối với bộ lưu trữ RDF.
- *Bộ lưu trữ thư viện số* bao gồm bộ lưu trữ RDF, bảng tham chiếu và kho dữ liệu lưu trữ các tài nguyên số.

### 4.4 Giao diện người sử dụng

Giao diện người dùng của ứng dụng được phát triển dựa trên bộ công cụ Web của Google. Đây là một nền tảng hỗ trợ việc phát triển các ứng dụng RIA, tức là các ứng dụng Web có khả năng tương tác tốt với người sử dụng. Một số trang chính của giao diện là trang quản lý siêu dữ liệu; và trang truy vấn siêu dữ liệu. Trong đó trang truy vấn siêu dữ liệu cho phép người dùng tạo ra và thực thi các truy vấn hội để khám phá tri thức có trong thư viện số.

## 5. KẾT LUẬN

Thông qua việc thiết kế và cài đặt một nguyên mẫu dựa trên mô hình lý thuyết chúng tôi đã kiểm chứng được tính khả thi của mô hình. Nguyên mẫu đã minh hoạ hầu hết các khía cạnh của mô hình và hoàn toàn có thể được mở rộng để trở thành một thư viện số chính thức với đầy đủ các chức năng.

Nguyên mẫu được phát triển dựa trên RDF và SPARQL, đây là ngôn ngữ biểu diễn và ngôn ngữ truy vấn được chấp nhận rộng rãi trong ngữ cảnh của các thư viện số và Semantic Web. Điều này cũng đảm bảo tính tương hợp và tính khả mở của nguyên mẫu.

## TÀI LIỆU THAM KHẢO

[1]. Anh Tuan Ly (2013)Accessing and Using Complex Multimedia Documents in a Digital Library, Ph.D. Thesis, Université Paris-Sud. Thesis Advisor: Prof. Nicolas Spyratos.

[2]. Carlo Meghini, Nicolas Spyratos, Tsuyoshi Sugibuchi, Jitao Yang (2014) A Model for Digital Libraries and its Translation to RDF. J. Data Semantics 3(2): 107-139.

[3]. Klyne G, Carroll JJ (2004) Resource description framework (RDF): concepts and abstract syntax.W3C Recommendation, WWWConsortium. http://www.w3.org/TR/rdf-concepts/

[4]. Prud'hommeaux E, Seaborne A (2008) Sparql query language for RDF. W3C Recommendation. http://www.w3.org/TR/rdf-sparql-query