# XÂY DỰNG KHO DỮ LIỆU MỞ CHO TRAO ĐỔI THÔNG TIN – THƯ VIỆN

**TS. Hoàng Lê Minh**
Trung Tâm Công nghệ Phần mềm ĐHQG TP. HCM

*Tóm tắt*

*Xây dựng các kho dữ liệu thông tin – thư viện theo hướng mở, chia sẻ tối đa các tài nguyên trên mạng Internet là một trong các nhiệm vụ phát triển ứng dụng CNTT. Trung tâm Công Nghệ Phần Mềm ĐHQG-HCM (UniSoft) trong thời gian qua đã triển khai thành công một số giải pháp và phần mềm tin học hoá trên nền các công nghệ nguồn mở trong lĩnh vực quản lý thông tin – thư viện. Bài viết này nhằm nhấn mạnh tầm quan trọng của việc ứng dụng các giải pháp công nghệ nguồn mở và các chuẩn mở trong xây dựng các kho dữ liệu mở để quản lý và chia sẻ các thông tin trên mạng Internet. Chúng tôi cố gắng trả lời các câu hỏi thường gặp sau của độc giả:*

1. *Tại sao nên xây dựng các **kho dữ liệu mở,** mô tả bằng các thẻ **metadata** ?*
2. *Chuẩn **Dublin Core Metadata** (Z.39.85-2001) là gì ?*
3. *Quản lý và truy xuất các kho dữ liệu mở trên mạng Internet như thế nào ?*

## I. Thông tin và các kho dữ liệu thông tin.

Trong thế giới hiện đại, mọi hoạt động của con người đều cần đến thông tin

1. Thông tin về giá cả và thị trường, khách hàng và đối tác, v.v... trong hoạt động sản xuất, kinh doanh, thương mại, tài chính.
2. Thông tin về chính sách và pháp luật, thuế và tiền tệ, dân số và lao động, đất đai và môi trường trong quản lý kinh tế, quản lý nhà nước.
3. Thông tin về văn hoá, khoa học, giáo dục, đặc biệt trong lĩnh vực hoạt động Thông tin - Thư viện.

Để có được các thông tin cần thiết và cập nhật về nội dung, phải xây dựng các kho dữ liệu. Đó là quy trình phức tạp và tốn nhiều chi phí, bao gồm nhiều công đoạn khác nhau từ thu thập, xử lý, phân tích, tổ chức lưu trữ thông tin theo một số tiêu chuẩn nào đó. Giải pháp CNTT thường được ứng dụng là sử dụng các hệ *quản trị cơ sở dữ liệu* và các phần mềm giao tiếp để quản lý và truy tìm các thông tin cần thiết một cách nhanh chóng.

Trong kế hoạch tổng thể trước đây và hiện nay về phát triển CNTT, ở trung ương và các địa phương đều chú trọng xây dựng các kho dữ liệu, thể hiện qua những dự án xây dựng, quản lý và khai thác các thông tin, các Trung tâm tích hợp dữ liệu. Riêng trong lĩnh vực thông tin – thư viện, đã có khá nhiều dự án đầu tư của nhà nước và nguồn vốn vay hay tài trợ của nước ngoài để xây dựng các Thư viện điện tử số hoá.

Trong thời đại cách mạng công nghệ thông tin, các kho thông tin cần được liên kết với nhau để giúp chia sẻ tài nguyên, cung cấp dịch vụ. Tuy nhiên không phải hệ quản trị CSDL nào cũng có thể đáp ứng tốt nhu cầu tích hợp, chia sẻ thông tin từ các nguồn phân tán, đôi khi không thật chuẩn mực! Lựa chọn giải pháp khả thi để trao đổi, liên thông các cơ sở dữ liệu là điều rất khó thực hiện.

Trên thực tế đã có hai xu hướng giải quyết vấn đề như sau:

**1.** Thống nhất sử dụng chung **một** phần mềm hay **một** cơ sở dữ liệu.

**2.** Xây dựng Trung tâm **tích hợp dữ liệu** để thống nhất dữ liệu từ nhiều nguồn.

Giải pháp *dùng chung* phần mềm và cơ sở dữ liệu tuy có làm giảm được chi phí phải chuyển đổi và nhập lại dữ liệu, nhưng rất khó thực hiện ở quy mô rộng, tại nhiều địa điểm khác nhau, bởi vì nó đòi hỏi môi trường và trình độ người sử dụng phải tương đối thống nhất. Mặt khác khó có thể đầu tư và triển khai trên diện rộng một phần mềm dùng chung có chất lượng cao.

Xây dựng các trung tâm tích hợp thông tin dữ liệu cũng đòi hỏi thời gian và chi phí khá lớn, và trên thực tế cũng khó lòng giải quyết được các vấn đề nêu trên. Thực tế áp dụng CNTT của ngành GD-ĐT trong tuyển sinh đại học năm 2002 cho thấy giải pháp tích hợp dữ liệu từ các trường và các địa phương để có được kết quả xét tuyển chung đã không thành công.

Vấn đề đặt ra ở đây là làm sao để các nguồn vốn đầu tư xây dựng các kho dữ liệu dùng chung phải phát huy hiệu quả thực sự. Để làm được, chỉ có một cách duy nhất là *phải đưa thông tin đến tận tay người sử dụng* (thông qua mạng Internet), đồng thời mở rộng khả năng *cung cấp và cập nhật thông tin trực tiếp từ nguồn*, không chờ đợi thông tin chuyển đến các trung tâm tích hợp rồi mới xử lý.

Xuất phát từ thực tế trên, UniSoft đề xuất mô hình có thể giúp xây dựng các kho dữ liệu dùng chung một cách đơn giản và nhanh chóng như sau:

3. **Thống nhất chuẩn** mô tả dữ liệu bằng ngôn ngữ **XML**, dựa trên các thẻ **metadata**; sử dụng mạng **Internet/Intranet** để xây dựng các **kho dữ liệu mở**; xây dựng phần mềm quản lý bằng công nghệ nguồn mở và tra cứu dữ liệu trên cơ sở các máy tìm kiếm thông tin (search engine).

Các **kho dữ liệu mở** là tập hợp các tập tin được gắn các thẻ mô tả metadata, được lưu trữ trong các thư mục và có thể truy xuất toàn văn bản (full-text) từ các địa chỉ website. Đặc điểm khác biệt của hệ thống các dữ liệu này so với các trang web thông thường là ở chỗ:

a. Mỗi dữ liệu được kèm theo các thẻ mô tả metadata; phần mô tả này được đặt ở phần đầu (header) của tập tin dữ liệu,

b. Phần mềm quản trị sẽ đọc các thông tin metadata trong phần header và trình bày kết quả (ngắn gọn hay đầy đủ) dưới dạng các trang web.

c. Cơ chế tìm kiếm thông tin được xây dựng trên nguyên tắc hoạt động của các search engine, nghĩa là ta thực hiện tìm kiếm theo thông tin metadata và full-text trước để tạo ra các bảng chỉ mục (index), giúp người sử dụng tìm kiếm thông tin một cách nhanh chóng.

Với giải pháp công nghệ như trên, chúng tôi tin tưởng rằng có thể nhanh chóng hình thành được những kho dữ liệu rất lớn với chi phí thấp, không bị các hạn chế như trường hợp sử dụng các cơ sở dữ liệu đóng. Đồng thời việc thêm và cập nhật dữ liệu có thể thực hiện rất đơn giản, bằng nhiều công cụ soạn thảo trang web và trình bày dữ liệu quen thuộc như MS FrontPage, MS Word, v.v… Thành phần quan trọng nhất của hệ thống là phần mềm trình diễn dữ liệu và phần mềm tìm kiếm thông tin đều được thiết kế để chạy trên máy chủ và trên mạng Internet.

Dưới đây chúng ta sẽ tìm hiểu kỹ hơn quá trình mô tả tài liệu (bao gồm cả tài liệu dạng toàn văn bản) theo chuẩn ISO Z39.85-2001 Dublin Core Metadata cho các thư viện điện tử số hoá trên mạng Internet.

## II. Ngôn ngữ đánh dấu mở rộng XML và các thẻ Dublin Core Metadata.

Ngôn ngữ đánh dấu mở rộng (XML) là giải pháp thích hợp cho vấn đề trao đổi dữ liệu tự động giữa các kho thông tin trên mạng Internet.

Bản thân ngôn ngữ XML (eXtensible Markup Language) có nguồn gốc giống như ngôn ngữ định dạng siêu văn bản HTML, từ chuẩn ngôn ngữ định dạng văn bản tổng quát có cấu trúc SGML (Structured General

Markup Language). Mỗi văn bản XML bao gồm các thẻ (tag) với tên gọi của các phần tử (element). Nhưng khác với ngôn ngữ HTML, số lượng và tên gọi các phần tử trong XML là không hạn chế.

Các văn bản XML với thông tin mô tả kèm theo, gọi là các thẻ metadata, có thể dễ dàng chuyển giao trên mạng Internet, tương tự như các tập tin HTML. Đặc biệt các văn bản này có thể được các phần mềm đọc (các bộ XML Parser) xử lý tự động để trình diễn trên web. Các phần mềm này cũng có vai trò tương tự như các trình duyệt thông tin, chỉ khác là các bộ XML Parser cho phép thực hiện quá trình tự động xử lý nhập/xuất dữ liệu thay cho con người. Tuy ý tưởng về một ngôn ngữ văn bản giúp tự động hoá quá trình nhập/xuất/trao đổi dữ liệu giữa các kho dữ liệu như XML là hết sức đơn giản, nhưng hiệu quả mà XML đem lại rất to lớn, giúp chúng ta có thể giải quyết những vấn đề còn vướng mắc trong quá trình trao đổi thông tin trên mạng Internet.

Chuẩn trao đổi các dữ liệu thông tin – thư viện trên Internet hiện nay đã được tổ chức tiêu chuẩn quốc gia của Mỹ thông qua nhằm thay thế cho các chuẩn cũ không còn phù hợp (như Z39.50) là chuẩn ANSI/NISO Z.39.85-2001. Nội dung chủ yếu của chuẩn mô tả dữ liệu này gồm 15 trường dữ liệu, còn gọi là Dublin Core Metadata. Đây là các trường dữ liệu phổ biến và hữu ích nhất kèm theo một tài liệu số hoá để trao đổi trên mạng Internet.

Thư viện Quốc gia Hoa Kỳ về Y học (National Library of Medicine) bắt đầu từ năm 2001 đã hoàn toàn thực hiện trao đổi thông tin và các dữ liệu bằng ngôn ngữ XML, sau khi chuyển đổi hơn 10 triệu bản ghi MARC sang dạng XML. Các thư viện của Pháp và Nhật bản cũng đang bắt đầu chuyển qua sử dụng ngôn ngữ XML cho công tác biên mục và trao đổi thông tin, v.v….

**Metadata là gì ?**

Metadata dùng để mô tả một tài nguyên thông tin. Thuật ngữ "meta" xuất xứ là một từ Hy Lạp dùng để chỉ một cái gì đó có bản chất cơ bản hơn hoặc cao hơn. Vì vậy metadata là dữ liệu về những dữ liệu khác. Nó được những thủ thư truyền thống đặt vào trong các biên mục và được sử dụng thông thường nhất là để mô tả thông tin về các tài nguyên Web.

Một bản ghi metadata bao gồm một tập các thuộc tính hoặc tập các phần tử cần thiết để mô tả các tài nguyên theo yêu cầu. Ví dụ, một hệ thống metadata thông thường trong biên mục thư viện chứa một tập các bản ghi metadata dùng để mô tả sách như: tác giả, nhan đề, ngày xuất bản, tiêu đề đề mục, số gọi để chỉ vị trí trên giá sách, v.v….

Mối liên kết giữa các bản ghi metadata và tài nguyên có thể ở một trong hai dạng sau:

- Các phần tử metadata được chứa trong một bản ghi riêng tách rời với tài liệu, ví dụ trường hợp bản ghi biên mục của thư viện truyền thống.
- Metadata có thể được gắn vào trong tài liệu. Ví dụ trong biên mục xuất bản nước ngoài, dữ liệu mô tả này được gắn vào trang sau của trang tiêu đề sách. Đối với các thư viện số hoá, việc gắn các thông tin biên mục metadata vào ngay trong tài liệu toàn văn bản là yêu cầu bắt buộc.

Có nhiều chuẩn mô tả biên mục mang tính chất metadata khá thông dụng, thí dụ MARC21/UNIMARC, ISO 2709, Dublin Core Metadata,… Các dữ liệu này thường được gắn vào phần đầu cho mỗi tài liệu điện tử đặt trên website và rất thích hợp cho các search engine tìm kiếm, lọc ra các thông tin metadata để tổ chức thành các kho dữ liệu mà không cần dùng đến hệ quản trị cơ sở dữ liệu truyền thống. Nói cách khác, ngay bản thân ngôn ngữ XML tự nó đã hỗ trợ cho việc hình thành một cơ sở dữ liệu toàn văn bản, phi cấu trúc rất tiện lợi cho việc tìm kiếm và trao đổi thông tin.

**Dublin Core Metadata là gì ?**

Dublin Core Metadada là chuẩn dùng mô tả nội dung của biểu ghi và dữ liệu. Nó đơn giản hơn MARC Format rất nhiều vì chỉ có 15 phần tử. Mục đích thiết kế metadata này là để sử dụng trên mạng Internet, mô tả các tài liệu chuyên ngành trong các thư viện và tài liệu tại các Web site khác nhau. Các phần tử dữ liệu MARC và Dublin Core có thể trao đổi lẫn nhau theo một giản đồ quy định khi hiển thị cho người dùng. Các phần tử này là:

1. **Nhan đề** - tên gọi chính thức của tài liệu (Title)
2. **Tác giả** **-** tên của một hay một số tác giả chính (Creator)
3. **Đề mục** **-** tên tiêu đề đề mục dùng để phân loại tài liệu (Subject)
4. **Mô tả** **-** mô tả vắn tắt nội dung tài liệu (Description)
5. **Nhà xuất bản** **-** tên gọi, nơi ban hành tài liệu (Publisher)
6. **Tác giả phụ** **-** tên của những tác giả cộng tác (Contributor)
7. **Ngày tháng** **-** ngày tháng ban hành tài liệu (Date)
8. **Loại tài liệu** **-** phân loại tài liệu (Type)
9. **Mô tả vật lý** **-** các thông tin về dạng vật lý (Format)
10. **Định danh** **-** các thông tin định danh của tài liệu (Identifier)
11. **Nguồn gốc** **-** các thông tin về xuất xứ của tài liệu (Source)
12. **Ngôn ngữ** **-** các thông tin về ngôn ngữ (Language)
13. **Liên kết** **-** các thông tin liên hệ của tài liệu (Relation)
14. **Nơi chứa** **-** các thông tin liên quan vị trí lưu trữ (Coverage)
15. **Bản quyền** **-** các thông tin liên quan quyền tác giả (Rights)

Việc sử dụng các phần tử của Dublin Core là tùy chọn và có thể lập lại. Mỗi phần tử cũng có một tập giới hạn các thuộc tính để giới hạn ý nghĩa của nó. Dublin Core Metadata có những đặc tính sau:

- Việc tạo lập và duy trì khá dễ dàng: Cho phép những người không chuyên nghiệp có thể tạo các bản ghi mô tả đơn giản cho các tài nguyên thông tin và việc truy xuất chúng trên môi trường mạng một cách dễ dàng.
- Ngữ nghĩa dễ hiểu và thông dụng: Việc khai thác thông tin trên mạng Internet diện rộng thường gặp trở ngại bởi những sự khác nhau về thuật ngữ và sự mô tả thực tế. Dublin Core Matadata giúp những người dò tìm thông tin không chuyên có thể tìm thấy vấn đề của mình bằng cách hỗ trợ một tập các phần tử thông dụng mà ngữ nghĩa của chúng được hiểu phổ biến. Ví dụ, những nhà khoa học thì liên quan đến các bài báo của tác giả nổi tiếng, những nhà nghiên cứu nghệ thuật thì quan tâm đến các tác phẩm nghệ thuật của nghệ nhân nổi tiếng.
- Phạm vi quốc tế: Tập các phần tử Dublin Core Metadata lúc đầu được phát triển bằng tiếng Anh, nhưng các phiên bản sau của nó được tạo bởi các ngôn ngữ khác nhau. Đến tháng 11 năm 1999 đã có khá nhiều phiên bản với trên 20 ngôn ngữ, chủ yếu là tiếng Phần Lan, tiếng Na Uy, tiếng Thái, tiếng Nhật, tiếng Pháp, tiếng Bồ Đào Nha, tiếng Đức, tiếng Hy Lạp, tiếng In-đô-nê-xi-a và tiếng Tây Ban Nha.
- Tính mở rộng : Những nhà phát triển Dublin Core Metadata đã cung cấp một cơ chế cho việc mở rộng tập các phần tử Dublin Core, phục vụ nhu cầu khai thác các tài nguyên bổ sung. Các phần tử metadata từ những tập các phần tử khác nhau có thể liên kết với Metadata của Dublin Core. Điều này cho phép các tổ chức khác nhau có thể dùng các phần tử Dublin Core để mô tả thông tin thích hợp cho việc sử dụng trên Internet.

**Cú pháp của Dublin Core Metadata:**

HTML hỗ trợ một khuôn dạng dễ hiểu dùng để mô tả các khái niệm cơ bản của Dublin Core nhưng với những ứng dụng phức tạp thì việc sử dụng RDF/XML (Resource Description Framework – Cơ cấu mô tả tài nguyên bằng cách sử dụng XML) có hiệu quả hơn dùng HTML. HTML sử dụng hai nhãn: "<META>" và "<LINK>" để sao chép metadata. Nếu có thêm vào việc tạo metadata thì các nhãn này phải xuất hiện trong phần HEAD của tài liệu HTML. Mỗi phần tử là tùy chọn và có thể lập lại. Các phần tử metadata có thể xuất hiện theo bất kỳ thứ tự nào. Ví dụ:

```
<HTML>
<HEAD>
<TITLE>Tuyên ngôn đảng cộng sản</TITLE>
<META NAME="DC.Creator" CONTENT="Marx, Karl">
<META NAME="DC.Creator" CONTENT="Engels, Friedrich">
<META NAME="DC.Title" CONTENT="Tuyên ngôn đảng công sản" >
<META NAME="DC.Date" CONTENT="1887" >
<META NAME="DC.Type" CONTENT="document">
<META NAME="DC.Format" CONTENT="text/html">
<META NAME="DC.Identifier" CONTENT="http://www.nhandan.org.vn/tndcs.htm">
</HEAD>
<BODY>
<H1>Tuyên ngôn Đảng Công sản</H1>
<P>....</P>
</BODY>
</HTML>
```

**III. Quản lý và truy xuất các kho tài liệu chứa thẻ Metadata**

Với việc áp dụng rộng rãi thẻ Metadata trong mô tả tài liệu, các kho dữ liệu thông tin có thể được hình thành nhanh chóng mà không tốn chi phí nhập và xử lý dữ liệu. Các quá trình quản lý, tìm kiếm thông tin đều có thể thực hiện hoàn toàn tự động dựa vào nguyên lý làm việc của phần lớn các hệ thống thông tin website hiện nay. Thí dụ các phần mềm đọc XML để xử lý và trình diễn đều hiểu rằng các thẻ metadata bắt đầu sau dòng "<HEAD>" và kết thúc trước dòng "</HEAD>". Do đó ta có thể trích lấy metadata một cách tự động trong khi các trình duyệt Web có thể bỏ qua nó. Những máy tìm kiếm thông tin (search engine) hiện nay đều có khả năng đọc và sử dụng các thông tin trong nhãn <META> của HTML trong các tài liệu web. Như vậy quy trình quản lý các kho dữ liệu mở này trở nên hết sức đơn giản:

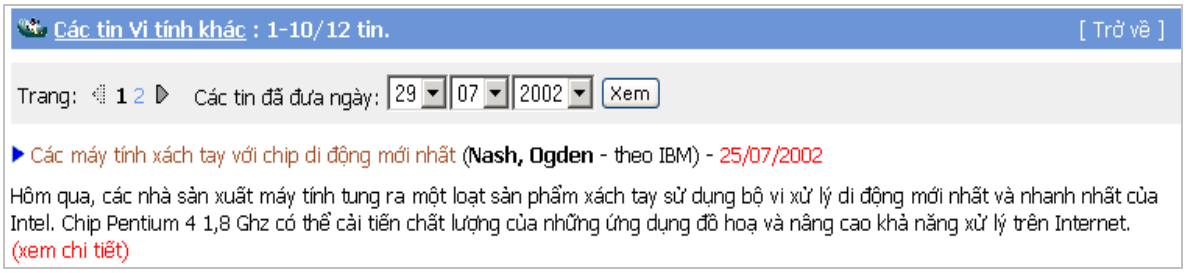
1. **Nhập thêm dữ liệu** mới vào kho chỉ là công đoạn sao chép thêm một tập tin tài liệu có chứa thẻ metadata mô tả từ trước. Quá trình này có thể thực hiện tự động bằng các thao tác tải lên (upload) các tập tin, hay sử dụng e-mail để gửi tài liệu đến kho.
2. **Cập nhật dữ liệu** bằng các sửa lại các thông tin trong các thẻ metadata bằng các công cụ tạo trang web, hoàn toàn không cần dùng đến phần mềm riêng biệt giao tiếp cơ sở dữ liệu.
3. **Xoá / di chuyển dữ liệu** chỉ đơn giản là các thao tác xoá / di chuyển tập tin trên một website hay giữa các website với nhau.

Với các thao tác đơn giản, việc tập hợp dữ liệu để hình thành các kho thông tin, tài liệu hữu ích cho người sử dụng là không phức tạp, hợp với trình độ của đa số người quản trị các hệ thống thông tin - thư viện hiện nay. Khai thác và trao đổi dữ liệu giữa các kho thông tin mở này cũng rất đơn giản, với sự trợ giúp của các phần mềm tìm kiếm quản lý thông tin Internet.

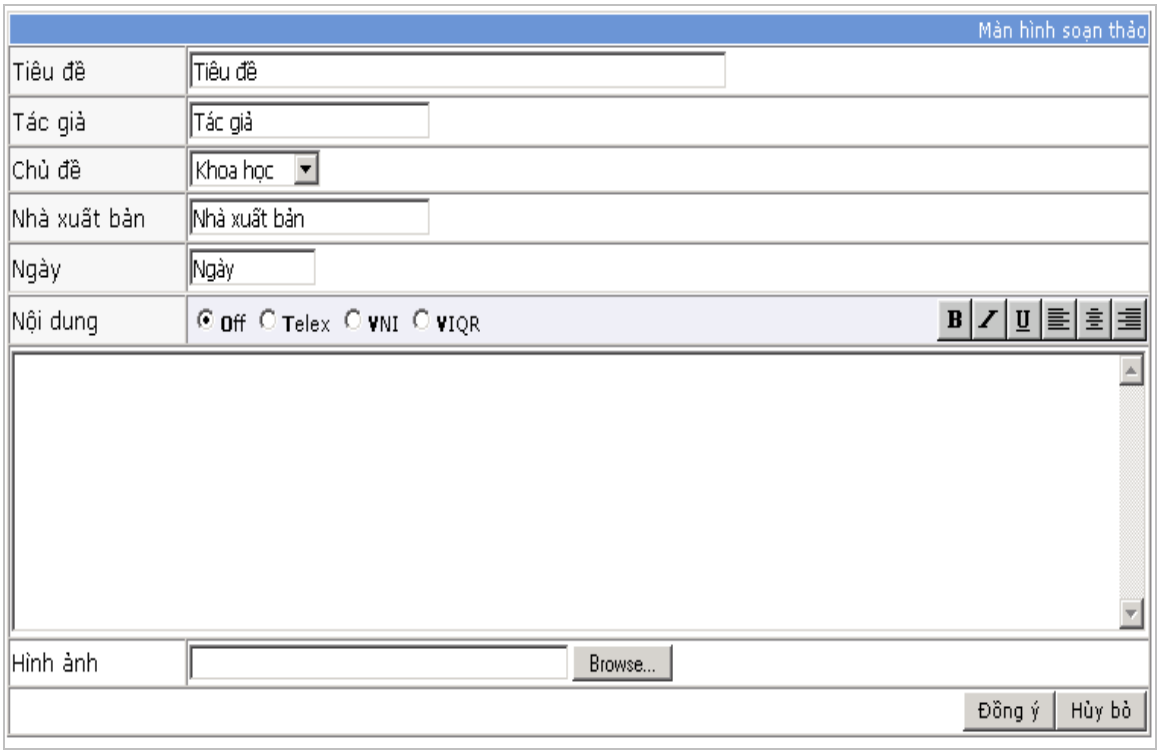
Một trong các kết quả khá thành công của UniSoft theo hướng ứng dụng ngôn ngữ XML, mô tả Dublin Core Metadata để xây dựng các kho thông tin mở là phần mềm quản lý tin tức **iNews** được phát triển trên nền cổng thông tin **iPortal** cho hệ thống thông tin website của ĐHQG TP.HCM. **iNews** bao gồm ba phân hệ ứng dụng:

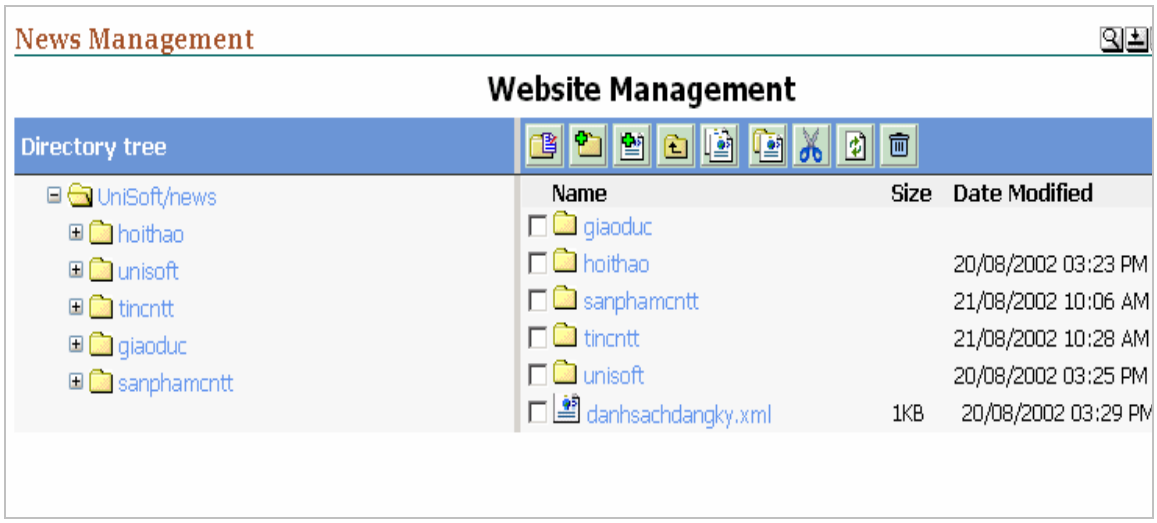
**iNewsReader** dùng để đọc các tin tức trên website. Do được thiết kế như các kênh ứng dụng trong iPortal, hệ thống này có thể phục vụ đồng thời nhiều mục đích khác nhau: thông tin công cộng (dành cho tất cả mọi người, không cần đăng nhập và xác thực) và các thông tin nội bộ (cần đăng nhập và xác thực để có thể truy cập được từng loại thông tin dành riêng cho mỗi đối tượng người dùng). Các thông tin nội bộ như các báo cáo, lịch công tác hàng tuần, hàng tháng dành cho cán bộ lãnh đạo và các cá nhân có liên quan, của từng bộ phận, phòng ban. Do việc cung cấp tin phải qua kiểm duyệt và được phân quyền rõ ràng, nên đảm bảo chất lượng và các yêu cầu khác về quản lý thông tin. Giao diện xem các tin tức của iNewsReader trong trình duyệt có khuôn dạng giống nhau cho các bộ phận của tổ chức, tuy nhiên để truy cập được từng hệ thống tin dành riêng, người sử dụng phải được xác thực (đăng nhập vào website)

Các thông tin, bản tin được tổ chức theo chủ đề (trong minh họa trên thông tin được tổ chức theo chủ đề *Văn hóa* và *Vi tính*) bao gồm tên chủ đề, tên bản tin, tác giả, ngày đăng tin, hình ảnh minh họa kèm theo nếu có và thông tin chi tiết (khi click vào tên bản tin). Ngoài ra, hệ thống iNews còn hỗ trợ chức năng lưu trữ các thông tin đã được đăng và cho phép cơ chế tìm kiếm lại thông tin theo thời gian.



Phân hệ **iNewsEditor** giúp biên soạn một bản tin mới cho hệ thống iNewsReader trong website. Sau khi biên soạn xong một tin mới, người sử dụng sẽ lưu vào trong thư mục định trước trên website dưới dạng tập tin chứa metadata. Giao diện iNewsEditor có dạng như sau :




***iNewsManager*** *là phân hệ quản trị hệ thống thông tin website, cho phép người quản trị có thể tạo mới các thư mục lưu trữ thông tin, sửa chữa, cập nhật các tin, xoá các tin cũ hay đưa vào vị trí lưu trữ (không hiển thị khi người dùng xem tin). Giao diện và các chức năng của kênh ứng dụng này được thiết kế gần giống với một cửa sổ quản trị thư mục của hệ điều hành Windows, giúp người dùng dễ dàng làm quen nhanh chóng với việc sử dụng.*