THƯ VIỆN TRƯỜNG ĐẠI HỌC KHOA HỌC XÃ HỘI & NHÂN VĂN TP. HỒ CHÍ MINH TRÊN LỘ TRÌNH XÂY DỰNG THƯ VIỆN SỐ

PGS. TSKH. Bùi Loan Thùy
Trường Đại học KHXH&NV TP. Hồ Chí Minh

Trình bày nhu cầu và phân tích thực trạng của việc số hoá tài liệu tại Thư viện Trường Đại học KHXH&NV TP. Hồ Chí Minh. Nêu một số bài học kinh nghiệm của Trường trong việc số hoá tài liệu trên lộ trình xây dựng thư viện số.

1. Hiện trạng số hoá tại Thư viện

Nhận thức được tầm quan trọng của việc xây dựng nguồn lực thông tin điện tử, ngoài việc bổ sung tài liệu điện tử từ nguồn mua, Thư viện trường Đại học KHXH&NV TP. Hồ Chí Minh đã và đang thực hiện số hóa nguồn tài liệu quý hiếm và tải xuống các tài liệu điện tử trên mạng Internet.

Lượng tài liệu quý hiếm cần số hóa tại thư viện phần lớn là tài liệu tiếng Việt cũ đã bị chuyển màu vàng, độ giòn cao (386.602 trang tạp chí nghiên cứu, 4643 tên sách tiếng Việt: 2.030.407 trang, 5329 tên sách một bản: 1.748.176 trang, 290 luận án trước 1975: 52.961 trang).

Việc số hóa tài liệu bắt đầu được tiến hành từ năm 2005. Tính đến 6/2007 Thư viện đã số hóa được 1.028.551 trang tài liệu các loại. Năm đầu tiên số hóa thử nghiệm 66.797 trang, đưa ra khai thác sử dụng đạt hiệu quả cao. Trên cơ sở này Thư viện đã lập dự án "Nâng cao năng lực cung ứng thông tin KHXH&NV phục vụ nghiên cứu khoa học", trong đó số kinh phí dành cho việc số hóa tài liệu quý hiếm được Đại học Quốc gia TP. Hồ Chí Minh phê duyệt là 637,3 triệu đồng. Trong hai năm 2006, 2007 Thư viện đã đẩy nhanh tốc độ số hóa (năm 2006 số hóa được 397.659 trang, 6 tháng đầu năm 2007 số hóa được 564.095 trang).

Số hóa tài liệu là một công nghệ phức hợp đòi hỏi đầu tư nhiều tiền bạc, công sức cho việc trang bị các thiết bị và phần mềm tương thích, xử lý tài liệu trong quá trình số hóa. Vì vậy, Thư viện đã chọn lọc kỹ các tài liệu cần ưu tiên số hóa. Thư viện đã xây dựng cơ chế quản lý dữ liệu số hóa, đặc biệt chú ý đến các công cụ tìm kiếm theo các tiêu chí khác nhau (tác giả, nhan đề, từ khóa, chủ đề, môn loại tri thức, nhà xuất bản, loại hình, thể loại ...) và tìm kiếm toàn văn.

Phần mềm Quản trị Nội dung số tích hợp được Thư viện lựa chọn là phần mềm LIBOL. Phần mềm này là một giải pháp tích hợp cho việc xây dựng và phát triển thư viện số, trong đó đối tượng tư liệu là một kho dữ liệu đa phương tiện bao gồm văn bản, hình ảnh, đồ hoạ, âm thanh, video, tệp máy tính, có thể được lưu trữ và khai thác trực tuyến qua mạng máy tính. Phần mềm đã bảo đảm được các yêu cầu của Thư viện đề ra như: tính tích hợp, tính mở và tuỳ biến; Hỗ trợ chuẩn biên mục MARC21, AACR2, ISBD, MARC21 VN, Dublin Core; Hỗ trợ các khung phân loại DDC, BBK, Subject Headings; Nhập/xuất dữ liệu theo chuẩn ISO 2709; Hỗ trợ đa ngữ Unicode với dữ liệu và giao diện làm việc; Hỗ trợ các bảng mã tiếng Việt như TCVN3, VNI, Unicode (TCVN 6909); Bảo mật và phân quyền chặt chẽ đến từng đối tượng trong CSDL; Thống kê tra cứu đa dạng, chi tiết phục vụ mọi nhóm đối tượng; Vận hành hiệu quả trên

những CSDL lớn.

Trong thực tế, việc biên mục tư liệu số được thực hiện dễ dàng, có thể tùy biến các mẫu biên mục, xuất/nhập dữ liệu trực tuyến.

Việc quản lý tài liệu số theo cơ chế đối tượng hệ thống các tập tin (File System Object), cho phép cán bộ thư viện tự xác định tên các bộ sưu tập tài liệu số. Cơ chế quản lý dữ liệu số hóa giúp thư viện quản lý các tư liệu số một cách tập trung, bố trí chương trình theo cấu trúc thư mục của Windows Explore thân thiện với người dùng. Thư viện phân quyền sử dụng tài liệu số của người dùng bằng cách phân cấp các tài liệu số thành các mức độ mật khác nhau theo chính sách phục vụ của Thư viện đối với các nhóm bạn đọc- người dùng tin khác nhau.

Cơ chế tra cứu toàn văn có các cấp độ rõ ràng, dễ hiểu, đơn giản, dễ sử dụng, không bị bó buộc về trình độ và kỹ năng sử dụng của bạn đọc-người dùng tin, cho phép tìm kiếm theo nhiều dấu hiệu tìm kiếm khác nhau bằng các công cụ tìm tin đáp ứng các chuẩn quốc tế về tìm tin như sử dụng toán tử logic, toán tử lân cận, toán tử chặt cụt, toán tử so sánh, các dấu ngoặc, cùng với các khả năng viết các biểu thức tìm tin phức hợp, tìm các nội dung có liên quan, tìm theo một chủ đề nhất định, sắp xếp kết quả theo các tiêu chí khác nhau... thỏa mãn những yêu cầu tìm tin đa dạng, khác nhau của người sử dụng.

Cho đến nay, Thư viện mới chỉ thực hiện qui trình số hóa đối với dữ liệu văn bản (text), và dữ liệu hình ảnh (image /scanned images), chưa thực hiện số hóa với dữ liệu âm thanh (sound), dữ liệu phim tư liệu (video), dữ liệu bài giảng trực tuyến (courseware), dữ liệu hỗn hợp (hybrid). Qui trình số hóa trải qua các bước: Lựa chọn tài liệu quý, hiếm; Quét tài liệu; Phân trang và tự động hoá xử lý hình ảnh; Kiểm tra, sắp xếp và xem lại các tập tin hình ảnh; Chuyển đổi dữ liệu số; Nhận dạng chữ viết (OCR); Tạo ra tài liệu phức hợp.

Quá trình số hóa tài liệu tại thư viện ĐHKHXH&NV có đặc điểm là phải xử lý số hoá với nhiều loại font chữ và chất lượng tài liệu xấu. Do đó, phần mềm nhận dạng chữ viết tiếng Việt hiện nay VNDORC của Viện Khoa học và Công nghệ Việt Nam (để khôi phục nội dung text của tài liệu từ các file ảnh) chưa đáp ứng được yêu cầu nhận dạng với độ chính xác cao. Vì vậy, nếu sử dụng bản text thu được qua nhận dạng thì không tránh khỏi sai sót và khác biệt so với bản gốc (nếu chất lượng tài liệu quá xấu chỉ nhận dạng được khoảng dưới 50%). Trong trường hợp này Thư viện khắc phục bằng cách phục vụ tài liệu dạng file ảnh.

Thư viện đang trong thời gian thử nghiệm định dạng phức hợp (ảnh và text) là định dạng tiên tiến nhất mới được ứng dụng ở các thư viện số trên thế giới trong vài năm gần đây. Định dạng này cho phép người dùng sử dụng file phức hợp (cả ảnh và text), giúp người dùng đọc được dạng hình ảnh mà vẫn tìm kiếm, trích được nội dung text của tài liệu, cung cấp cho người dùng những tính năng ưu việt của tư liệu số.

Thời gian đầu Thư viện thực hiện số hóa bằng máy Scan thông thường nên tốc độ số hóa rất chậm, chất lượng kém. Từ khi được trang bị máy Scan chuyên dụng OMISCAN với phần mềm nhận dạng Vndorc, Omipage, trong quá trình số hóa có những thuận lợi cơ bản:

- Máy scan OMISCAN có tốc độ cao gấp 4 lần máy thường, có nhiều chức năng như hiệu chỉnh kích cỡ, hiệu chỉnh độ sáng tối, độ tương phản, tẩy, xóa phần đen, làm trắng trang văn bản, dàn trang, điều chỉnh độ dày của sách; có khả năng Scan được tài liệu dày mà không cần phải tháo rời tài liệu, Scan được tài liệu khổ lớn A_2.

- Phần mềm Omipage nhận dạng nhanh, chính xác tài liệu tiếng Anh.

- Những tài liệu mới được scan và nhận dạng rất nhanh.

Những khó khăn thường gặp trong thực tiễn số hóa tài liệu là:

- Khi Scan những trang sách quá mỏng: chữ in của trang trước sẽ hiện lên ở trang sau.

- Đối với những tạp chí bị đóng quá chặt và sát gáy khi Scan không thấy được hết trang tạp chí.

- Đối với tài liệu cũ khi Scan chữ bị mờ.

- Đối với giấy bị đen, vàng ố chất lượng Scan rất kém.

- Đối với những cuốn sách tờ bị rời chi phí thời gian Scan cao hơn sách còn nguyên gáy.

- Những cuốn sách quá dày: phải chỉnh sửa lại trang Scan do bị lệch trong quá trình Scan.

Những khó khăn thường gặp trong quá trình nhận dạng là:

- Tài liệu bị mờ chữ, không rõ chữ.

- Tài liệu có nhiều cột.

- Tài liệu có nhiều bảng và hình.

- Phần mềm nhận dạng VNDORC chưa hoàn chỉnh (là khó khăn lớn nhất).

2. Một số bài học kinh nghiệm trong tiến trình thực hiện số hoá nguồn tư liệu

Từ quá trình số hoá nguồn tư liệu, Chúng tôi rút ra một số bài học kinh nghiệm sau :

- Phải xác định rõ các nhiệm vụ của trang thiết bị và dây chuyền công nghệ phục vụ hoạt động số hoá, ví dụ: chuyển nội dung tài liệu hiện có thành nội dung số, cho phép các thao tác xử lý nội dung: tìm kiếm, trích dẫn, truy cập trực tuyến, đảm bảo giữ nguyên vẹn hình thức ấn phẩm; Tạo nhiều định dạng khai thác: ảnh ấn phẩm lưu trữ trên hệ thống máy tính, nội dung số hoá, vi phim; Số hoá được các dạng tài liệu phi văn bản có trong thư viện: băng đĩa, ảnh ....

- Phải xác định rõ các yêu cầu về phần mềm số hóa và quản lý kho tư liệu số, ví dụ: phải đảm bảo thực hiện các thao tác nghiệp vụ số hóa tài liệu như scan, nhận dạng tài liệu, biên mục, quản lý phần nội dung của ấn phẩm, gán quyền truy nhập cho người đọc; Phải đảm bảo các tính năng tìm kiếm theo các thông tin biên mục, tìm kiếm toàn văn, truy nhập đến nội dung ấn phẩm; Quản lý kết quả của quá trình số hóa tài liệu; Nhận dạng được các tài liệu cũ trước năm 1975 với độ chính xác trên 80%; Tạo và quản lý được tài liệu số hóa tích hợp: tài liệu chứa đồng thời cả hình ảnh và nội dung text của tài liệu, cho phép người dùng đồng thời đọc bản gốc và tìm kiếm được nội dung tài liệu; Quản lý thông tin thư mục của tài liệu; Đảm bảo tích hợp với hệ thống quản lý nghiệp vụ thư viện, tạo điều kiện thuận lợi cho người sử dụng trong các thao tác tra cứu thông tin; Quản lý được nội dung của tài liệu; Tìm kiếm chính xác và đầy đủ theo nội dung của tài liệu, kể cả trong trường hợp chất lượng nhận dạng chỉ đạt 80%.

- Việc số hóa tài liệu phải đảm bảo tính chính xác, trung thực của nội dung số hóa thu được, không làm ảnh hưởng đến ấn phẩm nguyên gốc.

- Phải yêu cầu nhà cung cấp phần mềm số hóa bảo đảm chuyển giao công nghệ Số hóa tài liệu và quản trị tư liệu số, đào tạo cán bộ thư viện trên thực tế công việc số hóa (on-the-job training) để họ làm chủ và nắm bắt kỹ năng tạo ra và quản trị tư liệu số.

- Việc quản lý tư liệu số đòi hỏi phải nghiên cứu kỹ Luật Sở hữu trí tuệ và luật Bản quyền tác giả để không vi phạm bản quyền.

Đối với việc tải xuống tài liệu điện tử trên mạng Internet nhằm mục đích xây dựng các bộ sưu tập số về các chủ đề phục vụ học tập,

nghiên cứu, thư viện chia thành hai loại:

+ Nguồn tài liệu điện tử đã trả phí (ví dụ các tạp chí online do thư viện mua, được phép tải xuống): là tài liệu có chất lượng tốt về nội dung cũng như về hình thức, vấn đề bản quyền của tài liệu được đảm bảo, có thể tải xuống được một cách đầy đủ nhất nội dung của tài liệu.

+ Nguồn tài liệu điện tử miễn phí.

Trong quá trình tải xuống tài liệu điện tử trên mạng Internet, việc sử dụng các công cụ hỗ trợ như phần mềm Internet Download Manager (IBM), Download Accelerator Plus giúp tải xuống nhanh chóng, tải một loạt tài liệu về cùng một lúc và cùng một lúc tải xuống được nhiều kiểu file khác nhau (file .pdf, .tif, .prc, .lit, .mp3, .html v.v...). Phần mềm IDM có hỗ trợ giao thức FTP trong quá trình tải xuống, phần mềm DAP không có hỗ trợ giao thức FTP, tuy nhiên DAP có hỗ trợ tìm kiếm các trang khác khi đường siêu liên kết hiện tại bị mất kết nối.

Khi tải xuống tài liệu điện tử trên Internet Thư viện đã gặp một số khó khăn như sau:

- Muốn tải xuống được tài liệu điện tử có trả phí Thư viện phải thông qua địa chỉ IP tĩnh và proxy của Đại học Quốc gia TP. Hồ Chí Minh phân giải thì mới xem được toàn văn và tải xuống về được.

- Đối với tài liệu điện tử miễn phí: nhiều lúc nội dung tài liệu không đầy đủ, hoặc bị thiếu file, chất lượng nội dung và hình thức thấp. Vấn đề bản quyền không được đảm bảo vì ai cũng có thể tải xuống được.

- Tốc độ đường truyền của mạng quá chậm nên những file tài liệu có dung lượng lớn sẽ rất khó tải về; đường siêu liên kết phức tạp, đường siêu liên kết không đúng hoặc đường siêu liên kết nhiều tầng, nhiều lớp cũng khó tải xuống tài liệu.

- Cấu hình máy của Thư viện còn thấp so với yêu cầu, có lúc tải xuống tài liệu về rồi nhưng máy tính không đủ dung lượng để chứa tài liệu.

- Khi càng có nhiều người cùng click vào một file cần tải thì tốc độ dành cho file đó sẽ tăng cao và ngược lại thì file cần tải sẽ chậm hoặc không thể tải về máy của mình được.

- Khi sử dụng các công cụ phần mềm hỗ trợ tải xuống có thể gặp trường hợp không muốn tải xuống tài liệu bằng phần mềm đó nhưng các đường siêu liên kết vẫn cứ chạy ra nên mất nhiều thời gian.

Hiện tại, hướng giải quyết của Thư viện về việc tải xuống tài liệu điện tử trên Internet là máy tính dùng để tải xuống có kết nối "Băng thông rộng" và có thời gian online khoảng 6-10 tiếng trở lên để đẩy nhanh tiến độ tải xuống.

3. Kết luận

Việc xây dựng thư viện số là một quá trình lâu dài. Tổ chức nguồn tài nguyên thông tin điện tử của Thư viện Đại học KHXH&NV TP. Hồ Chí Minh đang ở giai đoạn đầu cho một thư viện số trong tương lai. Những kết quả đạt được tuy còn ít ỏi nhưng hiệu quả khai thác, sử dụng các bộ sưu tập tư liệu số toàn văn theo các chủ đề phục vụ thiết thực cho công tác nghiên cứu khoa học và đào tạo của trường bằng cả nguồn tài liệu nội sinh và trên mạng đã chứng minh rằng xây dựng thư viện số là xu thế tất yếu.

Tài liệu tham khảo

Dự án nâng cao năng lực cung ứng thông tin KHXH&NV phục vụ nghiên cứu khoa học.- Tp.HCM.: ĐHKHXH&NVTP.HCM, 2006.-49 tr.