

MỘT PHƯƠNG PHÁP CHUYỂN ĐỔI TÀI LIỆU XML SANG TÀI LIỆU RDF SỬ DỤNG NGÔN NGỮ XSL

Nguyễn Dũng*, Nguyễn Mậu Hân

Khoa Công nghệ Thông tin, Trường Đại học Khoa học, ĐH Huế

*Email: nguyendung@hueuni.edu.vn

Ngày nhận bài: 23/10/2018; ngày hoàn thành phản biện: 01/3/2019; ngày duyệt đăng: 10/3/2019

TÓM TẮT

Ngày nay, ngôn ngữ đánh dấu mở rộng XML (Extensible Markup Language) đã trở thành một định dạng trao đổi dữ liệu được chấp nhận rộng rãi. Cùng với việc gia tăng lượng Dữ liệu Liên kết (Linked Data), bằng cách sử dụng Khung mô tả Tài nguyên RDF (Resource Description Framework), thì nhiều công cụ nhằm biến đổi một tài liệu XML thành tài liệu RDF đã được phát triển. Trong bài báo này, chúng tôi sử dụng ngôn ngữ định kiểu mở rộng XSL (Extensible Stylesheet Language) để biến đổi tài liệu XML thành tài liệu RDF như là một giải pháp hiệu quả để chuyển dữ liệu hiện có của các hệ thống thành dữ liệu RDF được sử dụng trong các hệ thống web ngữ nghĩa.

Từ khóa: dữ liệu liên kết, rdf, xml, xsl, web ngữ nghĩa.

1. GIỚI THIỆU

XML là một bộ qui luật về cách chia một tài liệu ra làm nhiều phần, rồi đánh dấu và ráp các phần khác nhau lại để dễ nhận diện chúng. XML được phát triển bởi W3C (World Wide Web Consortium), một nhóm cộng tác (working group) do Jonh Bosak ở Sun Microsystem thành lập năm 1996, nhằm khắc phục những hạn chế của HTML (Hyper Text Markup Language, ngôn ngữ đánh dấu siêu văn bản, là cơ sở để mọi trang Web sử dụng để hiển thị nội dung).

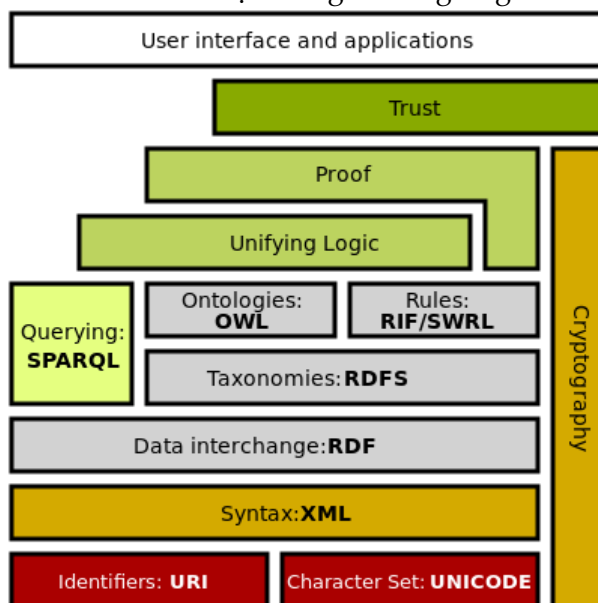
XML cho phép chia sẻ và sử dụng thông tin phân tán trên các hệ thống khác nhau và hỗ trợ người dùng thông qua khả năng tạo nội dung động, phát triển ứng dụng và tích hợp trên nhiều qui mô khác nhau. XML khá giống HTML, hai ngôn ngữ này có cùng luật cú pháp, cả hai đều được lưu trữ dưới dạng các tập tin văn bản với phần mở rộng html hoặc xml. Tuy nhiên, tính linh hoạt của XML cho phép tạo và sử dụng tập thể, thuộc tính riêng để nhận biết các phần tử cấu trúc và nội dung tài liệu. XML không chỉ đơn thuần là ngôn ngữ đánh dấu, mà còn là phương pháp định ra nội dung tài liệu, tương tự như HTML định hình thức tài liệu trên Web. Với HTML, người

thiết kế đánh dấu văn bản, hình ảnh cùng các thành phần khác của trang Web bằng tập thẻ mà không liên quan gì tới ý nghĩa tài liệu; đoạn mã HTML chỉ nói lên cách hiển thị nội dung tài liệu qua trình duyệt. XML không chỉ định ra hình thức mà còn cả nội dung tài liệu. Tóm lại, XML và HTML đều là các ngôn ngữ đánh dấu nhưng XML được thiết kế để mang và lưu trữ dữ liệu chứ không phải để biểu diễn dữ liệu. Còn HTML được thiết kế để hiển thị dữ liệu, được tập trung vào cách dữ liệu được hiển thị như thế nào.

Tuy nhiên, việc sử dụng định dạng XML để trao đổi dữ liệu giữa các hệ thống nó mới chỉ phục vụ cho các hệ thống Web 2.0, tức là mối quan hệ ngữ nghĩa giữa các đối tượng chưa được xét đến. Các loại dữ liệu này chỉ dừng lại mức con người có thể hiểu và xử lý. Với mục đích biểu diễn dữ liệu sao cho đồng thời máy tính và con người cùng hiểu và xử lý được thì Web ngữ nghĩa, hay Web 3.0, đã ra đời. Việc trao đổi thông tin trên Web ngữ nghĩa người ta dùng định dạng RDF. Do đó việc chuyển đổi dữ liệu từ hệ thống hiện tại, theo định dạng XML, sang hệ thống Web ngữ nghĩa, theo định dạng RDF, là nhu cầu hết sức bức thiết.

2. KHUNG MÔ TẢ TÀI NGUYÊN WEB RDF

Khung mô tả tài nguyên Web RDF là một nhóm các đặc tả của tổ chức World Wide Web Consortium (W3C) được thiết kế như là một ngôn ngữ siêu dữ liệu để biểu diễn dữ liệu trên Web, đồng thời cung cấp một mô hình để mô tả và tạo các mối quan hệ giữa các tài nguyên. Theo mô tả trong hình 1 dưới đây, dữ liệu trong Web ngữ nghĩa được tổ chức dựa trên cơ sở XML và được mô hình hóa bằng RDF. Như vậy RDF được chọn như là chuẩn trao đổi dữ liệu trong Web ngữ nghĩa.



Hình 1. Kiến trúc phân tầng của Web ngữ nghĩa

RDF định nghĩa một nguồn tài nguyên (resource), sau này chúng tôi gọi là chủ thể (subject) như là một đối tượng bất kỳ có khả năng xác định duy nhất bằng cách sử dụng một định danh tài nguyên thống nhất URI (Uniform Resource Identifier). Các nguồn tài nguyên có các thuộc tính (property) đi kèm, sau này chúng tôi gọi là các vị từ (predicate). Các vị từ biểu diễn các mối quan hệ giữa các chủ thể và cũng được xác định bởi một URI. Giá trị của các vị từ gọi là đối tượng (object).

Như vậy mô hình dữ liệu của RDF là một tập các bộ ba (triple), sau này chúng tôi gọi là một tập các phát biểu (statement) gồm 3 thành phần: (Chủ thể, Vị từ, Đối tượng), trong đó:

- Chủ thể (Resource/Subject): là các tài nguyên được mô tả. Các tài nguyên này được xác định bởi URI cụ thể.
- Vị từ (Property/Predicate): Thuộc tính của các tài nguyên hay chủ thể. Các vị từ này cũng được xác định bởi các URI.
- Đối tượng (Object): Giá trị của các vị từ. Có thể là một giá trị (literal) hoặc một chủ thể khác được xác định bởi một URI

Các phát biểu trong tài liệu RDF cũng có thể được biểu diễn dưới dạng đồ thị, gọi là đồ thị RDF. Có thể hình dung đồ thị RDF đơn giản như sau:



Hình 2. Đồ thị RDF

Trong đó:

- Hình oval đại diện cho một chủ thể
- Mũi tên đại diện cho một vị từ với nhãn là định danh của vị từ
- Hình chữ nhật đại diện cho giá trị của một vị từ là literal. Nếu giá trị của vị từ là một chủ thể khác thì chúng ta lại sử dụng hình oval như trên.

Ngoài cách thể hiện trên, các tài liệu RDF có thể sử dụng các định dạng khác nhau như: N3, N-Triples, Turtle, RDFa, JSON-LD, TriG và RDF/XML nhằm để biểu diễn, trao đổi và liên kết dữ liệu. Trong đó định dạng RDF/XML là một định dạng được sử dụng phổ biến được định nghĩa bởi W3C để thể hiện một đồ thị RDF như là một tài liệu XML. Cú pháp của RDF/XML như sau:

```
<?xml version="1.0" encoding="UTF-8"?>
```

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
```

```
<rdf:Description rdf:about="subject">
```

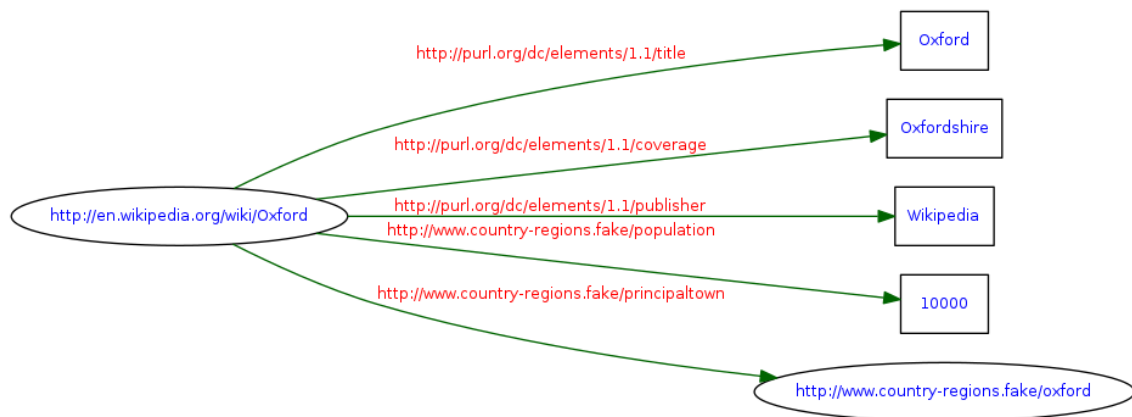
```
<predicate rdf:resource="object" />
```

```
<predicate>literal value</predicate>
<rdf:Description>
</rdf:RDF>
```

Trong đó, dòng đầu tiên là phần mô tả tài liệu XML. Theo sau phần mô tả là thẻ gốc `rdf:RDF`. Thuộc tính `xmlns:rdf` chứa không gian tên `http://www.w3.org/1999/02/22-rdf-syntax-ns#` là một không gian chuẩn của tổ chức W3C. Không gian tên này sẽ chỉ cho các bộ máy đọc hiểu rằng, nội dung bên trong là một tài liệu RDF và thẻ `rdf:RDF` thuộc không gian tên này. Trong tài liệu RDF có thể chứa một hoặc nhiều phát biểu. Để mô tả các chủ đề của phát biểu, chúng ta sử dụng thẻ `rdf:Description` và thuộc tính `rdf:about` sẽ chứa định danh URI của chủ đề. Nếu các phát biểu có cùng chủ đề, thì các phát biểu này sử dụng chung một thẻ `rdf:Description`. Cụ thể, cho một tài liệu RDF/XML như sau:

```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:dc="http://purl.org/dc/elements/1.1/"
  xmlns:region="http://www.country-regions.fake/">
  <rdf:Description rdf:about="http://en.wikipedia.org/wiki/Oxford">
    <dc:title>Oxford</dc:title>
    <dc:coverage>Oxfordshire</dc:coverage>
    <dc:publisher>Wikipedia</dc:publisher>
    <region:population>10000</region:population>
    <region:principaltown rdf:resource="http://www.country-
regions.fake/oxford"/>
  </rdf:Description>
</rdf:RDF>
```

Tài liệu RDF/XML trên là kết quả của đồ thị RDF sau:



Hình 3. Một ví dụ về đồ thị RDF.

3. PHƯƠNG PHÁP CHUYỂN ĐỔI

Không mất tính tổng quát, chúng tôi minh họa phương pháp chuyển đổi bằng tài liệu XML như dưới đây:

```
<?xml version='1.0' ?>
<?xml-stylesheet type='text/xsl' href='stylebook.xsl'?>
<bookstore>
  <book category="programing">
    <title lang="en">C LANGUAGE</title>
    <author>Giada De Laurentiis</author>
    <year>2005</year>
    <price>30.00</price>
  </book>
  <book category="web">
    <title lang="en">XQuery</title>
    <author>James McGovern</author>
    <author>Per Bothner</author>
    <year>2003</year>
    <price>49.99</price>
  </book>
</bookstore>
```

3.1. Biến đổi các thuộc tính

Các thuộc tính trong các thẻ của tài liệu XML có một sự tương ứng tự nhiên với bộ ba của tài liệu RDF. Trong đó, mỗi thuộc tính XML sẽ tương ứng với một thuộc tính trong RDF, thẻ chứa thuộc tính sẽ tương ứng với chủ thể (subject) và giá trị của thuộc tính tương ứng với đối tượng trong RDF.

Ví dụ: Xem xét thẻ `<book category="programing">`. Thẻ book này có thuộc tính là category, giá trị thuộc tính là programing. Do đó, chúng ta sẽ có được đồ thị RDF như sau:



Hình 4. Minh họa chuyển đổi thuộc tính

3.2. Biến đổi các giá trị văn bản trong các thẻ

Các giá trị văn bản trong các thẻ XML được biến đổi như là các literal. Các literal này được xem như là đối tượng trong các phát biểu RDF. Lúc này có hai trường hợp xảy ra:

- Nếu thẻ chứa giá trị văn bản không có thuộc tính thì chủ thể của literal này chính là thẻ cha của thẻ chứa giá trị này và vị từ của các literal chính là thẻ chứa giá trị văn bản.

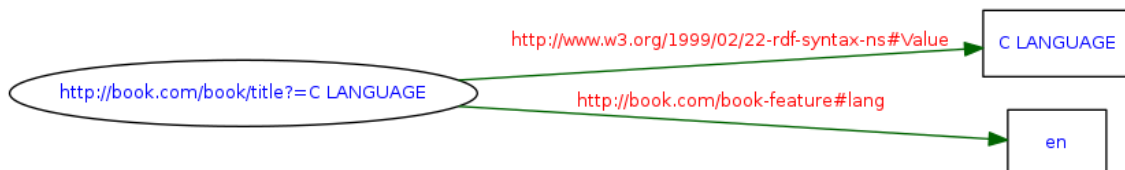
Ví dụ: Xem xét thẻ `<price>30.00</price>`. Thẻ này có giá trị là 30.00, không có thuộc tính và thẻ cha của nó là book. Do đó, chúng ta có đồ thị RDF như sau:



Hình 5. Minh họa chuyển đổi giá trị văn bản trong thẻ không có thuộc tính

- Nếu thẻ chứa giá trị văn bản nhưng có thuộc tính thì chủ thể của literal này là thẻ chứa giá trị và vị từ của các literal là **rdf:value**.

Ví dụ: Xem xét thẻ `<title lang="en">C LANGUAGE</title>`. Thẻ này có thuộc tính là lang (giá trị là en) và giá trị văn bản của nó là C LANGUAGE. Do đó, chúng ta có đồ thị RDF như sau:



Hình 6. Minh họa chuyển đổi giá trị văn bản trong thẻ có thuộc tính

3.3. Biến đổi các chú thích

Các chú thích XML được giữ nguyên trong các bộ ba của RDF. Lúc này trong các phát biểu của RDF, chúng sẽ được biến đổi thành đối tượng, còn vị từ bổ sung cho nó là: **xs:comment** và chủ thể chính là nút cha chứa chú thích.

3.4. Biến đổi các thẻ XML

Đối với các thẻ không có thẻ con, tức là thẻ chứa giá trị là literal thì lúc này chúng ta quay lại trường hợp 3.2, tức là biến đổi giá trị văn bản của các thẻ.

Đối với các thẻ có danh sách các thẻ con thì lúc này ta xem thẻ là chủ thể và lúc này các thẻ con sẽ được biến đổi theo hình thức đệ quy theo các trường hợp đã nêu ở mục 3.1, 3.2, 3.3 và cả 3.4.

3.5. Sử dụng ngôn ngữ XSL để biến đổi

Áp dụng các phương pháp biến đổi trên, chúng tôi sử dụng ngôn ngữ XSL để tiến hành biến đổi tài liệu XML đã cho ở trên như sau:

```
<?xml version="1.0"?>
<xsl:stylesheet version="1.0"
  xmlns:xsl="http://www.w3.org/1999/XSL/Transform"
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:book-feature="http://www.bookstore.com/book-features#">
  <xsl:template match="/">
    <rdf:RDF>
      <rdf:Description rdf:about="http://www.bookstore.com/">
        <xsl:apply-templates/>
      </rdf:Description>
    </rdf:RDF>
  </xsl:template>
  <xsl:template match="book">
    <xsl:variable name="title"><xsl:value-of select="title"/></xsl:variable>
    <xsl:variable name="author"><xsl:value-of select="author"/></xsl:variable>
    <xsl:variable name="year"><xsl:value-of select="year"/></xsl:variable>
    <xsl:variable name="price"><xsl:value-of select="price"/></xsl:variable>
    <book-feature:hasBook>
```

Một phương pháp chuyển đổi tài liệu XML sang tài liệu RDF sử dụng ngôn ngữ XSL

```
<rdf:Description rdf:about="http://www.bookstore.com/book/{title}">
  <book-feature:category><xsl:value-of select="@category"/>
</book-feature:category>
  <book-feature:title>
    <rdf:Description
rdf:about="http://www.bookstore.com/book?title={title}">
      <rdf:value><xsl:value-of select="title"/></rdf:value>
      <book-feature:lang><xsl:value-of select="title/@lang"/>
    </book-feature:lang>
  </rdf:Description>
</book-feature:title>
  <book-feature:price><xsl:value-of select="price"/></book-feature:price>
  <book-feature:author>
    <rdf:Bag>
      <xsl:for-each select="author">
        <rdf:li><xsl:value-of select="."/></rdf:li>
      </xsl:for-each>
    </rdf:Bag>
  </book-feature:author>
</rdf:Description>
</book-feature:hasBook>
</xsl:template>
</xsl:stylesheet>
```

Và cuối cùng chúng ta nhận được tài liệu kết quả RDF/XML sau khi biến đổi tài liệu XML bằng ngôn ngữ XSL:

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
  <rdf:Description rdf:about="http://www.bookstore.com/">
    <book-feature:hasBook xmlns:book-feature="http://www.bookstore.com/book-features#">
      <rdf:Description rdf:about="http://www.bookstore.com/book/C LANGUAGE ">
```



```

<book-feature:title>
  <rdf:Description rdf:about="http://www.bookstore.com/book?title=C
LANGUAGE">
    <book-feature:category>programing</book-feature:category>
    <rdf:value>C LANGUAGE</rdf:value>
    <book-feature:lang>en</book-feature:lang>
  </rdf:Description></book-feature:title>
<book-feature:price>30.00</book-feature:price>
<book-feature:author>
  <rdf:Bag><rdf:li>Giada De Laurentiis</rdf:li></rdf:Bag>
</book-feature:author>
</rdf:Description>
</book-feature:hasBook>
<book-feature:hasBook xmlns:book-feature="http://www.bookstore.com/book-
features#">
  <rdf:Description rdf:about="http://www.bookstore.com/book/XQuery">
    <book-feature:category>web</book-feature:category>
    <book-feature:title>
      <rdf:Description rdf:about="http://www.bookstore.com/book?title=XQuery">
        <rdf:value>XQuery</rdf:value>
        <book-feature:lang>en</book-feature:lang>
      </rdf:Description>
    </book-feature:title>
    <book-feature:price>49.99</book-feature:price>
    <book-feature:author>
      <rdf:Bag>
        <rdf:li>James McGovern</rdf:li>
        <rdf:li>Per Bothner</rdf:li>
      </rdf:Bag>
    </book-feature:author>

```

```
</rdf:Description>  
</book-feature:hasBook>  
</rdf:Description>  
</rdf:RDF>
```

Đồng thời sau khi có kết quả trên, chúng tôi tiến hành kiểm chứng bằng cách sử dụng công cụ Validator của tổ chức W3C tại địa chỉ: <https://www.w3.org/RDF/Validator/> và nhận được kết quả là hoàn toàn chính xác.

4. KẾT LUẬN

Trong bài báo này, chúng tôi thực hiện việc chuyển đổi thành công dữ liệu trong các tài liệu XML sang tài liệu RDF theo định dạng RDF/XML. Việc làm này sẽ giúp cho việc chuyển đổi mô hình dữ liệu theo dạng truyền thống sang dạng dữ liệu mà máy tính có thể đọc được một cách nhanh chóng bằng ngôn ngữ XSL, phục vụ cho tiến trình phát triển của Web ngữ nghĩa. Trong tương lai, chúng tôi sẽ cố gắng phát triển thuật toán cho phép chuyển đổi một tài liệu XML có cấu trúc bất kỳ sang định dạng RDF một cách tự động hóa, từ đó giảm thiểu thời gian và công sức chuyển đổi như hiện nay.

TÀI LIỆU THAM KHẢO

- [1]. [1]. W3C. Extensible Markup Language (XML). Website: <http://www.w3.org/XML/>
- [2]. [2]. W3C. Resource Description Framework (RDF). Website: <http://www.w3.org/RDF/>
- [3]. [3]. Joe Fawcett, Liam R.E. Quin, Danny Ayers (2012), *Begin XML*, Fifth edition, John Wiley & Sons, Inc.
- [4]. [4]. Jyun-Yao Huang, Christoph Lange, Sören Auer (2015). Streaming Transformation of XML to RDF using XPath-based Mappings, ResearchGate, Website: <https://www.researchgate.net/publication/301377290>
- [5]. [5]. S. Bischof (2012). Mapping between RDF and XML with XSPARQL, *Journal on Data Semantics 1.3*, pp. 147–185, Website: <http://dx.doi.org/10.1007/s13740-012-0008-7>

A METHOD FOR TRANSFORM XML DOCUMENTS TO RDF DOCUMENTS USING LANGUAGE XSL

Nguyen Dung*, Nguyen Mau Han

Faculty of Information Technology, University of Sciences, Hue University

*Email: nguyendung@hueuni.edu.vn

ABSTRACT

Today, the Extensible Markup Language (XML) has become a widely accepted form of data exchange. In addition to increasing the amount of Linked Data by using the Resource Description Framework (RDF), many tools for transforming an XML document into an RDF one have been developed. In this article, we use Extensible Stylesheet Language (XSL) to transform XML documents into RDF documents as an effective solution to transfer existing data from systems to data RDF that is used in semantic web systems.

Keywords: linked data, RDF, semantic web, XML, XML.



Nguyễn Dũng sinh ngày 13/06/1988 tại Thừa Thiên Huế. Năm 2010, ông tốt nghiệp cử nhân ngành Tin học và năm 2013 ông tốt nghiệp thạc sĩ chuyên ngành Khoa học máy tính tại Trường Đại học Khoa học, ĐH Huế. Hiện nay là giảng viên Khoa Công nghệ Thông tin, Trường Đại học Khoa học, Đại học Huế.

Lĩnh vực nghiên cứu: giải thuật di truyền, ứng dụng di động, xml và ứng dụng...



Nguyễn Mậu Hân sinh năm 1957 tại Thừa thiên Huế. Năm 1981, ông tốt nghiệp cử nhân toán tại trường Đại học Tổng hợp Huế. Năm 1998 nhận bằng thạc sĩ về Khoa học máy tính tại Trường đại học Bách khoa Hà Nội. Năm 2003, nhận bằng tiến sĩ chuyên ngành Khoa học Máy tính tại Viện Công nghệ Thông tin Hà nội. Hiện là Phó Giáo sư, Giảng viên cao cấp tại Khoa Công nghệ Thông tin, Trường Đại học Khoa học, Đại học Huế.

Lĩnh vực nghiên cứu: Công nghệ phần mềm, Cơ sở dữ liệu, Xử lý song song và phân tán, tính toán lưới.

