XÂY DỰNG THƯ VIỆN HIỆN ĐẠI BẰNG DỊCH VỤ WEB & XML

ThS. ĐOÀN HỒNG NGHĨA
Integrated e-Solutions Ltd.

"Chúng ta cần các công cụ để mở rộng và các chuẩn hệ thống để kết nối chuyển đổi thông tin giữa các thư viện lưu trữ"
Margaret Hedstrom (Giáo sư Đại Học Wisconsin, Chủ Tịch New York State Forum on Information Resources Management)

Trong bài viết này chúng tôi xin giới thiệu với các bạn những kinh nghiệm về việc sử dụng các dịch vụ Web và XML để xây dựng một thư viện số quản lý thông tin thực phẩm và hàng nông nghiệp đã được triển khai tại Phần Lan. Bài viết chú trọng đến việc xây dựng cấu trúc chuyển đổi thông tin XML Information Bus (XIB) nhằm hỗ trợ cho việc khai thác các dữ liệu từ các nguồn lưu trữ (information sources) dưới các dạng khác nhau, thuộc nhiều ngôn ngữ khác nhau. Việc đảm bảo tính độc lập của nguồn dữ liệu để dễ dàng cho thư viện "tiến hoá" khi thêm/bớt các nguồn dữ liệu. Ngoài ra, các dịch vụ Web và chuẩn đóng gói XML nâng cao tính độc lập của hệ thống, cho phép sử dụng trên các nền phần cứng và phần mềm khác nhau.

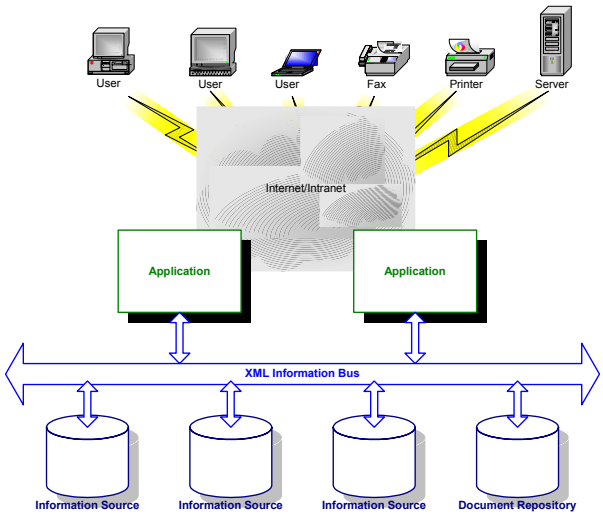
Giới thiệu

Sự phát triển của tính toán phân tán trên mạng (distributed network computing) đã cung cấp các nền tảng công nghệ cơ bản cho việc truy cập dữ liệu và ứng dụng từ xa. Sự phát triển đồng thời và đi sâu của các hệ thống khác nhau đã làm tăng tính hữu ích của các hệ thống này, tuy nhiên không giải quyết được vấn đề thao tác chuyển đổi (interoperability) giữa các ứng dụng trên các hệ thống này. Các ứng dụng được xây dựng không nhằm mục đích kết nối chuyển đổi với nhau, vì thế chúng định nghĩa các dạng dữ liệu khác nhau, sử dụng các giao thức trao đổi (communication protocol) khác nhau và được phát triển trên các nền (platform) khác nhau. Việc thao tác chuyển đổi vẫn là vấn đề lớn trong tính toán phân tán trên mạng.

Ngày nay, việc cho phép thao tác chuyển đổi giữa các tài nguyên thông tin khác nhau về dạng và nội dung là một trong những vấn đề then chốt của các cộng đồng và công ty lớn. Người sử

dụng và ứng dụng (application) có những nhu cầu ngày càng tăng về việc truy cập và thao tác trên các dữ liệu từ một số lượng lớn và đa dạng của các tài nguyên thông tin. Tuy nhiên các tài nguyên thông tin này được tạo ra và quản trị hoàn toàn độc lập, về mặt vật lý, nguyên tắc và phương thức. Các vấn đề nảy sinh do liên kết với những môi trường đó bao gồm tính không đồng nhất và tự quản của hệ cơ sở dữ liệu, sự mâu thuẫn trong phương thức nhận dạng và phân tích (identification and resolution), ngữ nghĩa biểu diễn của dữ liệu, việc xác định vị trí và cách xác định cơ bản thông tin quan trọng, cách truy cập và tính thống nhất của dữ liệu từ xa, các phương pháp xử lý truy vấn (query processing) và quan trọng nhất: việc tiến hoá có kế thừa của hệ thống.

Một trong các ví dụ trực quan cho các hệ thống thư viện là Thư Viện Quản Lý Thực Phẩm và Hàng Nông Nghiệp (Fin FAO Library – FFL) của chính phủ Phần Lan. Thư viện này hỗ trợ việc hiện đại hoá và mở rộng các ngành nông, lâm và ngư nghiệp, đảm bảo việc cung cấp lương thực đúng tiêu chuẩn chất lượng. Việc thu thập, phân tích và phổ biến thông tin là một chức năng quan trọng nhằm hỗ trợ chính phủ trong việc đảm bảo nguồn cung cấp thức ăn đầy đủ, đa dạng và an toàn. Một số lượng lớn các loại hình thông tin khác nhau được tạo mới và cập nhật hằng ngày và lưu trữ tại các nguồn dữ liệu hoàn toàn độc lập. Tuy vậy không hề có các chuẩn nội dung, ngôn ngữ, dạng dữ liệu để định nghĩa các thông tin này. Người sử dụng có nhu cầu truy cập và thao tác trên thông tin được lưu trữ phân tán trong các nguồn khác nhau trong và ngoài chính phủ. Tiêu chí quan trọng của hệ thống là việc sử dụng chung (share) các thông tin nhanh chóng và tiện lợi mà không cần xây dựng lại các hệ thống sẵn có một cách quá phụ thuộc lẫn nhau. Nói khác đi, điều thiết yếu nhất là các hệ thống sẵn có cần tìm ra một ngôn ngữ và giao thức để dễ dàng trao đổi thông tin. Bài viết này sẽ đề cập đến XML [2] (ngôn ngữ) và các dịch vụ Web [1] (giao thức) nhằm phục vụ cho việc liên kết các nguồn dữ liệu độc lập.



Đây là một giải pháp đòi hỏi chi phí thấp, dựa trên công nghệ XIB cho phép trao đổi các thông tin giữa các nguồn khác nhau bằng các kỹ thuật khác nhau. XIB truy cập bằng một phương pháp thống nhất các thông tin lưu trữ trong các dạng dữ liệu khác nhau, lưu trữ trong các ngôn ngữ khác nhau. Việc truy cập này được hỗ trợ bằng các siêu dữ liệu (metadata) như các mô hình mẫu sử dụng trong chuyển đổi dữ liệu. XIB hỗ trợ việc tạo lập các báo cáo thống kê và phương thức theo dõi hoàn toàn trên các dữ liệu động, cho phép người quản lý nắm rõ tài nguyên thông tin thư viện đang cung cấp vào thời điểm hiện tại.

Các vấn đề đặt ra

FFL có hơn 200 nguồn thông tin khác nhau: các website, cơ sở dữ liệu và các kênh thông tin qua giao thức riêng của chính phủ, doanh nghiệp và các tổ chức phi lợi nhuận (non-profit organization). Việc nối kết và trao đổi dữ liệu bao gồm

- ⇨ Việc chuẩn hoá dữ liệu chung,
- ⇨ Kết nối với các website đã được xây dựng qua các công nghệ: HTML, Microsoft ASP và Java Servlet/JSP
- ⇨ Kết nối với các cơ sở dữ liệu SQL: Oracle 8.1.6, Microsoft SQL Server 2000, IBM DB2, PostgresQL, …
- ⇨ Kết nối với các ứng dụng sử dụng giao thức (interface) riêng biệt như MARC dành cho các thư viện sẵn có

Cơ sở hạ tầng của các hệ thống có sẵn bao gồm các nguồn thông tin lưu trữ trong các cơ sở dữ liệu khác nhau, và sử dụng 5 loại ngôn ngữ: Phần Lan, Anh, Pháp, Nga và Đức. Các dữ liệu này được lưu trữ trong các dạng dữ liệu khác nhau, các văn bản khác nhau về cấu trúc và ngoài ra còn có các yếu tố tham chiếu (reference), dữ liệu thống kê, bản đồ và hình ảnh, tin mới, sự kiện từ các ngành nghề và mảng kinh doanh khác nhau, …

Người sử dụng hệ thống bao gồm từ các nhà nghiên cứu, doanh nghiệp tư nhân, bộ phận lập kế hoạch của chính phủ và các thành phần khác. Người sử dụng đa phần dùng các website sẵn có và các ứng dụng trong đơn vị, tổ chức để truy cập và thao tác (hạn chế) trên các dữ liệu. Quá trình sử dụng khá phức tạp vì cần tổng hợp thông tin từ một số lượng lớn các nguồn khác nhau và truy cập rất nhiều các website, cơ sở dữ liệu trực tuyến khác nhau. Một phần lớn thời gian của người sử dụng dành cho việc tìm thông tin theo các liên kết (link) sẵn có, nhưng cần thao tác thủ công để truy cập đến thông tin cần thiết, cũng như thao tác hoàn toàn thủ công (cut-and-paste) để chuyển được thông tin từ các trang này sang ứng dụng của mình.

Các hệ thống sẵn có cung cấp một số tính năng mở nhất định nhằm liên kết với các ứng dụng bên ngoài, nhưng các hạn chế quá lớn do không đủ kinh phí xây dựng, các vấn đề kỹ thuật phức tạp vượt quá khả năng giải quyết, thiếu tính linh động, chuẩn hoá, khả năng cung cấp dịch vụ cho một số lượng lớn người sử dụng cùng lúc, thiếu tính mở rộng và các yếu tố khác làm cho việc kết nối chuyển đổi dữ liệu không thể thống nhất và quá tốn kém khi xây dựng lại.

Điều quan trọng ở đây là vấn đề công nghệ nào có thể đáp ứng được các yêu cầu sau:

- ⇨ Chi phí thấp,
- ⇨ Dễ dàng triển khai (implement),
- ⇨ Dễ dàng quản trị,
- ⇨ Sử dụng các chuẩn (standard) sẵn có,
- ⇨ Sử dụng đòn bẩy trên điểm tựa của sự am hiểu và các tài nguyên sẵn có mà không cần tạo mới lại toàn bộ các hệ thống

Các công nghệ cần thiết này cần phải đáp ứng khả năng thao tác chuyển đổi giữa các nguồn dữ liệu sẵn có trên cơ sở các biến thể khác nhau về ngôn ngữ và cấu trúc dữ liệu mà không đòi hỏi thay

đổi các cơ sở dữ liệu hay các giao thức sẵn có. Một trong các vấn đề hiện có là việc thay đổi cơ sở dữ liệu khi cần hỗ trợ ngôn ngữ mới hoặc có dạng dữ liệu mới (hình ảnh, âm thanh, phim, video, …) Tính phức tạp của các dạng dữ liệu sẵn có không thể đồng nhất hoá các cơ sở dữ liệu và không thể thay đổi xuyên suốt các hệ thống khi một hệ thống có nhu cầu đổi mới cấu trúc dữ liệu hay giao thức đối với người sử dụng.

Một trong những vấn đề khác là việc cung cấp dạng thông tin và giao thức thống nhất nhằm hỗ trợ người sử dụng hệ thống mới dễ dàng thu thập dữ liệu, phổ biến thông tin trong thời gian ngắn nhất. Các nhu cầu mới về việc tra cứu nhanh, tìm xuyên suốt các hệ cơ sở dữ liệu, xây dựng từ điển liệt kê ngữ nghĩa (thesaurus) trực tuyến và liên kết đến tận sản phẩm và đơn vị sản xuất / xuất-nhập khẩu là có thực.

Giải pháp

Phương pháp tiếp cận để giải quyết bài toán nêu trên là việc đặt ra hướng giải quyết dựa trên

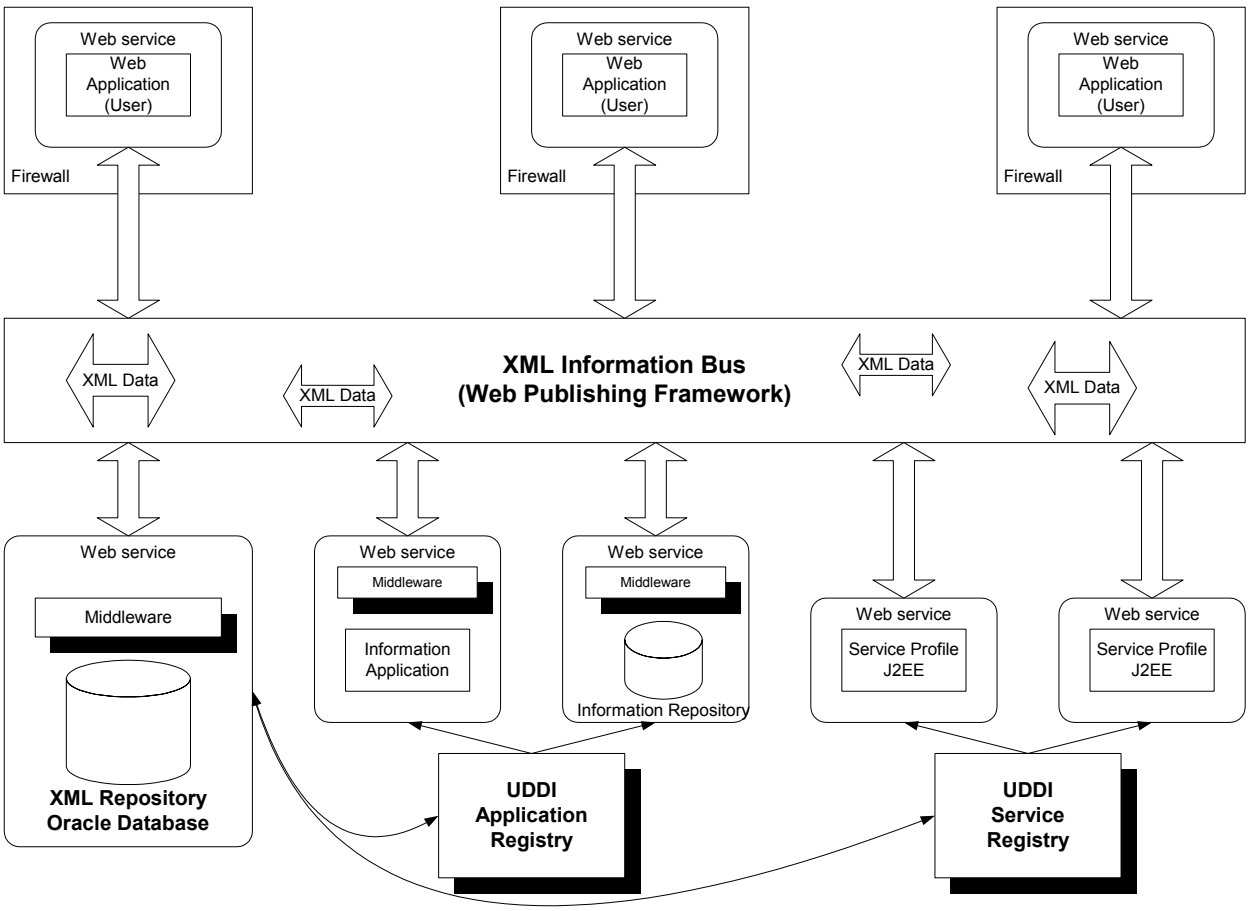
- ⇨ Việc nối kết và chuyển đổi dữ liệu với nhiều hệ thống của các nhà cung cấp khác nhau (về kỹ thuật giao thức và dạng nội dung dữ liệu).
- ⇨ Mục tiêu thứ hai trong quá trình tiếp cận là hạn chế tối đa việc thay đổi giao thức, cơ sở dữ liệu và cách hoạt động các hệ thống sẵn có.
- ⇨ Mục tiêu thứ ba là đảm bảo tính liên tục của các dịch vụ được cung cấp hiện nay

Giải pháp bao gồm:

1. Phát triển một hệ thống dịch vụ Web nhanh và dễ dàng
2. Kết nối với hơn 200 nguồn thông tin hiện có
3. Hỗ trợ dạng ngôn ngữ và các cấu trúc dữ liệu khác nhau
4. Hỗ trợ việc thống kê và theo dõi các thông tin động trên toàn bộ các nguồn
5. Phát triển một tập hợp các siêu dữ liệu XML nhằm phục vụ cho việc thu thập, cung cấp thông tin tự động với các ứng dụng bên ngoài khác

XML Information Bus

Giảp pháp được xây dựng trên nền tảng XML Information Bus, nhằm liên kết các nguồn thông tin dạng khác nhau vào một chuẩn đóng gói dữ liệu duy nhất. Giao thức trao đổi (interface) tại các nguồn thông tin có thể khác nhau tùy theo yếu tố của nguồn thông tin địa phương. Các thông tin được đóng gói thành các dữ liệu có cấu trúc XML chặt chẽ. Việc đảm bảo giao thức địa phương của XIB và các nguồn thông tin địa phương tuân theo giao thức cung cấp thông tin của nguồn thông tin đó.



Cấu Trúc Hệ Thống và XML Information Bus

Ý tưởng chính của XIB là việc tất cả các dữ liệu trao đổi qua kênh thông tin đều có cấu trúc XML. Các cấu trúc này tuân theo các giản đồ XML (XML Schema). Các giản đồ này được sử dụng để tạo ra dữ liệu, mô tả cấu trúc dữ liệu khi luân chuyển, và kiểm tra cấu trúc dữ liệu và tính đúng đắn của các thành phần (về hình dạng và nội dung) dữ liệu khi xử lý. Hoàn toàn không phụ thuộc vào dạng dữ liệu tại các nguồn thông tin đã có, XIB sử dụng dạng dữ liệu XML chuẩn cho mọi truy xuất dữ liệu tại các đầu ra vào của hệ thống đối với người sử dụng và thành phần bên trong của hệ thống. Ví dụ như tất cả các dữ liệu về đất nước, tiền tệ và ngôn ngữ đều tuân theo dạng XML duy nhất với các chuẩn ISO-3166 (3 ký tự) cho quốc gia, ISO-639-1 (2 ký tự) cho ngôn ngữ và ISO-4217 dành cho tiền tệ. Mặc dù việc sử dụng các chuẩn này là không bắt buộc đối với các hệ thống đang được sử dụng, nhưng đây là điều bắt buộc với các giao thức của dịch vụ Web và các tầng phần mềm giữa (middleware) để đảm bảo tính thống nhất tại các giao thức và giao diện chuyển đổi thông tin, từ đó đảm bảo tính thống nhất về hình thức và nội dung dữ liệu thông tin bên trong XIB.

```
<soap:Envelope
 xmlns:xsi="http://www.w3.org/2001/
                          XMLSchema-instance"
 xmlns:xsd="http://www.w3.org/2001/
                          XMLSchema"
 xmlns:soap="http://schemas.xmlsoap.org/soap/
                          envelope/">
 <soap:Body>
  <Query xmlns="http://tempuri.org/">
    <Country>SEN</Country>
    <Language>EN</Language>
   <Keyword>Forestry</Keyword>
  </Query>
 </soap:Body>
</soap:Envelope>
```

Ví dụ thông tin XML được chuyển đổi trong XML Information Bus

Việc quản lý thông tin, bao gồm cả thông tin trên các ngôn ngữ khác nhau, cũng dựa trên nền tảng XML. Các cấu trúc thông tin dữ liệu trong các bảng và trường của cơ sở dữ liệu (database table & field) cần được mô tả lại bằng các cấu trúc XML thích hợp, các liên kết trực tiếp (direct), gián tiếp (indirect) hoặc tham chiếu (reference) của các dữ liệu bên trong đều có thể được mô tả qua các phần tử con (sub-element), thuộc tính (attribute), hoặc các reference qua XLink. Vì thế ngôn ngữ XML có thể mô tả rõ ràng và chính xác bất kỳ cơ sở dữ liệu SQL nào. Việc chuyển đổi các thông tin từ HTML/XHTML đều có thể khai thác thông qua XML bằng việc định nghĩa các thành phần quan trọng của trang và các trích dẫn các nội dung và thiết lập các ánh xạ giữa các thành phần này và các phần tử của trang XML. Việc chuyển đổi có thể hoàn toàn được tự động hoá thông qua XML StyleSheet Language (XSL). Toàn bộ công việc này nhằm xây dựng một ngôn ngữ mô tả bằng XML nhằm phục vụ các ứng dụng hiện có trao đổi thông tin với nhau và với người sử dụng qua cầu nối XIB. XIB đóng nhiệm vụ người biên dịch hai chiều cho bất kỳ hai thành phần nào sử dụng XIB. Việc liên kết các hệ thống đang tồn tại vào XIB chỉ đòi hỏi việc thiết lập "ngữ pháp" để trao đổi giữa XIB và hệ thống đó.

```
<BBCNewsDS
 xmlns ="http://www.fao.org/waicent/cpmis/
                          BBCNewsDS.xsd">
 <BBCNews>
   <headline>Blair blasts green pacesetters
   </headline>
   <intro>In 1997 Labour undertook to be the
&amp;#34;first truly green government&amp;#34;,
but has that promise been fulfilled?</intro>
   <newsdate>23/10/2000</newsdate>
   <link>http://news.bbc.co.uk/hi/english/sci/tech/
        newsid_987000/987400.stm</link>
 </BBCNews>
 <BBCNews>
   <headline>Labour: A green government?
   </headline>
   <intro>In 1997 Labour undertook to be the
&amp;#34;first truly green government&amp;#34;,
but has that promise been fulfilled?</intro>
   <newsdate>23/10/2000</newsdate>
   <link>http://news.bbc.co.uk/hi/english/sci/tech/
        newsid_986000/986532.stm</link>
 </BBCNews>
</BBCNewsDS>
```

Ví dụ thông tin trả về từ hệ thống đang sử dụng sau khi đã kết nối với XIB

Đi sâu hơn nữa vào cách chuyển đổi của các thông tin thư viện được xây dựng trên chuẩn MARC, các hỗ trợ sẵn có hiện nay bao gồm RDF [3], RDF Schema [4], Dublin Core elements version 1.1 [5] và XML Topic Maps [6]. RDF được sử dụng để mô tả các siêu dữ liệu dành cho tài nguyên, ví dụ như giá trị của các đặc tính riêng của các miền. RDF Schema dành cho việc định nghĩa các lớp tài nguyên và các đặc tính phụ thuộc mà các dữ liệu cụ thể có thể sử dụng. Ngoài ra việc sử dụng đồng thời RDF Schema, Dublin Core và XML Topic Maps có thể định nghĩa được các bản thể học (ontopology) của các quan hệ giữa các lớp, tài nguyên và đặc tính để tạo nên một bảng từ vựng (vocabulary). Áp dụng XML Schema,

bảng từ vựng này định nghĩa các giá trị có thể có của các đặc tính mà các tài nguyên sử dụng. Các giá trị có thể có và giới hạn của các trường thường được định nghĩa bên trong các hệ thống đang tồn tại và không hề hiển hiện cho người sử dụng. Với XML Schema, các giá trị này được kiểm tra và quản lý ngay tại các giao diện dịch vụ Web, vì thế giảm thiểu khả năng nhập/xuất dữ liệu sai và tăng tính ổn định và an toàn của hệ thống.

```
<FSCollectionChoices
xmlns="http://tempuri.org/CollectionChoices.xsd">
 <Country diffgr:id="Country231"
          msdata:rowOrder="230">
   <COUNTRY>Zimbabwe</COUNTRY>
   <FS_COUNTRYCODE>181
          </FS_COUNTRYCODE>
   <ISOCODE>ZWE</ISOCODE>
 </Country>
 <Item diffgr:id="Item1"    msdata:rowOrder="0">
   <ITEM>Abaca (Manila Hemp)</ITEM>
   <FS_ITEMCODE>809</FS_ITEMCODE>
 </Item>
 <Element diffgr:id="Element1"
          msdata:rowOrder="0">
   <ELEMENT>Seed</ELEMENT>
   <FS_ELEMENTCODE>111
     </FS_ELEMENTCODE>
 </Element>
 <Year diffgr:id="Year1"   msdata:rowOrder="0"
       diffgr:hasChanges="inserted">
   <YEAR>1961</YEAR>
 </Year>
</FSCollectionChoices>
```

Ví dụ thông tin trả về từ hệ thống hỗ trợ của XIB

Để truy cập trực tiếp các dữ liệu dạng văn bản sẵn có và chuyển đổi sang cấu trúc XML và ngược lại, XML:DB API [7] là chuẩn giao thức cho phép truy cập các văn bản này và các siêu dữ liệu đi kèm với các văn bản. Cùng với dịch vụ Web sử dụng để kết nối vào XIBđ, XML:DB API cho phép việc sử dụng các văn bản đang tồn tại như sử dụng các cơ sở dữ liệu thông thường: tìm kiếm, tra cứu, trích lục, sao chép một phần và chuyển đổi thông tin cho các hệ thống khác.

Nhằm theo dõi xuyên suốt các dịch vụ và nội dung dịch vụ trong hệ thống, các mô tả sơ lược dịch vụ (service profile) được lưu trữ trong hệ thống nhằm mô tả các khả năng, tính năng, phương thức giao tiếp, cấu trúc dữ liệu vào/ra của mỗi dịch vụ Web và tài nguyên thông tin mà dịch vụ đó cung cấp.

```
<ServiceDetails
 xmlns:xsi="http://www.w3.org/2001/
                        XMLSchema-instance"
 xmlns="http://tempuri.org/ServiceDetails.xsd">
 <ServiceDetail
     d2p1:ServiceName="GeneralMaps"
     d2p1:ServiceID="900"
xmlns:d2p1="http://tempuri.org/ServiceDetails.xsd">
   <ServiceDescription>
     Description for GeneralMaps
   </ServiceDescription>
   <Param d2p1:name="Country">
     <value>GBR</value>
   </Param>
   <Param d2p1:name="Language">
     <value>EN</value>
   </Param>
   <Param d2p1:name="Category">
     <value>16</value>
     <value>19</value>
   </Param>
 </ServiceDetail>
</ServiceDetails>
```

Ví dụ mô tả sơ lược của một dịch vụ Web

Điều hành hệ thống, khám phá thông tin mới, dịch vụ mới, cũng như việc dự liệu (provision) cung cấp dịch vụ cho người sử dụng khi có dịch vụ mới, thông tin mới là hai hệ khám phá tổng quát, mô tả và giao diện (Universal Discovery, Description and Interface).

Kết luận

Sử dụng công nghệ phù hợp là điều quan trọng nhất trong việc kết nối và đưa vào sử dụng ngay trong thời gian ngắn nhất các dịch vụ sẵn có. Sử dụng đúng công cụ khi tạo mới các hệ cung cấp thông tin nhằm phục vụ tính mở để dễ dàng nâng cấp, kết nối đi đến việc xây dựng một mạng lưới tài nguyên thông tin mang tính kế thừa và phát triển nhanh. Việc thiết kế tạo mới hay kết nối các nguồn thông tin hiện có cần tuân theo các tiêu chí này. Các công cụ hiện nay hoàn toàn có khả năng tạo lập mạng lưới nguồn thông tin tài nguyên giàu có, dễ sử dụng chỉ trong khoảng thời gian ngắn và chi phí thấp. Bài viết này giới thiệu một trong các giải pháp đã được minh chứng trong thực tế và có thể được áp dụng vào tình hình các thư viện chúng ta hiện nay.

TÀI LIỆU THAM KHẢO

1. Graham,S.,Simeonov,S.,Boubez,T.,Davis,D., Daniels,G.,Nakamura,Y.and Neyama,R.,2002. Building Web Services with Java: Making Sense of XML,SOAP,WSDL,and UDDI.SAMS Publishing, 2002.
2. Bray,T.,Paoli,J.,Sperberg-McQueen,C.M.and Maler,E.,2000.Extensible Markup Language (XML) 1.0,Second Edition,W3C Recommendation,October 2000. http://www.w3.org/TR/2000/REC-xml-20001006.
3. Lassila,O.and Swick,R.R.,1999.Resource Description Framework (RDF) Model and Syntax Specification.February,1999. http://www.w3.org/TR/REC-rdf-syntax
4. Brickley,D.and Guha,R.V.,2002.Resource Description Framework (RDF)Schema Specification 1.0,March 2002 http://www.w3.org/TR/rdf-schema
5. Dublin Core..http://dublincore.org
6. Pepper,S.and Moore,G.XML Topic Maps (XTM)1.0 http://www.topicmaps.org/xtm/1.0
7. XML::DB http://www.xmldb.org