XÂY DỰNG HỆ THỐNG THÔNG TIN PHỤC VỤ TRA CỨU VĂN HÓA QUẢNG NGÃI

Phan Thị Thanh Tuyền; Võ Trung Hùng* **

TÓM TẮT

Quảng Ngãi là vùng đất có bề dày lịch sử về văn hóa, điển hình là nền văn hóa Sa Huỳnh, văn hóa Chăm Pa, hệ thống thành lũy Chàm… Thực tế tại Quảng Ngãi cho thấy vấn đề bảo tồn và phát triển về văn hóa còn nhiều hạn chế.

Vì vậy, để góp phần vào việc bảo tồn văn hóa Quảng Ngãi chúng tôi đã tiến hành khảo sát thực trạng tại địa phương và đưa ra giải pháp xây dựng hệ thống thông tin để phục vụ tra cứu văn hóa, áp dụng công nghệ thông tin trong việc xây dựng kho dữ liệu, xây dựng mô hình tổng thể với các mô đun hoạt động độc lập giúp cho việc tra cứu thông tin được thuận lợi.

BUILDING AN INFORMATION SYSTEM FOR QUANG NGAI CULTURAL QUERY

SUMMARY

Quang Ngai is a land of history and rich culture with historic treasures such as Sa Huynh Culture, Cham Pa Culture and Cham Ramparts, etc. However, the current practice of cultural preservation and development in Quang Ngai has many limitations.

With the aim of contributing to the preservation of Quang Ngai culture, I have analyzed the local situation and built an information system that helps to store and access to the cultural information quickly and automatically. In this system, I have developed a framework with independent modules and a database warehouse to enable efficient information access.

1. ĐẶT VẤN ĐỀ

Trong giai đoạn hội nhập và giao thoa văn hóa, có nhiều phong tục, địa điểm, làng nghề văn hóa… đã dần bị mai một hoặc lãng quên do nhiều lý do cả về chủ quan lẫn khách quan. Thực tế tại Quảng Ngãi cho thấy vấn đề bảo tồn và phát triển văn hóa còn nhiều hạn chế và thực hiện một cách thủ công, có nguy cơ thất lạc hay bị mai một trong nền kinh tế thị trường. Công nghệ thông tin hiện đang phát triển mạnh và được nhiều nước trên thế giới sử dụng trong bảo tồn, giới thiệu văn hóa một cách hiệu quả. Tuy nhiên, ở Việt Nam nói chung và Quảng Ngãi nói riêng, việc ứng dụng công nghệ thông tin vào quá trình giữ gìn và phát triển bản sắc văn hóa vẫn còn nhiều hạn chế. Xuất phát từ lý do đó, chúng tôi đã chọn thực hiện đề tài: *"Xây dựng hệ thống thông tin phục vụ tra cứu văn hóa Quảng Ngãi"*.

* GV. Khoa Công nghệ - Trường Đại học Công nghiệp Tp.HCM – CS Quảng Ngãi
** PGS.TS. Đại học Đà Nẵng

2. MỤC ĐÍCH, ĐỐI TƯỢNG VÀ PHƯƠNG PHÁP NGHIÊN CỨU

2.1. Mục đích nghiên cứu

Xây dựng hệ thống thông tin để bảo tồn và phát triển văn hóa Quảng Ngãi, ứng dụng vào quá trình lưu trữ, khai thác dữ liệu và ứng dụng công nghệ thông tin trong việc giữ gìn và phát triển bản sắc văn hóa của Quảng Ngãi, góp phần vào công cuộc xây dựng và bảo tồn văn hoá dân tộc của đất nước.

2.2. Đối tượng nghiên cứu

Từ mục đích nghiên cứu xác định đối tượng và phạm vi nghiên cứu: nghiên cứu tổng quan về văn hóa Quảng Ngãi, đặc biệt chú trọng về văn hóa Sa Huỳnh. Nghiên cứu lý thuyết về công nghệ xây dựng hệ thống thông tin, Data warehouse; các kỹ thuật hỗ trợ để xây dựng ứng dụng giúp cho việc lưu trữ, khai thác và bảo tồn và phát triển bản sắc văn hoá Quảng Ngãi ngày càng phong phú đa dạng.

2.3. Phương pháp nghiên cứu

Nghiên cứu lý thuyết về công nghệ xây dựng kho dữ liệu, xử lý phân tích trực tuyến (OLAP), xây dựng website về văn hóa. Khảo sát, phân tích dữ liệu từ nhiều nguồn khác nhau. Từ kết quả phân tích, tiến hành xây dựng kho dữ liệu chuyên ngành văn hóa cùng các module tích hợp dữ liệu và cuối cùng phát triển trang web để cung cấp thông tin cho mọi người dùng

3. NGHIÊN CỨU TỔNG QUAN

3.1. Tình trạng lưu trữ dữ liệu về văn hóa ở Quảng Ngãi

Hiện nay, việc bảo tồn và phát triển văn hóa Quảng Ngãi mà điển hình là nền văn hóa Sa Huỳnh còn nhiều hạn chế, thông tin phân tán, không tập trung, việc ứng dụng công nghệ thông tin để đưa văn hóa Quảng Ngãi đến với mọi người còn rất sơ sài và vẫn chưa được các cấp, các đơn vị quan tâm đúng mức.

Để có thể xây dựng hệ thống thông tin phục vụ người dùng tra cứu về văn hóa Quảng Ngãi cần phải xây dựng kho dữ liệu về văn hóa, tuy nhiên hiện nay các thông tin về văn hóa Quảng Ngãi nằm rải rác trong nhiều cơ sở dữ liệu khác nhau trong hệ thống các trang web của tỉnh, các trang web của các báo và các viện nghiên cứu. Thông tin về văn hóa Sa Huỳnh chủ yếu được lưu trữ bằng hiện vật, bằng hình ảnh hoặc thông tin liệt kê trong các tập tin Excel, Access. Các địa điểm văn hóa về làng nghề, di chỉ thì do Sở Văn hóa, Thể thao và Du lịch quản lý… Tất cả thông tin này không được liên kết về một điểm, gây không ít khó khăn cho người dùng khi có nhu cầu tra cứu thông tin hay tìm hiểu về văn hóa Quảng Ngãi hay văn hóa Sa Huỳnh.

Việc lưu trữ dữ liệu rất hạn chế, trang web của Sở Văn hóa, Thể thao và Du lịch quản lý tỉnh Quảng Ngãi chỉ có nhiệm vụ đưa tin lên trang web dưới hình thức điểm tin hàng tuần hoặc hàng tháng, những sự kiện văn hóa được mô tả sơ sài, ngắn gọn. Những đề tài nghiên cứu cấp tỉnh về văn hóa Quảng Ngãi của Sở Khoa học và Công nghệ tỉnh không được công bố rộng rãi và tính ứng dụng thấp. Bảo tàng tỉnh Quảng Ngãi lưu trữ nhiều hiện vật cũng như những thông tin về văn hóa của vùng cao, vùng xa Quảng Ngãi nhưng chỉ giới thiệu những hình ảnh tiêu biểu còn thông tin chi tiết không được công bố, và được quản lý nghiêm ngặt mang tính cục bộ địa phương. Khi cần nghiên cứu người dùng sẽ rất khó khăn để có được dữ liệu vì các cơ quan ban ngành không có cơ sở dữ liệu riêng để phục vụ những đối tượng này.

3.2. Công nghệ sử dụng

3.2.1. Kho dữ liệu

Kho dữ liệu (Data Warehouse - DW) là tập hợp của các cơ sở dữ liệu tích hợp, hướng chủ đề, thường thiết kế để hỗ trợ cho chức năng trợ giúp quyết định, khai phá dữ liệu.

Kho dữ liệu thường rất lớn tới hàng trăm GB hay thậm chí hàng Terabyte.

Dữ liệu phát sinh từ các hoạt động hàng ngày và được thu thập xử lý để phục vụ công việc nghiệp vụ cụ thể của một tổ chức thường được gọi là dữ liệu tác nghiệp (operational data) và hoạt động thu thập xử lý loại dữ liệu này được gọi là xử lý giao dịch trực tuyến (On Line Transaction Processing - OLTP). Kho dữ liệu còn phục vụ cho việc phân tích với kết quả mang tính thông tin cao. Các hệ thống thu thập xử lý dữ liệu loại này còn gọi là hệ xử lý phân tích trực tuyến (On Line Analytical Processing - OLAP).

3.3. Tính chất kho dữ liệu

- Hướng chủ đề (Subject-oriented).

- Dữ liệu có tính tích hợp (Intergated).

- Dữ liệu gắn thời gian và có tính lịch sử (Time-variant).

- Dữ liệu chỉ đọc (Read Only).

- Dữ liệu không biến động (Nonvolatile).

3.4. Các vấn đề kỹ thuật

Hướng tiếp cận công nghệ: Cơ sở dữ liệu tập trung là một nền tảng cơ bản của môi trường Data Warehousing. Cơ sở dữ liệu này hầu hết được cài đặt dựa trên công nghệ của hệ thống quản trị cơ sở dữ liệu quan hệ (RDBMS). Những cách tiếp cận đó bao gồm:

- Thiết kế cơ sở dữ liệu quan hệ song song.

- Một cách tiếp cận mới để làm tăng tốc độ RDBMS truyền thống bằng cách sử dụng một cấu trúc chỉ số bỏ qua kiểm tra các bảng quan hệ.

- Các cơ sở dữ liệu đa chiều (MDDBS), dựa trên công nghệ cơ sở dữ liệu phổ biến hoặc được cài đặt sử dụng trên nền RDBMS quen thuộc đã xuất hiện trên thị trường. Cơ sở dữ liệu đa chiều được thiết kế để khắc phục những giới hạn tồn tại trong DW gây ra do bản chất của mô hình dữ liệu quan hệ. Cách tiếp cận này gắn liền với các công cụ xử lý phân tích trực tuyến thực hiện như một đối tác của các kho dữ liệu đa chiều. Các công cụ này gộp lại thành một nhóm công cụ truy vấn, tạo báo cáo, phân tích và đào xới dữ liệu.

Mô hình dữ liệu đa chiều: Bản chất đa chiều của dữ liệu thể hiện nhiều trong việc báo cáo, phân tích của một doanh nghiệp.

3.5. Xử lý phân tích trực tuyến – OLAP

OLAP là một kỹ thuật sử dụng các thể hiện dữ liệu đa chiều gọi là các khối (cube) nhằm cung cấp khả năng truy xuất nhanh đến dữ liệu của kho dữ liệu. Tạo khối (cube) cho dữ liệu trong các bảng chiều (dimension table) và bảng sự kiện (fact table) trong kho dữ liệu và cung cấp khả năng thực hiện các truy vấn tinh vi và phân tích cho các ứng dụng client.

3.6. Trích lọc dữ liệu

Một trong những mục tiêu chính của kho dữ liệu là tích hợp dữ liệu (Data Intergration) từ các nguồn khác nhau. Do đó, sau khi xác định được các hệ thống nguồn và tạo cấu trúc các bảng của kho dữ liệu, chúng ta cần trích lọc và nạp dữ liệu từ các nguồn đó về kho dữ liệu.

Trong phần này, chúng tôi xin giới thiệu về nguyên lý chung của việc trích dữ liệu (data extraction) và công cụ SSIS.

Giới thiệu về ETL

ETL là chữ viết tắt từ Extract (trích), Transform (chuyển đổi) và Load (nạp). Đây là quá trình lấy dữ liệu về từ hệ thống nguồn, chuyển đổi nó rồi đưa nó vào kho dữ liệu.

Có một số nguyên tắc cơ bản cần tìm hiểu khi trích dữ liệu từ hệ thống nguồn. Trước hết, lượng dữ liệu lấy về rất lớn, có thể đến hàng trăm megabyte hay hàng chục gigabyte. Hệ thống OLTP được thiết kế để dữ liệu được lấy về từng lượng nhỏ chứ không phải lượng lớn

như thế. Vì vậy, bạn cần cẩn thận không làm cho hệ thống nguồn bị chậm lại quá nhiều. Chúng ta muốn trích dữ liệu nhanh đến mức có thể, chẳng hạn năm phút chứ không phải ba tiếng đồng hồ. Chúng ta cũng muốn dữ liệu nhỏ đến mức có thể, chẳng hạn 10MB mỗi ngày nếu có thể chứ không phải 1GB mỗi ngày. Thêm vào đó, chúng ta muốn nó hiếm khi xảy ra, chẳng hạn một lần một ngày nếu có thể thay vì mỗi năm phút. Chúng ta muốn sự thay đổi trong hệ thống nguồn tối thiểu đến mức có thể thay vì lưu giữ lại những thay đổi dữ liệu trong mỗi bảng đơn (single table).

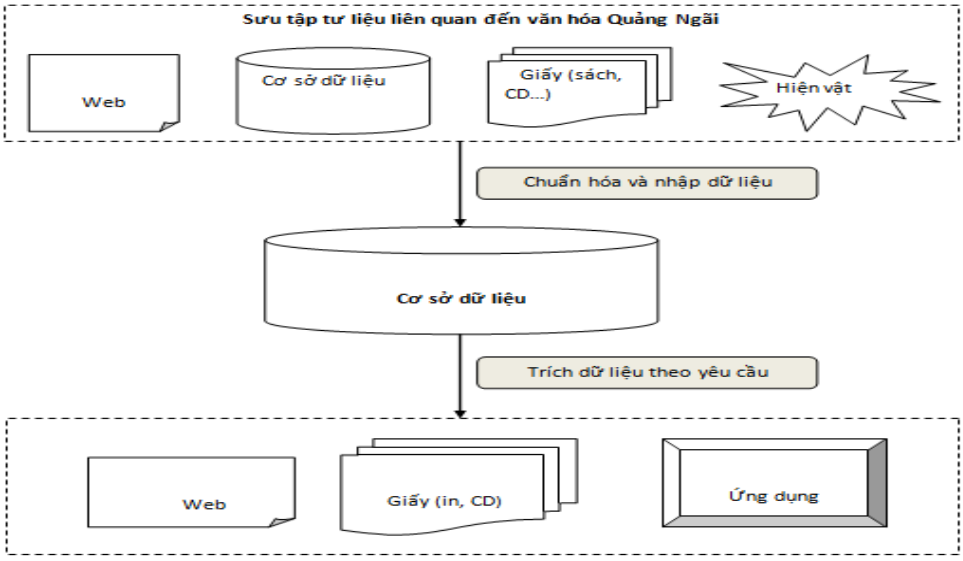
Sau khi trích dữ liệu, chúng ta muốn nạp nó vào thẳng kho dữ liệu càng sớm càng tốt mà không đụng chạm gì đến ổ đĩa cả (nghĩa là không phải lưu trữ tạm thời dữ liệu trong cơ sở dữ liệu hay trong các file). Chúng ta cần thực hiện một vài chuyển đổi dữ liệu để cho nó phù hợp với định dạng và cấu trúc dữ liệu trong NDS và DDS. Đôi khi sự chuyển đổi dữ liệu chỉ là định dạng, chuẩn hóa, chuyển định dạng ngày tháng hay số, cắt bỏ ký tự trống phía sau hay chèn thêm số 0 đằng trước. Có khi là thay đổi sự tra cứu, Sự chuyển đổi khác cũng hay gặp trong kho dữ liệu là sự tập hợp, nghĩa là tổng kết dữ liệu ở mức cao hơn.

Chúng ta cũng muốn dữ liệu đưa vào kho là dữ liệu sạch và chất lượng tốt. Thí dụ, chúng ta không muốn có mã cổ vật bất hợp lệ, một địa chỉ không tồn tại hay một cổ vật không có năm khai quật. Với mục đích đó, chúng ta cần nhiều kiểm tra khác nhau trước khi đưa dữ liệu vào kho.

Hai nguyên tắc quan trọng khác là sự rò rỉ (leakage) và khả năng phục hồi (recoverability). Sự rò rỉ xảy ra khi ta tưởng rằng tiến trình ELT đã tải về toàn bộ dữ liệu từ hệ thống nguồn nhưng thực ra đã bị sót vài bản ghi. Một tiến trình ELT tốt không được có bất cứ sự rò rỉ nào. Khả năng phục hồi là khi có hỏng hóc, tiến trình ELT có thể hồi phục mà không làm mất hay hỏng dữ liệu.

4. PHÂN TÍCH THIẾT KẾ HỆ THỐNG

4.2. Giới thiệu hệ thống



Hình 1: *Kiến trúc tổng quát của hệ thống*

Khảo sát các hệ thống nguồn

Loại	Hệ thống đơn vị sử dụng	Chức năng	Hệ quản trị cơ sở dữ liệu
Website	Website chuyên ngành Sở VHTT&DL http://www.quangngai.gov.vn/quangngai/tiengviet/chuyennganh/sovhtt/134867/	Website giới thiệu tin tức, các hoạt động của ngành	MS SQL Server
	Website địa chí Quảng Ngãi http://www.quangngai.gov.vn/userfiles/file/dudiachiquangngai/Trangchu.htm	Giới thiệu Các vấn đề về văn hóa xã hội, tự nhiên, con người Quảng Ngãi	MS SQL Server
CT	Chương trình quản lý bảo tàng của Bảo tàng tỉnh Quảng Ngãi	Quản lý cổ vật trong bảo tàng	MS Access
File	File báo cáo hàng quý của các đơn vị văn hóa cơ sở	Báo các các hoạt động văn hóa từng quý	File Word File Excel
	Hình ảnh cổ vật được sưu tập tại nhà một số nghệ nhân	Cổ vật được khai quật và sưu tầm	File Word File ảnh
	Sách viết về văn hóa Sa Huỳnh	Giới thiệu về văn hóa Sa Huỳnh qua các niên đại…	File Word (nhập thủ công)

4.3. Phân tích thiết kế hệ thống

Mô tả người dùng

Người dùng của hệ thống chia làm ba nhóm chính:

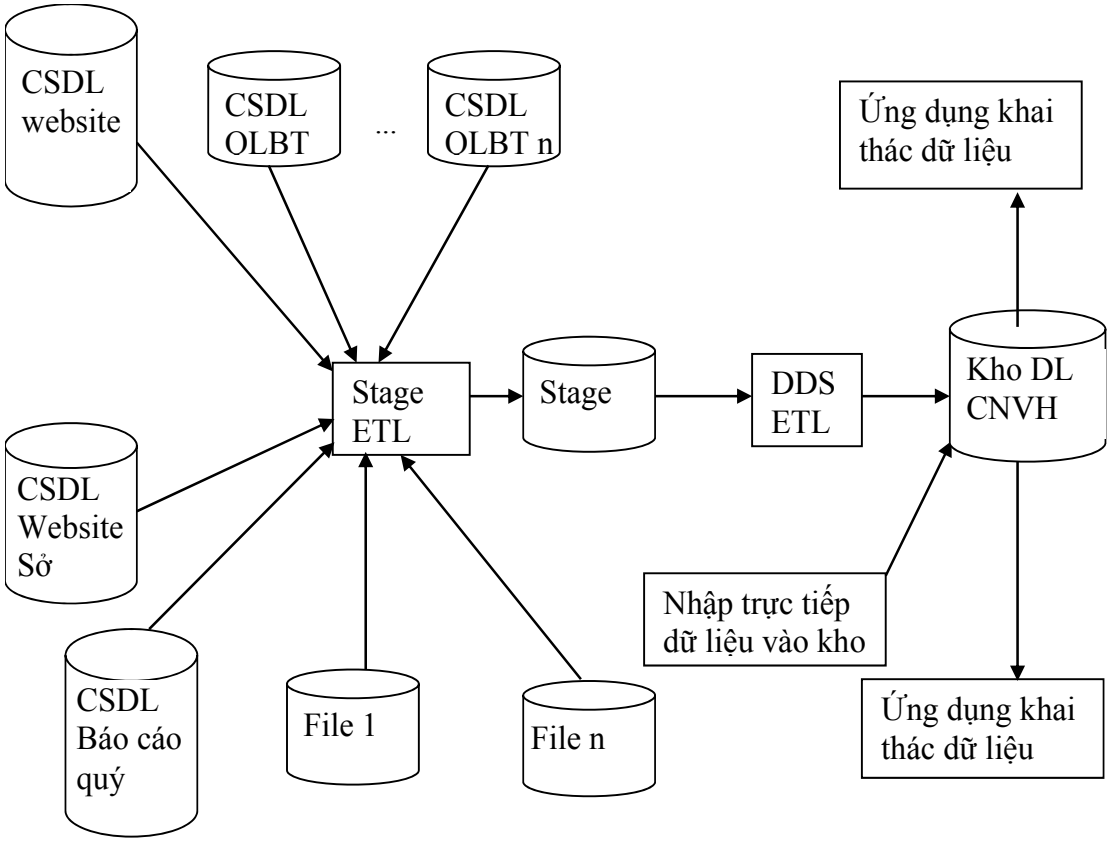
- Người dùng cấp 1: tiếp cận hệ thống thông qua giao diện chính hệ thống cung cấp: website cung cấp thông tin và cho phép tìm kiếm dữ liệu.

- Người dùng cấp 2: cho phép khai thác dữ liệu hệ thống ở mức độ nâng cao, ví dụ tải dữ liệu khối lượng lớn, tra cứu chuyên sâu…

- Quản trị: quản lý các tài khoản của người sử dụng (cấp quyền, hủy quyền), thiết lập định kỳ cập nhật dữ liệu, kết nối các nguồn dữ liệu mới…

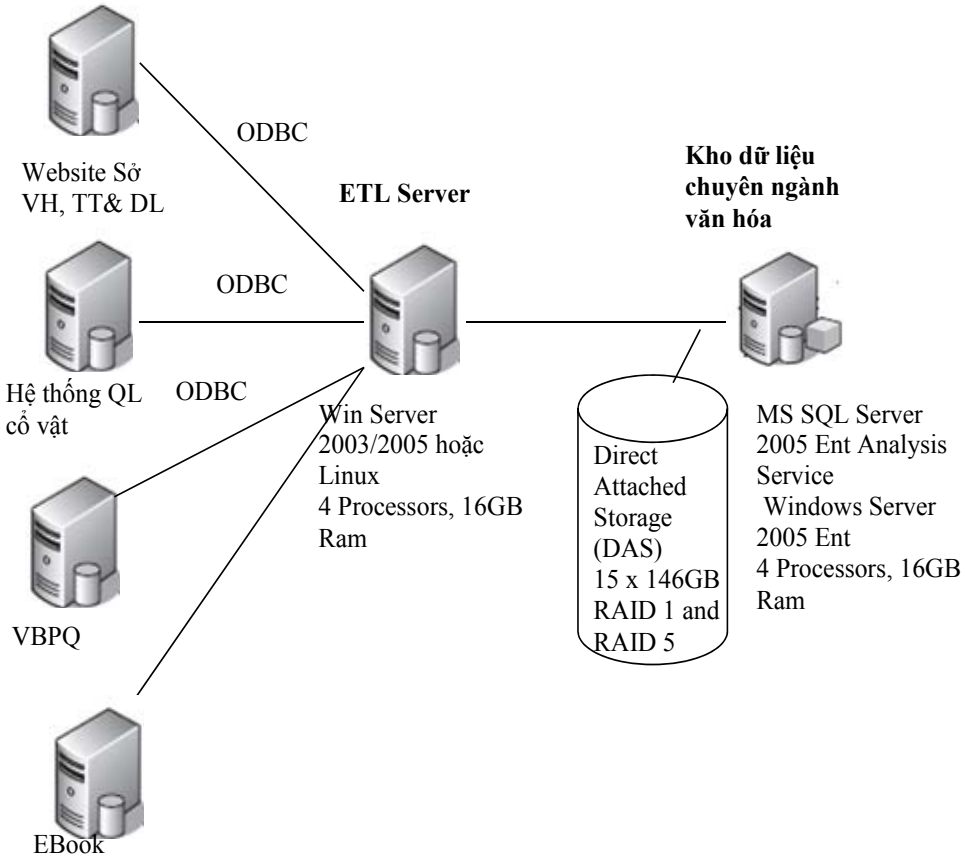
Hình 2: *Sơ đồ Use Case*

Kiến trúc luồng:



Hình 3: *Kiến trúc luồng dữ liệu của kho dữ liệu chuyên ngành văn hóa*

Kiến trúc hệ thống



Hình 4: *Kiến trúc hệ thống của kho dữ liệu chuyên ngành văn hóa*

4.4. Đặc tả cấu trúc kho dữ liệu văn hóa

Căn cứ trên các hệ thống nguồn đã khảo sát và nhu cầu xây dựng một kho dữ liệu chuyên phục vụ tra cứu về văn hóa, tiến hành xây dựng kho dữ liệu bằng cách xây dựng các kho dữ liệu cục bộ (Data mart - DM) tương ứng sau:

- Cổ vật (Antiques)
- Tin tức (News)
- Ebook
- Văn bản pháp quy (Rule Documents)

4.5. Mô hình tích hợp và cập nhật dữ liệu từ nhiều nguồn khác nhau

Dữ liệu hệ thống được lấy từ ba nguồn chính là website, báo cáo tháng và ứng dụng quản lý cổ vật ở bảo tàng. Do đó, mô hình tích hợp dữ liệu (data integration) được thiết kế với ba chức năng chính:

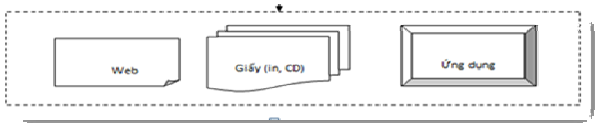
- Tích hợp dữ liệu từ website;
- Tích hợp dữ liệu từ hệ thống quản lý cổ vật ở bảo tàng;
- Tích hợp dữ liệu từ các báo cáo tháng,

các file dữ liệu.

Dữ liệu sau khi được lấy về từ các hệ thống nguồn sẽ được chuyển đổi và nạp vào trong DW. Nếu là lần lấy đầu tiên, dữ liệu sẽ được nạp toàn bộ vào trong DW. Đối với trường hợp cập nhật dữ liệu từ các nguồn dữ liệu trước đó, phương pháp trích lọc dữ liệu

tăng dần (incremental extraction) sẽ được sử dụng để giảm chi phí cập nhật dữ liệu.

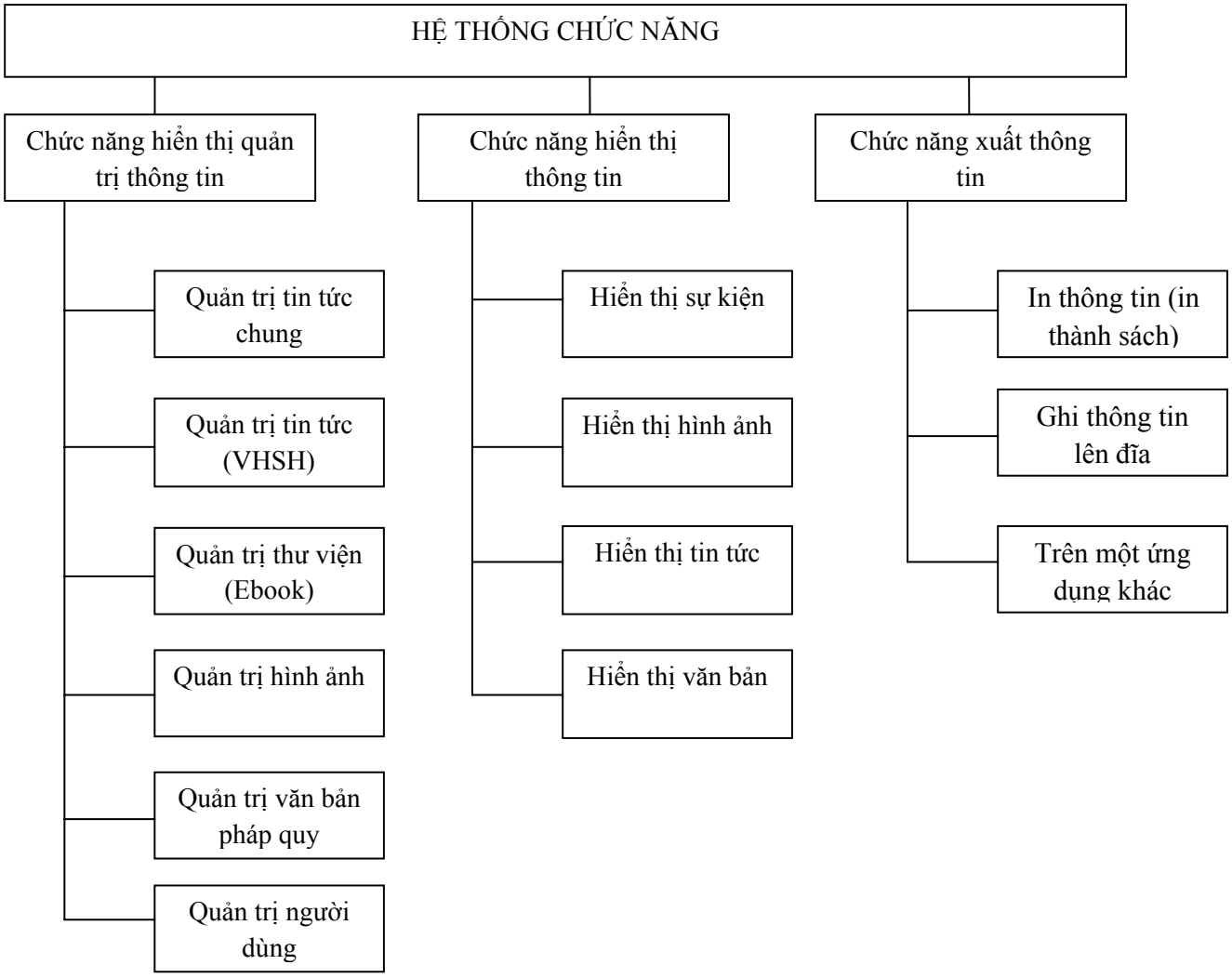
4.6. Mô hình khai thác



Mô hình chức năng:

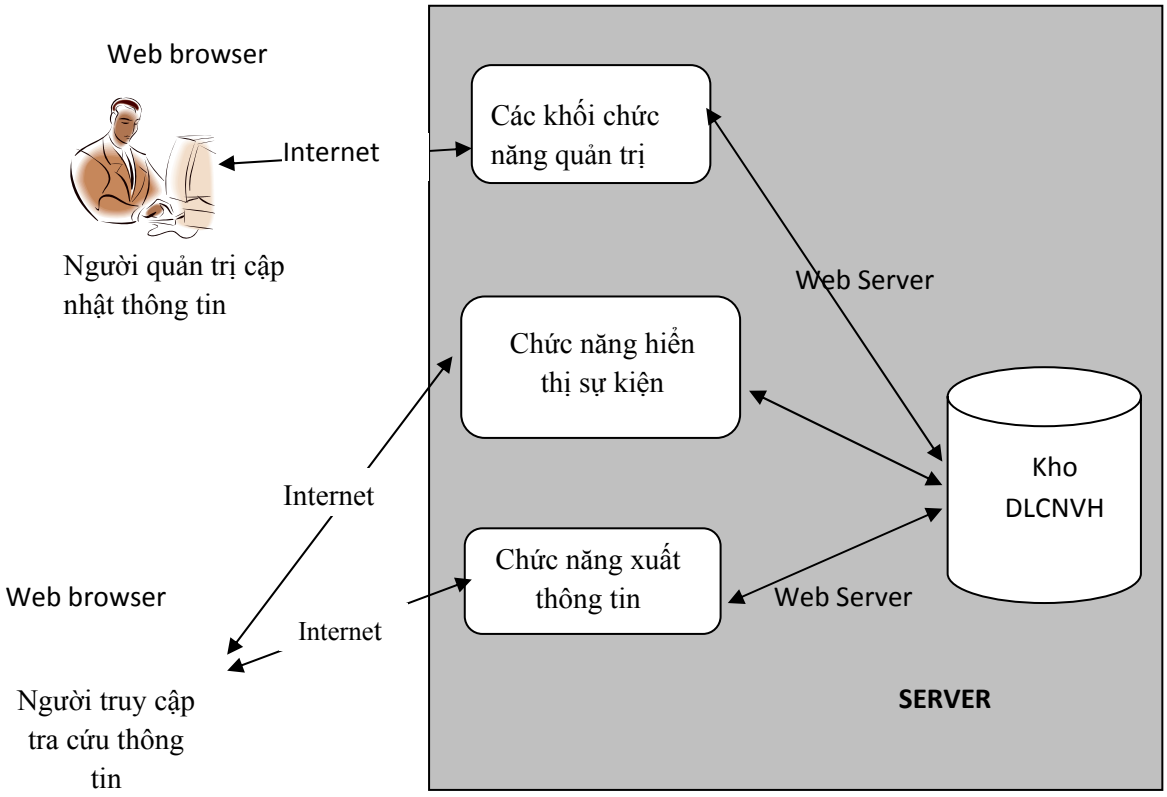
5. XÂY DỰNG HỆ THỐNG

Phát triển các Mô đun để tích hợp và cập nhật dữ liệu; tích hợp dữ liệu từ websites, tích hợp dữ liệu từ hệ thống quản lý cổ vật ở bảo tàng, tích hợp dữ liệu từ các báo cáo. Phát triển các môđun để khai thác dữ liệu từ Data Warehouse.



Hình 5: *Mô hình chức năng*

Kiến trúc tổng thể của Website văn hóa Quảng Ngãi



Hình 6: *Kiến trúc tổng thể của Website văn hóa Quảng Ngãi*

6. KẾT QUẢ ĐẠT ĐƯỢC

Quá trình nghiên cứu đã nêu được giải pháp kỹ thuật để xây dựng hệ thống thông tin phục vụ tra cứu về văn hóa Quảng Ngãi.

Về lý thuyết đã nêu được khái niệm, kỹ thuật xây dựng kho dữ liệu, cách xây dựng các mô đun tích hợp dữ liệu từ kho.

Về thực tiễn ứng dụng, luận văn đã xây dựng được Kho dữ liệu chuyên ngành văn hóa, các module tích hợp dữ liệu và trang web khai thác thông tin từ kho dữ liệu. Tuy nhiên, dữ liệu tích hợp được chỉ mang tính trình diễn, thử nghiệm về văn hóa Sa Huỳnh, chưa tích hợp một cách đầy đủ và kịp thời.

TÀI LIỆU THAM KHẢO

Tiếng Việt

[1] Báo cáo hàng tháng của Sở Văn hóa, Thể thao và Du lịch tỉnh Quảng Ngãi.

[2] Hình ảnh Bộ sưu tập hiện vật văn hóa Sa huỳnh - Nghệ nhân Lâm Dũ Xênh, Bình Sơn Quảng Ngãi.

Tiếng Anh

[3] G.G. Shanks, P.A. O'Donnell and D.R. Arnott (2003), Data warehousing.

[4] Jiawei Han and Micheline Kamber (2006), Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, Elsevier Inc.

[5] Vincent Rainardi (2007), *Building a Data warehouse with Examples in SQL Server,* Apress.

Trang web

[6] http://www.quangngai.gov.vn/