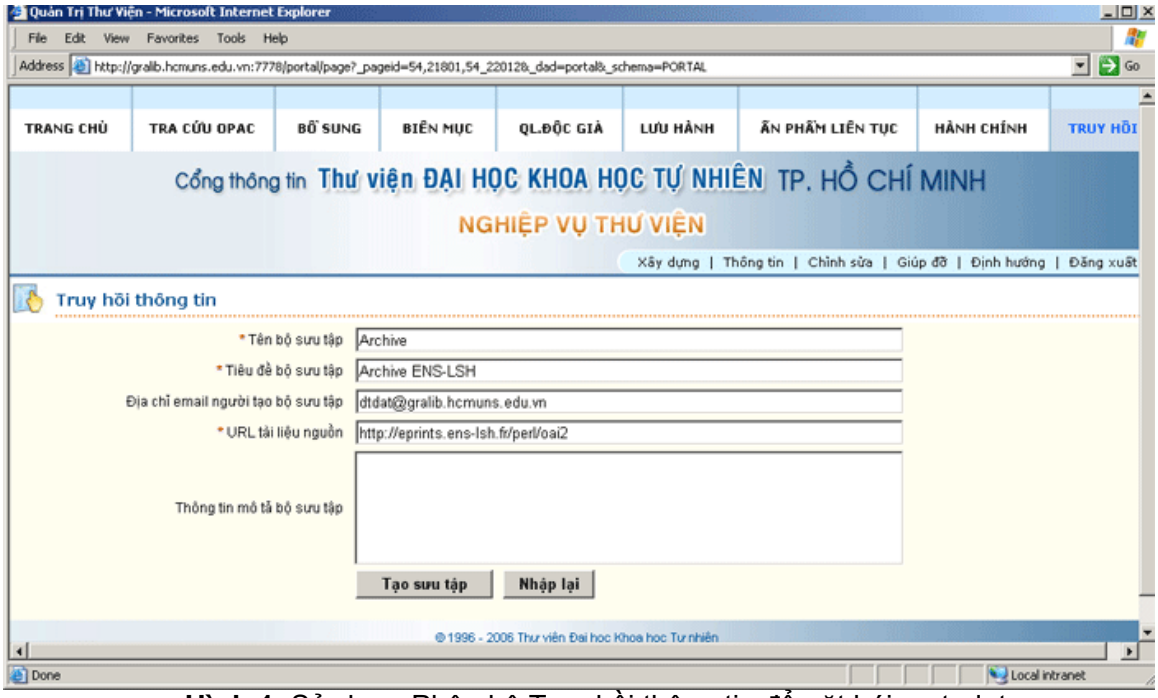
GẶT HÁI SIÊU DỮ LIỆU

Metadata Harvesting

LƯƠNG MINH HÒA
DƯƠNG TÍCH ĐẠT
Phòng Công tác Kỹ thuật
Thư viện ĐH Khoa học Tự nhiên TP. HCM

Để xây dựng những Bộ sưu tập thư viện số, chúng ta cần phải *tạo lập metadata* và *gặt hái metadata.* **Tạo lập metadata** là chủ động biên mục tài liệu sưu tầm được để xây dựng những Bộ sưu tập với đầy đủ nội dung được lưu trữ trên server của thư viện mình; trong khi **Gặt hái metadata** được dùng để xây dựng những Bộ sưu tập chỉ bao gồm metadata, tuy nhiên qua từng siêu dữ liệu thư tịch trong Bộ sưu tập ta có thể truy cập đến nội dung của tài liệu ở khắp nơi – Đây là một hình thức Thư viện ảo.

Thư viện ĐH Khoa học Tự nhiên TP. HCM sử dụng Phân hệ Truy hồi thông tin trong Hệ thống quản lý thư viện để gặt hái thông tin (Hình 1). Đây là phân hệ giúp chúng ta xây dựng Bộ sưu tập số từ địa chỉ liên kết URL mà người sử dụng cung cấp. Phân hệ này cho phép tập hợp các siêu dữ liệu thư tịch (*bibliographic metadata)* theo chuẩn OAI-PMH từ địa chỉ liên kết URL được cung cấp, sau đó xây dựng thành Bộ sưu tập số để cho phép độc giả tìm kiếm trên dữ liệu đã lấy về. Mỗi bộ sưu tập là tập hợp các biểu ghi OAI từ một hoặc nhiều địa chỉ liên kết URL. Thư viện ĐH Khoa học Tự nhiên đã tiên phong trong việc ứng dụng giao thức OAI để gặt hái metadata.



Hình 1: Sử dụng Phân hệ Truy hồi thông tin để gặt hái metadata

OAI-PMH là thuật ngữ viết tắt của *Open Archives Initiative - Protocol for Metadata Harvesting* (Sáng kiến lưu trữ mở - Giao thức gặt hái siêu dữ liệu). Thuật ngữ này chỉ

một giao thức theo mô hình khách chủ dựa trên HTTP *(Hyper Text Transfer Protocol)*, cho phép một hệ thống cung cấp dịch vụ (*Service Provider*) có thể truy vấn, lọc và gặt hái siêu dữ liệu để truy hồi các nguồn tài nguyên trên các hệ thống cung cấp dữ liệu *(Data Provider)* theo định dạng chuẩn XML. Thông tin tài nguyên sẽ được lưu trữ, phân loại và hiển thị theo dạng chuẩn Dublin Core, giúp cho người sử dụng nắm bắt thông tin dễ dàng và hiệu quả.

- Những nhà cung cấp dữ liệu
  Danh sách các nhà cung cấp dữ liệu trực tuyến hỗ trợ giao thức OAI-PMH. Danh sách này do tổ chức OAI phát triển và duy trì. Các cơ sở dữ liệu này thuộc nhiều lĩnh vực khoa học khác nhau. Địa chỉ liên kết: http://www.openarchives.org/Register/BrowseSites
- Tổ chức OAI
  Trang chủ của Open Archives Initiative, tổ chức phát triển và duy trì chuẩn OAI-PMH. Địa chỉ liên kết: http://www.openarchives
- OAI-PMH: Phiên bản 2.0
  Toàn văn phiên bản 2.0 của chuẩn OAI-PMH. Địa chỉ liên kết: http://www.openarchives.org/OAI/openarchivesprotocol.html

Phân hệ Truy hồi thông tin trong Hệ thống phần mềm của Thư viện được xây dựng từ 3 thành phần chính: Oracle Portal, Greenstone, OAI Protocol. Việc kết hợp này có một số ưu điểm sau:

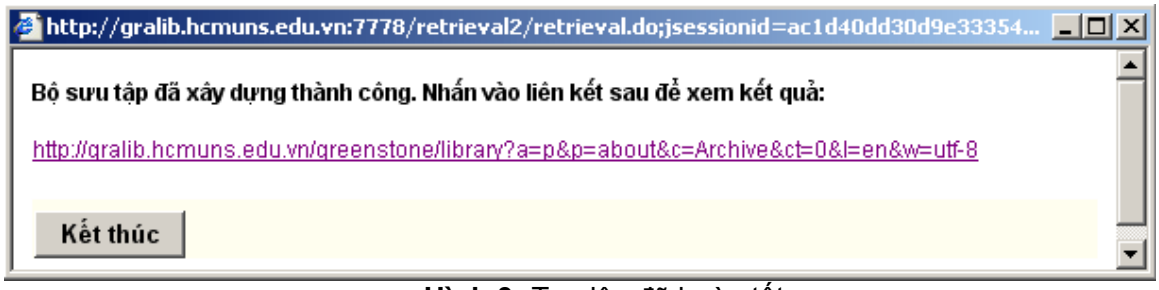
- Nhúng Greenstone như là một ứng dụng thành phần của Portal nhờ khả năng tích hợp ứng dụng của Oracle Portal
- Kế thừa các tính năng nổi bật của Greenstone như tổ chức và tìm kiếm tài liệu
- Gặt hái tài liệu từ các nguồn trên Internet dựa vào giao thức OAI

Việc sử dụng phân hệ Truy hồi rất đơn giản, người dùng chỉ việc cung cấp đầy đủ thông tin cơ bản và nhấn nút **"Tạo bộ sưu tập".** Hình 1

- Tên bộ sưu tập: Bao gồm số và chữ, không vượt quá 8 ký tự.
- Tiêu đề bộ sưu tập: Nhập vào tên do người dùng định nghĩa.
- Email: Địa chỉ thư điện tử của người tạo bộ sưu tập.
- URL: Địa chỉ nguồn tài liệu mà người dùng đưa vào.
- Thông tin mô tả: Thông tin mô tả liên quan đến bộ sưu tập.

Khi màn hình xuất hiện cửa sổ cho phép người dùng xác nhận việc xây dựng bộ sưu tập nhấn nút **"Đồng ý"**. Hình 2.

Hình 2: Xác nhận thực hiện tạo lập bộ sưu tập



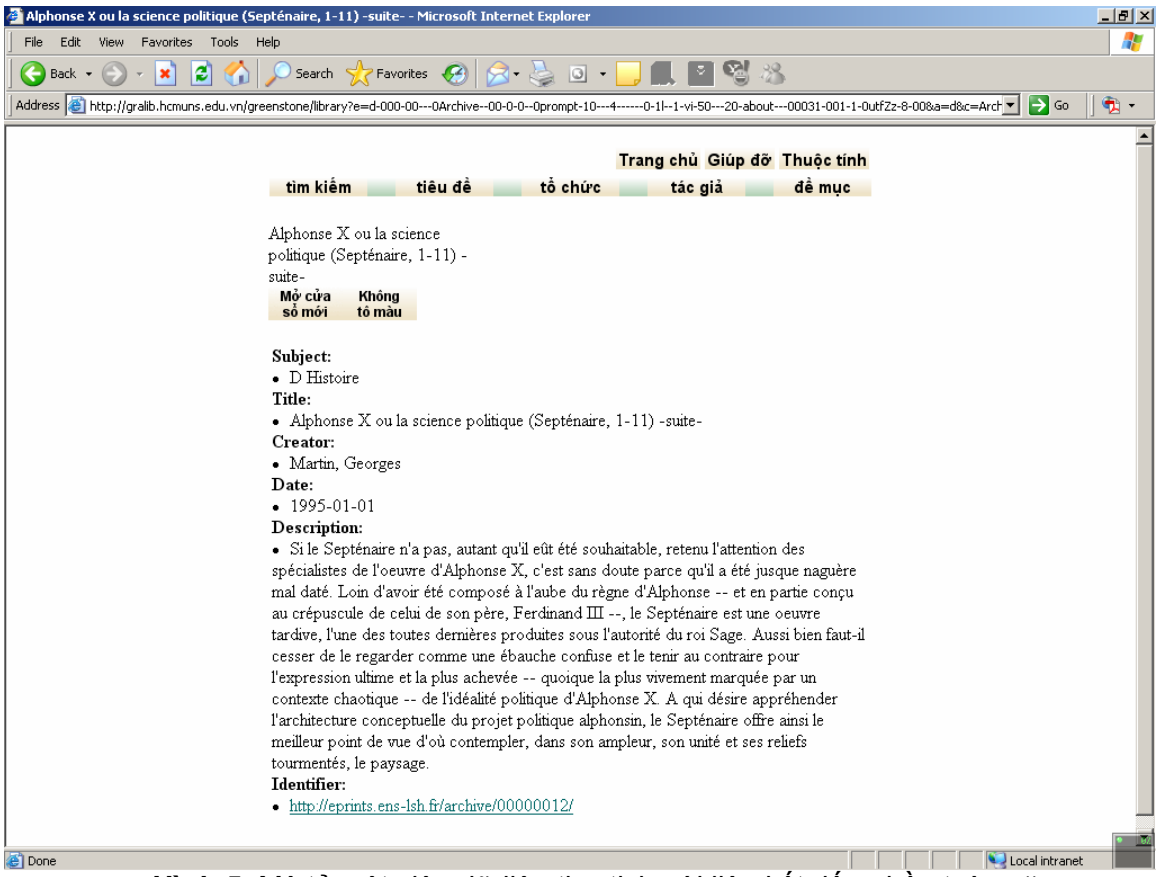
Hình 3: Tạo lập đã hoàn tất

Sau khi kết thúc quá trình tạo lập, chúng ta click chuột vào đường liên kết Hình 3. Giao diện bộ sưu tập xuất hiện trên trình duyệt web với dạng thức Greenstone (Hình 4).

Hình 4: Giao diện của bộ sưu tập

Bộ sưu tập chứa những siêu dữ liệu thư tịch. Mỗi siêu dữ liệu hay metadata bao gồm 15 thành phần của Dublin Core mô tả chi tiết nguồn tài nguyên kể cả tóm tắt nội dung với đầy đủ những tiêu đề (nhan đề, tác giả, đề mục) và những điểm truy cập khác (Hình 5). Nếu muốn xem phần toàn văn thì click chuột vào đường liên kết ở thành phần **Indentifier** đến server - nơi cung cấp bộ sưu tập.

Những bộ sưu tập được tạo nên do gặt hái metadata mặc dù thông tin rất nhiều nhưng tiết kiệm được không gian lưu trữ. Đây chính là điều mà những người làm thông tin luôn quan tâm.



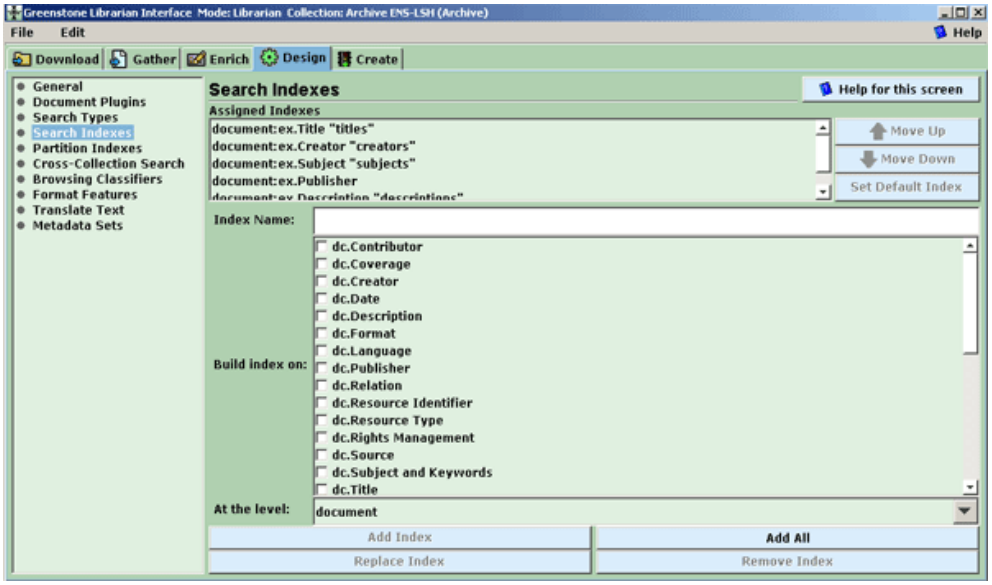
Hình 5: Mô tả một siêu dữ liệu thư tịch với liên kết đến phần toàn văn

Ngoài ra chúng ta có thể xây dựng, chỉnh sửa giao diện, thiết lập lại các điểm truy cập nhan đề, đề mục, tác giả… của bộ sưu tập.

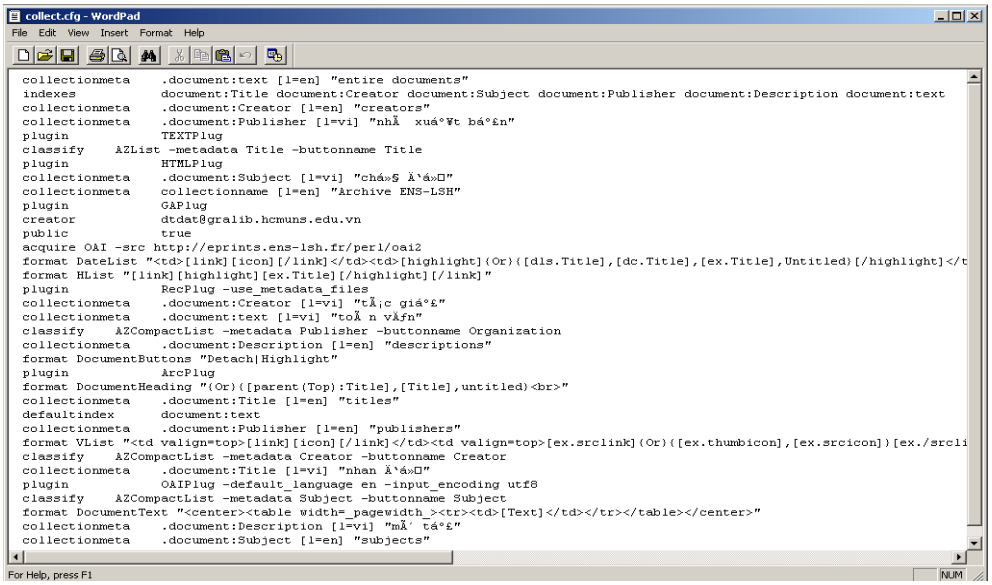
Có hai phương pháp để chỉnh sửa:

- Phương pháp thứ nhất: Dùng giao diện Greenstone Librarian Interface của phần mềm Thư viện số Greenstone, phương pháp này cho phép chúng ta tái biên mục, thiết lập các kiểu thể hiện biểu ghi, bổ sung các điểm truy cập. Xây dựng bộ sưu tập từ đầu dựa trên dữ liệu đã có sẵn. Hình 6.

- Phương pháp thứ hai: Chúng ta chỉnh sửa trực tiếp trên tập tin collect.cfg. Tập tin này chứa thông tin định dạng và được lưu trong thư mục ect của mỗi bộ sưu tập. Phương pháp này chỉ áp dụng khi chúng ta muốn thêm hình ảnh, logo, thay đổi giao diện và đòi hỏi cán bộ Thư viện phải có kiến thức về hệ điều hành MS-DOS. Hình 7



Hình 6: Bộ sưu tập được chỉnh sửa trên giao diện Greenstone Librarian Interface



Hình 7: Chỉnh sửa trên tập tin collect.cfg

Mỗi bộ sưu tập được tạo nên do chúng ta gặt hái metadata có thể chứa nhiều loại tài liệu khác nhau và nhiều định dạng khác nhau: văn bản, âm thanh, hình ảnh, video…Có thể nói OAI-PMH là một phương thức mới không chỉ giúp chúng ta gặt hái metadata để tạo nên nhiều bộ sưu tập Thư viên số mà còn giúp cho việc tổ chức lưu trữ dữ liệu trở nên đơn giản và tiện lợi.

Cơ sở dữ liệu của Thư viện ngày càng phong phú, tiết kiệm được không gian lưu trữ, đây là điều rất lý tưởng cho các Thư viện đã đang và sẽ xây dựng Thư viện số.

Gặt hái metadata để xây dựng những Bộ sưu tập thư viện số đồng nghĩa với việc xây dựng Thư viện ảo.