THẾ GIỚI THƯ VIỆN SỐ

ThS. NGUYỄN MINH HIỆP
GĐ Thư viện ĐH Khoa học Tự nhiên
ĐHQG TP. Hồ Chí Minh

Thư viện và thư viện số

Thư viện số là gì? Xây dựng thư viện số là xây dựng một cơ sở thư viện hay đơn giản chỉ hình thành một bộ phận công nghệ trong một cơ sở? Đây là điều chúng ta cần quán triệt trước khi bắt tay vào xây dựng thư viện số.

Ngày nay vẫn còn nhiều người cho rằng thư viện là một nơi yên tĩnh trong đó sách được cất giữ và người ta đánh giá thư viện theo tiêu chí số lượng sách được cất giữ nhiều hay ít. Đối với những người quản thủ thư viện có chuyên môn thì thư viện là một cơ sở có tổ chức để bảo quản tài liệu, sưu tập và để truy cập đến những thư viện khác; không những chỉ sách mà còn có phim ảnh, băng đĩa âm thanh, mẫu vật thực vật, sản phẩm văn hoá, vv… Đối với nhà nghiên cứu, thư viện là một mạng lưới cung cấp việc truy cập đến tri thức nhân loại được lưu giữ khắp mọi nơi. Nhiều sinh viên khoa học và công nghệ ngày nay trên thế giới thì cho rằng thư viện chính là World Wide Web. Đây là một quan niệm không đúng mặc dầu ngày nay Web là công nghệ quan trọng của thư viện. Sự khác nhau giữa thư viện số với World Wide Web thể hiện ở chỗ Web thiếu hẳn những đặc điểm quan trọng của việc sưu tầm và tổ chức thông tin; trong khi thư viện số ngày càng hoàn thiện việc tổ chức để người sử dụng tự hình thành tri thức với phương châm ***"Thư viện số là nơi sử dụng công nghệ để chuyển câu hỏi thành câu trả lời".***

Thư viện số không thực sự là một "thư viện được số hoá". Xây dựng thư viện số là xây dựng phương thức mới, công nghệ mới trong việc xử lý thông tin - tri thức. Đó là bảo quản, sưu tầm, tổ chức, quảng bá, và truy cập thông tin hay nói chính xác hơn là tri thức, tức là thông tin có ý nghĩa và hữu ích. Do đó, một thư viện số được xem như là nơi trình bày những bộ sưu tập thông tin có tổ chức. Bộ sưu tập tập trung vào đối tượng số hóa, bao gồm văn bản, hình ảnh và âm thanh cùng với phương thức truy cập, truy hồi, chọn lọc, tổ chức, bảo trì sưu tập đó. Sưu tập do chuyên gia thư viện tạo nên. Phần mềm thư viện số, chẳng hạn như "Hòn đá xanh – Greenstone" hỗ trợ người sử dụng tìm kiếm sưu tập, cũng như hỗ trợ cho chuyên gia thư viện xây dựng và bảo trì sưu tập có hiệu quả.

Đối với một thư viện truyền thống, điều quan trọng là việc bổ sung nguồn tài nguyên ngày càng nhiều trên giá kệ trong kho thư viện; nhưng ngày nay thông tin về những nguồn tài nguyên

đó chứa trong mục lục thư viện là quan trọng hơn. Chúng ta gọi những thông tin đó là *metadata* hay *siêu dữ liệu* – dữ liệu về dữ liệu – và đây là khái niệm nổi bật nhất trong thư viện số.

Sự thay đổi bộ mặt thư viện

Thư viện là kho tri thức của xã hội; có người còn cho rằng thư viện là đền đài của văn hoá và sự uyên thâm. Được hình thành trong thời kỳ nông nghiệp thống trị trong tư duy của nhân loại, thư viện đã trãi nghiệm qua một cuộc hồi sinh với việc phát minh ngành in trong thời kỳ Phục hưng, và thực sự bắt đầu khởi sắc khi cuộc cách mạng công nghiệp bùng phát với hàng loạt những phát minh cơ giới hoá quy trình in ấn.

Lịch sử thư viện đã trãi qua hơn 25 thế kỷ. Hình ảnh thư viện của thời xa xưa được hình dung như là một cơ sở vững chắc trong đó chứa hàng ngàn phiến đá khổng lồ được khắc chữ - thường được gọi là "rừng bia". Qua nhiều năm cùng với sự tiến hoá của nhân loại, con người càng tiến bộ trong nhận thức và thư viện ngày càng được phát triển. Giai đoạn *Quản lý tư liệu* đã trãi qua một thời gian dài theo sự phát triển đó. Cho đến một lúc, cũng xuất phát từ ý định ban đầu là làm tốt công việc lưu trữ và bảo quản, thư viện đã chú trọng đến việc xem người sử dụng là trung tâm, với sự nhấn mạnh đến việc trao đổi thông tin. Điều này cũng đồng thời để đáp ứng yêu cầu thông tin ngày càng gia tăng. Giai đoạn *Quản lý thông tin* được xem như bắt đầu. Và chúng ta sẽ nhận thức được rằng để xây dựng thư viện số là ta đã bắt đầu bước qua một giai đoạn phát triển mới của thư viện: Giai đoạn *Quản lý tri thức.*

Thư viện cổ đại chỉ hữu ích đối với một thiểu số những người biết chữ và bị giới hạn trong một tầng lớp, giai cấp theo điều kiện xã hội. Hoạt động Thư viện công cộng được bắt đầu phát triển trong thế kỷ 19. Nhưng vẫn là những thư viện đóng: sách được xếp theo kích cỡ trong những kho kín trong thư viện, độc giả chỉ tiếp cận với thủ thư ở quầy để yêu cầu mượn sách. Hầu hết những thư viện trong lục địa châu Âu đã áp dụng phương thức này trong một thời gian dài. Đến thế kỷ 20 một số quản thủ thư viện nhận thức được tiện ích của việc cho độc giả tiếp cận với kho sách đã đề xuất phương thức phục vụ kho mở với tài liệu được xếp theo môn loại. Phương thức này được bắt đầu áp dụng và nhanh chóng phát triển trong những quốc gia nói tiếng Anh hồi đó.

Ngày nay chúng ta đang đứng trước ngưỡng cửa của thư viện số. Cuộc cách mạng thông tin không những cung cấp năng lực công nghệ hướng đến thư viện số, mà còn đáp ứng một nhu cầu chưa từng có về lưu trữ, tổ chức, và truy cập thông tin. Nếu thông tin là tiền tệ trong nền kinh tế tri thức, thư viện số sẽ là ngân hàng, nơi được đầu tư. Quả vậy, Goethe – Đại thi hào Đức đã từng nói *"đến thư viện giống như đi vào một nơi phô hiện sự giàu sang tột đỉnh, ở đó lãi suất hậu hĩnh đang được thanh toán một cách thầm lặng"*.

Như chúng ta đã thấy, xây dựng thư viện số không phải là xây dựng một cơ sở thư viện mà là xây dựng một nền tảng công nghệ để tiến đến giai đoạn quản lý tri thức. Nền tảng công nghệ này được xây dựng trên một cơ sở thư viện mà cơ

sở thư viện này được xây dựng trên một nền tảng thư viện truyền thống. Vì thế muốn xây dựng thư viện số, trước hết phải củng cố nền tảng thư viện truyền thống: tuân thủ những tiêu chuẩn nghiệp vụ căn bản và thay đổi một số giá trị cũ cho phù hợp với việc ứng dụng công nghệ mới.

Hình thành Thư viện số

Hình thành thư viện số bằng cách sử dụng công nghệ để tạo lập một cách nhanh chóng những bộ sưu tập thông tin có tổ chức và làm tăng năng lực truy tìm và lướt tìm của người sử dụng. Phần mềm nguồn mở thư viện số "Hòn đá xanh – Greenstone" sẽ cung cấp phương cách xây dựng những bộ sưu tập và tổ chức thông tin để phục vụ trên Internet.

Một *sưu tập – collection* thông tin bao gồm nhiều tài liệu dưới nhiều hình thức. Một *tài liệu – document* là thông điệp mang thông tin dưới hình thức điện tử. Tài liệu là đơn vị cơ sở từ đó sưu tập thông tin được xây dựng, mặc dù chúng có thể có những cơ sở hạ tầng và những tập tin kết hợp riêng. Nói chung tài liệu bao gồm văn bản, mặc dù chúng có thể là tập tin hình ảnh, âm thanh hay video. Một sưu tập có thể chứa nhiều loại tài liệu khác nhau. Mỗi sưu tập cung cấp một giao diện đồng nhất qua đó tất cả tài liệu có thể được truy cập – mặc dù cách mà tài liệu đó hiển thị sẽ tuỳ thuộc vào phương tiện và hình thức của tài liệu đó. Một *thư viện* nói chung bao gồm nhiều sưu tập khác nhau, mỗi sưu tập tổ chức mỗi khác – mặc dù có sự giống nhau nổi bật trong phương cách sưu tập hiển thị.

Sưu tầm thông tin và tổ chức phục vụ rộng rãi trên Internet hoàn toàn khác hẳn với việc đơn thuần là trình bày thông tin trên Web. Sưu tập trở nên có thể được bảo hành, truy tìm, lướt tìm. Trước khi trình bày, mỗi sưu tập trải qua một quá trình hình thành, một khi được xây dựng xong, sưu tập hoàn toàn tự động. Quá trình này tạo nên tất cả những cấu trúc được dùng trong thời gian truy cập sưu tập. Việc truy tìm dựa trên những chỉ mục khác nhau bao gồm trong toàn văn và siêu dữ liệu. Việc lướt tìm dựa trên những siêu dữ liệu và thông tin được tóm tắt từ những toàn văn của tài liệu. Cấu trúc hỗ trợ cho cả hai việc truy tìm và lướt tìm được tạo ra trong suốt quá trình xây dựng sưu tập. Một khi tài liệu mới xuất hiện, nó có thể được kết nạp vào sưu tập bằng cách tái xây dựng.

Sở hữu trí tuệ và bản quyền

Trong một thư viện truyền thống, quyền sở hữu tài liệu là quan trọng; nhưng trong lĩnh vực lưu hành tài nguyên điện tử, quyền sở hữu trí tuệ, cụ thể là quyền tác giả hay bản quyền là quan trọng hơn.

Sưu tầm thông tin và làm cho thông tin đó trở nên phổ biến hơn đối những người khác là một điều liên quan đến vấn đề xã hội, và những người xây dựng thư viện số phải am hiểu quyền sở hữu trí tuệ để hành động một cách có trách nhiệm và đúng luật xung quanh những ứng dụng cụ thể của họ.

Thư viện số có thể làm cho việc truy cập trở nên rộng rãi hơn thư viện truyền thống. Và chính điều này đã nảy sinh ra nhiều vấn đề: truy cập thông tin

trong thư viện số, nói chung ít bị kiểm soát hơn truy cập sưu tập in ấn trong thư viện thường. Đưa thông tin vào thư viện số là có khả năng làm cho thông tin đó trở nên phổ biến ngay đối với một số lượng độc giả hầu như vô hạn.

Đối với người sử dụng, thông tin trên thế giới có thể truy cập bất cứ nơi đâu. Đối với tác giả, một độc giả có trình độ hơn có thể tiếp cận được nhiều thông tin hơn trước. Và đối với nhà xuất bản, nhiều thị trường mới mở ra vượt quá mọi giới hạn địa lý. Nhưng có một nghịch lý. Tác giả và nhà xuất bản hỏi có bao nhiêu cuốn sách sẽ được bán nếu những thư viện số nối mạng khiến cho bản điện tử của cuốn sách đó được truy cập rộng rãi trên thế giới? Cơn ác mộng cho họ khi câu trả lời là *một.* Có bao nhiêu sách sẽ được xuất bản trực tuyến nếu toàn bộ thị trường có thể bị phá hủy bởi việc bán một bản điện tử cho một thư viện công cộng?

Nhà xuất bản sẽ phản ứng như thế nào trước tình huống này? Lời cảnh báo cho người sử dụng là nhà xuất bản sẽ thông qua công nghệ và phương tiện luật pháp để thực thi chính sách hạn chế việc truy cập đến những thông tin họ bán – chẳng hạn như, bằng cách hạn chế việc truy cập của người mua, cho dùng thử trong một thời gian hạn định hoặc đánh thuế khi sử dụng quá hạn.

Sở hữu một cuốn sách chắc chắn không phải là xác lập được quyền sở hữu đối với tài liệu đó theo nghĩa của bản quyền. Mặc dù có nhiều bản của một tài liệu nhưng chỉ có một bản quyền. Điều này không chỉ áp dụng cho bản in mà cả cho bản điện tử, dù được số hoá từ bản in hay được tạo nên dưới dạng điện tử từ đầu. Khi mua một cuốn sách, ta có thể bán lại, nhưng chắc chắn không mua quyền tái phân phối. Quyền đó tùy thuộc vào bản quyền.

Ai làm chủ một tác phẩm cụ thể? Bản quyền đầu tiên là của người sáng tác trừ phi tác phẩm được thuê sáng tác. Trong trường hợp này bản quyền thuộc về cơ quan hay tổ chức thuê theo hợp đồng; bản quyền có thể được sang nhượng hay chuyển cho một đơn vị khác thông qua một hợp đồng cụ thể, được thực hiện bằng văn bản do người chủ ký tên.

Luật bản quyền là phức tạp. Tình trạng luật pháp đối với tập tin máy tính và tài liệu cụ thể được xuất bản trên World Wide Web lại mù mờ. Muốn xây dựng thư viện số thì phải cần số hoá tài liệu. Chúng ta phải làm như thế nào để tránh vi phạm bản quyền? Trước hết chúng ta phải xem xét:

- Nếu tác phẩm được số hoá ở trong miền (domain) công cộng thì chúng ta không phải xin phép ai hết. Dĩ nhiên kết quả số hoá của chúng ta cũng không được bảo vệ bản quyền, trừ phi kết quả của ta nhiều hơn bản gốc;
- Nếu tài liệu được tặng cho cơ sở của ta để số hoá và người tặng có bản quyền, thì chúng ta tiến hành số hoá, tuy nhiên cần phải yêu cầu người tặng cung cấp cho mình quyền được số hoá – có thể bằng một mẫu giấy có ghi *"quyền sử*

dụng tác phẩm với bất kỳ mục đích chung của cơ sở, dưới bất kỳ phương tiện nào".

Nếu ta muốn số hoá tài liệu mà không rơi vào hai trường hợp trên thì ta phải cân nhắc thử việc số hoá của chúng ta có phải là một việc làm có lợi ích chung mà không xâm phạm quyền lợi của người khác. Đây là một điều khó về mặt pháp lý. Cuối cùng nếu chúng ta không chắc chắn với điều cân nhắc trên thì ta phải tiến hành xin phép để được cấp phép thực hiện số hoá.

Tóm lại để tiến hành xây dựng thư viện số, ta phải lưu ý đến vấn đề bản quyền. Những người thực hiện phải cam kết hiểu biết đầy đủ về bản quyền và nhận thức sâu sắc rằng giấy phép là rất cần thiết để chuyển đổi tài liệu không thuộc trong miền công cộng.

Sưu tầm thông tin từ Web

Tất cả những vấn đề về bản quyền đều có một tác động thực tiển và tức thì vào thư viện số. Vì rằng thư viện số là sưu tập thông tin có tổ chức. Web thì đầy những thông tin không được tổ chức. Tải xuống từng phần để tổ chức thông tin đưa vào trong sưu tập có định hướng và làm cho những tài liệu đó hữu ích hơn, thông tin có ý nghĩa hơn đối với những đối tượng cụ thể là một phạm vi áp dụng chính yếu cho thư viện số.

Bộ máy tìm kiếm - Search engines, một trong những dịch vụ được dùng rộng rãi nhất trên Internet, là một minh hoạ cho việc sưu tầm thông tin từ Web. Người ta sử dụng phần mềm "robot" để tải xuống một cách liên tục phần lớn thông tin từ Web và tạo chỉ mục cho những thông tin đó. Hầu hết những thư viện số có mục tiêu cung cấp việc truy tìm và lướt tìm cao cấp hơn và hiệu quả hơn search engines. Trong thư viện số, tài liệu là có khả năng tìm thấy ở trong thư viện nhiều hơn là sản phẩm được tạo nên trên Web site.

Nguồn tài liệu

Nguồn tài liệu của mỗi thư viện mỗi khác tuỳ theo chính sách phát triển sưu tập và loại hình thư viện – thư viện Quốc gia, thư viện công cộng, thư viện đại học, thư viện chuyên ngành, và thư viện trường học. Ngoài ra, khi bắt tay xây dựng thư viện số cần quan tâm đến sự khác nhau về tự động hóa – việc ứng dụng công nghệ sẽ không đồng đều trong nhiều thư viện. Do đó vấn đề xác định mục tiêu là phải được xem xét trước, qua đó xác định nguyên tắc sưu tầm tài liệu. Có ba kịch bản xây dựng thư viện số liên quan đến việc hình thành nguồn tài liệu:

1. Xây dựng thư viện số trên cơ sở chuyển đổi một thư viện hiện hữu – số hoá tài liệu thư viện.
2. Xây dựng thư viện số bằng cách thiết lập một bộ sưu tập điện tử bên cạnh sưu tập in ấn;
3. Xây dựng thư viện số bằng cách cung cấp một cổng thông tin vào một sưu tập tài liệu điện tử đang hiện hữu trên Web.

Những kịch bản này không phải là độc nhất mà cũng không phải toàn diện, trong thực tế chúng ta thường gặp phải sự trộn lẫn. Chúng ta cần phải xác định rõ để tập trung giải quyết vấn đề trước khi tiến hành dự án xây dựng thư viện số.

Chuyển đổi một thư viện hiện hữu

Chuyển một thư viện thường sang dạng số là một cách làm đầy tham vọng và đắt tiền. Số hóa nội dung của một sưu tập in ấn thường là một công việc đồ sộ và chán nản. Thế nhưng có người cho rằng muốn xây dựng thư viện số thì phải số hoá toàn bộ tài liệu có trong thư viện. Đây là một quan niệm hết sức sai lầm, thực ra đây là một ảo tưởng vì thực tế không có một thư viện nào trên thế giới có đủ nhân lực tài lực để thực hiện công việc này.

Mặc dù thư viện số có ba thuận lợi chính hơn hẳn thư viện thường là: Dễ dàng truy cập từ xa, nâng cao hơn năng lực truy tìm và lướt tìm, và phục vụ với tính cách là một dịch vụ mang đến giá trị gia tăng cho người sử dụng, tuy nhiên trước khi bắt tay vào việc số hóa một sưu tập chúng ta cần phải cân nhắc thật kỹ lưỡng liệu có thật cần thiết để thực hiện không.

Một khi chúng ta đã quyết định tiến hành thì vấn đề then chốt là xác định độ ưu tiên của tài liệu để chuyển đổi. Tài liệu thư viện có thể chia làm ba loại: sưu tập đặc biệt và tài liệu một bản, chẳng hạn sách quý hiếm và bản viết tay; tài liệu được sử dụng cao, thường xuyên được yêu cầu cho giảng dạy và nghiên cứu; và tài liệu có mức độ sử dụng thấp bao gồm tài liệu nghiên cứu ít dùng thường xuyên.

Có sáu nguyên tắc được xác định nhằm chọn tài liệu để số hóa hướng đến việc phát triển sưu tập thư viện số:

1. *Tính hữu dụng*: Hữu dụng là lý do cơ bản trước tất cả mọi quyết định phát triển sưu tập. Tài liệu có tần suất sử dụng cao (như giáo trình, tài liệu tham khảo mà các giáo viên thường yêu cầu tất cả sinh viên tìm đọc);
2. *Nhu cầu nội bộ*: Sưu tập nội bộ được xây dựng để phục vụ nhu cầu nội bộ và chi phí cho tài nguyên nội bộ phải được thuyết minh vì lợi ích nội bộ – chẳng hạn như đối với thư viện đại học, yêu cầu học tập, giảng dạy, và nghiên cứu là ưu tiên;
3. *Tài liệu mới*: Mặc dù sưu tập cũ mang tính lịch sử là cần thiết cho nghiên cứu, nhưng tài liệu mới vẫn ưu tiên hơn;
4. *Tài liệu liên quan đến bản gốc*: Những tài liệu mà người muốn tìm hiểu không thể tiếp cận được bản gốc (ví dụ các văn bản viết tay – "manuscript" của các nhà thơ, nhà văn, các nhà chính trị, hoặc các bản tuyên ngôn có chữ ký cuả các lãnh tụ như bản tuyên ngôn độc lập của Hoa Kỳ hiện có tại Thư viện Quốc hội Hoa Kỳ, vv…). Trên thực tế, còn có rất nhiều thể loại viết tay trên những chất liệu khác nhau. Việc số hoá các bản viết tay đó tạo điều kiện tiếp cận thuận lợi hơn cho các nhà nghiên cứu;
5. *Tài liệu quý hiếm*: Tài liệu quí hiếm, lâu năm, độc giả không thể trực tiếp sử dụng, dễ hư hỏng – chẳng hạn như tài liệu chữ Nôm trên giấy bồi;
6. *Chuyển đổi nhận thức*: Ngày càng có nhiều thông tin chuyển

sang dạng số. Tài liệu giúp người sử dụng chuyển đổi nhận thức để làm quen việc sử dụng dạng thông tin này là ưu tiên.

Chúng ta cần phải cân nhắc mức độ ưu tiên đối với những nguyên tắc trên trong việc chọn tài liệu để số hóa.

Xây dựng một sưu tập mới

Xây dựng thư viện số bằng cách thiết lập những sưu tập tài liệu mới là phổ biến hơn việc số hoá thư viện hiện hữu. Để xây dựng một sưu tập mới ta thường phải đối mặt với cả hai loại tài liệu: tài liệu đã ở dạng điện tử rồi và tài liệu in ấn cần phải số hoá. Nếu toàn bộ tài liệu ở dạng điện tử thì công việc hết sức dễ dàng, ngay cả việc sưu tầm, tổ chức tập tin và chuyển đổi dạng thức; công việc này rẽ hơn nhiều so với việc số hoá tài liệu.

Vấn đề là chúng ta phải xác định metadata. Có được metadata cần thiết và chuyển đổi qua dạng điện tử thường là công việc chính trong vấn đề xây dựng sưu tập. Khi số hoá một thư viện hiện hữu thì metadata có sẵn rồi, nhưng khi xây dựng sưu tập mới việc xác định metadata là phức tạp hơn nhiều.

Thư viện ảo

Một loại thư viện số khác cung cấp một cổng thông tin nhằm vào thông tin điện tử ở nơi khác ngoài thư viện. Loại này đôi khi được gọi là *thư viện ảo* để nhấn mạnh rằng đây là thư viện mà bản thân không chứa nội dung. Những quản thủ thư viện đã dùng thuật ngữ này cách đây hơn mười năm để chỉ một loại thư viện chuyên cung cấp việc truy cập thông tin điện tử thông qua những chỉ điểm – pointers.

Như chúng ta đã đề cập ở trên, Web thiếu những đặc điểm sưu tầm và tổ chức thông tin như của thư viện số. Nhưng nó chứa một lượng khổng lồ thông tin có ích. Người ta sàng lọc thông tin đó và tổ chức lại để xây dựng những sưu tập phụ của Web. Thư viện số xây dựng theo cách này sẽ tạo ra những phân lớp quan trọng cung cấp việc truy cập thông tin đã có sẵn trên Web. Giá trị học thuật và giáo dục của nguồn tài liệu càng cao thì khối lượng thời gian của chuyên gia đầu tư cho việc mô tả càng lớn.

Cổng thông tin thường tập trung một đề tài chuyên biệt hoặc chú trọng đến đối tượng người sử dụng cụ thể. Chẳng hạn như đối với một thư viện đại học, công nghệ cổng thông tin tích hợp cung cấp những công cụ mới cho người quản lý thư viện, giảng viên, các nhà nghiên cứu khoa học và các chuyên gia công nghệ thông tin tạo nên một môi trường dạy, học và nghiên cứu với bốn (4) mục tiêu sau:

1. Mục tiêu thứ nhất là cung cấp kho tài nguyên trung tâm cho các tài nguyên được số hoá, hỗ trợ việc chia sẻ các nguồn tài nguyên và làm nơi bảo tồn các công trình số hoá này.
2. Mục tiêu thứ hai là cung cấp hệ thống thông tin số có khả năng tổ chức, phân loại, biên mục, chú dẫn và tổng hợp các tài nguyên theo các chuẩn IMS, DC, … trên nền hỗ trợ của siêu dữ liệu XML.

3. Mục tiêu thứ ba là tạo nên giao diện duy nhất và thống nhất cho các giảng viên, sinh viên và nghiên cứu sinh cùng truy cập, tra cứu, tìm kiếm các tài nguyên số hoá này để hỗ trợ nguồn thông tin cho các bài giảng, công việc học tập, tham khảo của sinh viên và công tác nghiên cứu khoa học.
4. Mục tiêu thứ tư nhằm đảm bảo việc thống nhất các quy trình bổ sung thông tin, cơ sở dữ liệu, bộ sưu tập điện tử cũng như việc lưu trữ và quản lý tập trung các nguồn tài nguyên này giữa thư viện và các khoa. Đây là khả năng tập trung và chia sẻ hợp lý các lực lượng lao động giữa thư viện và các khoa.

Số hóa tài liệu

Một trong những công việc đầu tiên mà ta quan tâm khi bắt đầu xây dựng một thư viện số là liệu ta có cần phải số hoá tài liệu hiện hữu trong thư viện hay không. Số hoá là tiến trình chuyển tài liệu thư viện truyền thống, cụ thể là sách và văn bản sang dạng điện tử và lưu trữ trên máy tính.

Có hai giai đoạn trong tiến trình số hoá. Giai đoạn đầu *quét hình – scanning* cho ra sản phẩm số hoá dạng hình. Giai đoạn hai cho ra một sản phẩm dạng số hoá văn bản bằng môt tiến trình gọi là *nhận dạng ký tự quang học –OCR (Optical Character Recognition)*. Trong nhiều hệ thống thư viện số, tài liệu chỉ ở giai đoạn đầu, nghĩa là những gì độc giả thấy chỉ là hình ảnh. Giai đoạn hai là cần thiết đối với những văn bản có chỉ mục được tạo ra một cách tự động để độc giả có thể định vị bất kỳ một tổ hợp từ nào hay đối với bất kỳ kỹ thuật trích dẫn metadata tự động được định trước, chẳng hạn xác định nhan đề của tài liệu bằng cách tìm trong văn bản.

Những yếu tố tổ chức tài liệu

Nếu tài liệu là những khối xây dựng căn bản của thư viện số, thì markup và metadata là những yếu tố tổ chức. Markup được dùng để chỉ rõ cấu trúc của tài liệu riêng lẻ và kiểm soát phương thức trình bày cho người sử dụng. Metadata được dùng để xúc tiến việc truy cập đến những phần thích hợp của sưu tập qua việc truy tìm và lướt tìm. Một phần công việc của markup là để xác định metadata.

Có một sự khác biệt quan trọng giữa metadata *hiện – explicit* với metadata *trích –extracted*. Metadata hiện được xác định bởi con người sau khi xem xét cẩn thận và phân tích tài liệu. Tạo nên một dẫn mục của mục lục thư viện truyền thống (biểu ghi MARC) là một công việc của một biên mục viên được đào tạo kỹ lưỡng: thường phải mất từ một đến hai tiếng đồng hồ để tạo nên một dẫn mục mới. Do đó để có một tập hợp lớn biểu ghi thư tịch dạng MARC thì phải đầu tư rất nhiều công sức và tiền của. Tác giả Ian H. Witten trong cuốn sách *How to Build a Digital Library* đã cho biết vào năm 1997 Thư viện Quốc hội Hoa Kỳ đã biên mục gần 300.000 biểu ghi thư tịch MARC, tiêu tốn vào khoảng 25 triệu đô la. OCLC, một cơ quan biên mục tập trung ở Mỹ, có vào khoảng 34 triệu biểu ghi MARC – biểu

hiện một đầu tư vào khoảng 30.000 năm nhân công!

Metadata trích có được một cách tự động từ nội dung tài liệu. Công việc này thường khó thực hiện chính xác. *Khai thác văn bản – text mining*, được định nghĩa như một tiến trình phân tích văn bản để trích thông tin hữu ích cho mục đích cụ thể, là một đề tài nghiên cứu nóng bỏng hiện nay.

Markup

Để hiểu được công nghệ mới, chúng ta cần biết đôi chút về lịch sử, công việc của những con người thầm lặng sau "hậu trường". Một consortium nổi tiếng W3C (World Wide Web Consortium) bao gồm các nhà nghiên cứu và chuyên gia CNTT, trong suốt thập niên 1970 và 1980 với một nỗ lực tiên khởi để phối hợp một dạng thức dữ liệu có thể trao đổi toàn cục với các khả năng lưu trữ thông tin phong phú đã tạo nên một hệ thống khái quát hoá việc đánh dấu cấu trúc được phát triển gọi là Ngôn ngữ đánh dấu khái quát hoá tiêu chuẩn hay SGML – Standard Generalised Markup Language; được thông qua như là một tiêu chuẩn quốc tế ISO vào năm 1986. SGML không phải là một ngôn ngữ đánh dấu mà là siêu ngôn ngữ (ngôn ngữ nói về ngôn ngữ khác) miêu tả dạng markup. Ngôn ngữ này phổ biến trong những tổ chức lớn như văn phòng chính phủ và quân đội. Tuy nhiên nó khá phức tạp và tỏ ra khó để phát triển những công cụ phần mềm linh hoạt. Ứng dụng nổi tiếng nhất của SGML chính là HTML; còn XML là phiên bản đơn giản hoá của SGML được thiết kế đặc biệt có thể thao tác trên Web. XML lưu trữ và tổ chức dữ liệu trong khi HTML cho phép hiển thị dữ liệu đó trong trình duyệt Web. Do đó XML và HTML có thể chuyển đổi cho nhau.

– Ngôn ngữ đánh dấu siêu văn bản: HTML

HTML, hay HyperText Markup Language, là dạng tài liệu cơ sở của World Wide Web, được thiết kế một cách đặc biệt để cho phép tham khảo hay *siêu kết nối – hyperlinks* đến những tập tin khác. Với tư cách là ngôn ngữ chung của Web, HTML là cơ sở cho tất cả giao diện thư viện số và giao diện chuẩn đối với Internet. Do đó, tài liệu nguồn thư viện số thường được trình bày dưới dạng HTML bao gồm hình ảnh, đồ hoạ, hình ảnh động, âm thanh, các chương trình tương tác hoàn chỉnh và văn bản. Hàng triệu trang Web được truy xuất mỗi ngày từ hàng ngàn máy tính Web server trên khắp thế giới.

– Ngôn ngữ đánh dấu mở rộng: XML

Hạn chế của SGML đã làm xúc tác phát sinh "ngôn ngữ đánh dấu mở rộng" XML – Extensible Markup Language. XML là một công cụ đầy năng lực. Nó cho phép các dạng thức tập tin trong phạm vi một tổ chức, như thư viện số, hợp lý hoá và được chia sẻ. Những tiêu chuẩn quan hệ khác nhau gia tăng năng lực XML và mở rộng khả năng ứng dụng. XML cung cấp một cú pháp biểu thị thông tin có cấu trúc hay metadata. Được kết hợp với những tiêu chuẩn phụ, XML còn được ứng dụng rất nhiều: hỗ trợ tái cấu trúc tài liệu, truy vấn, trích dẫn và định dạng thông tin, vv…

Siêu dữ liệu thư tịch

Những ai làm việc với thư viện số thì cần phải biết hai phương pháp chuẩn khác nhau về trình bày siêu dữ liệu tài liệu: Dạng biên mục máy đọc được MARC và Dublin Core. Dạng MARC được phát triển công phu, kiểm soát chặt chẽ, chi ly và bao hàm đến độ khá phức tạp, được tạo nên bởi những nhà biên mục học chuyên nghiệp chủ yếu để sử dụng trong thư viện truyền thống. Chuẩn Dublin Core chủ trương đơn giản hóa để có thể áp dụng rộng rãi cho tài liệu thư viện số đối với những người không cần được huấn luyện biên mục thư viện. Hai chuẩn này không những chú ý đến giá trị đặc thù của mình mà còn lưu tâm đến những triết lý căn bản đối nghịch nhau một cách tuyệt đối.

Chuẩn MARC được Thư viện Quốc hội Hoa Kỳ phát triển vào cuối thập niên 1960 để phục vụ việc trao đổi biểu ghi mục lục giữa các thư viện. MARC được giảng dạy khá kỹ lưỡng trong những chương trình đào tạo thư viện học trên thế giới. Chúng ta khá quen thuộc với biểu ghi MARC khi tiếp xúc với mục lục trực tuyến ở thư viện đại học.

Chuẩn Dublin Core là một tập hợp những thành phần metadata được thiết kế đặc biệt cho việc sử dụng không chuyên. Được dùng chủ yếu cho việc mô tả tài liệu điện tử. Đây là kết quả của một sự hợp tác nhiều người cùng xây dựng. Dublin là tên của thành phố ở Bang Ohio, Hoa Kỳ, nơi cuộc họp đầu tiên được tổ chức vào năm 1995. Từ đó đến nay đã có 11 lần hội nghị quốc tế tổ chức tại Anh, Úc, Phần Lan, Đức, Canađa, Nhật, và Hoa Kỳ để hoàn thiện. Dublin Core được Tổ chức Chuẩn Quốc gia Hoa Kỳ – ANSI phê chuẩn vào năm 2001.

So với dạng MARC, Dublin Core đơn giản một cách dễ chịu. Dublin Core chỉ bao gồm 15 thành phần so với hàng trăm của MARC. Như cái tên "core – nòng cốt" đã hàm ý rằng Dublin Core là một tập hợp những thành phần nòng cốt, ngoài ra còn có thể tăng thêm những thành phần phụ cho mục đích riêng. Hơn nữa, những thành phần hiện hữu có thể được cải tiến xuyên qua việc sử dụng. Tất cả những thành phần này đều có thể lập lại khi cần thiết.

Các bước cơ bản hình thành thư viện số

Sử dụng phần mềm thư viện số Hòn đá xanh – Greenstone là giải pháp thích hợp với điều kiện của chúng ta hiện nay. Các bước cơ bản hình thành thư viện số với Hòn đá xanh bao gồm:

1. Chọn các tài liệu muốn thêm vào;
2. Xác lập quyền hạn, bản quyền cho việc sử dụng các tài liệu này trong thư viện số;
3. Dùng máy quét để chuyển thể các tài liệu giấy tờ thành dạng kỹ thuật số;
4. Chuyển đổi các tài liệu này thành định dạng (có thể tích hợp giữa văn bản và hình) mà phần mềm Hòn đá xanh hiểu được (tốt nhất là HTML, các tài liệu soạn bởi Microsoft Word, riêng một số định dạng khác cũng có thể được chấp nhận nhờ vào một phần gọi là plug-in nhưng với mức độ chính xác khác nhau);

5. Đánh nhãn cho các chương, các đoạn và hình ảnh cho tài liệu;
6. Tổ chức tập hợp này thành thư viện số có cấu trúc tối ưu hóa;
7. Xây dựng thư viện số bằng phần mềm Hòn đá xanh;
8. Xuất bản tập hợp này thành CD-ROM và/hay phân phối trên Internet.

Để tạo ra một thư viện số, sản phẩm sau khi xuất bản phải ở dạng kỹ thuật số. Nếu tài liệu là sách, tờ bản tin hoặc các tài liệu giấy tờ khác thì chúng cần phải được quét (scan) để chuyển thành dạng máy tính hiểu được (bước 3). Thông thường công việc này được thực hiện nhờ vào bộ nhận dạng kí tự, nhưng cũng có thể được đánh máy lại. Bước 5 làm cho các phần khác nhau của tài liệu có thể được người sử dụng chọn và xem một cách độc lập trong thư viện số. Còn bước 6 liên quan đến việc gán các thuộc tính cho các tài liệu chẳng hạn như tiêu đề đề mục, từ khóa, và các dữ liệu thư tịch khác giúp sắp xếp thứ tự và tìm kiếm trong thư viện.

Những điều cần quan tâm trước tiên

Để bắt đầu một tiến trình biên tập tạo ra thư viện số từ tài liệu, văn bản giấy, chúng ta nên quan tâm đến những câu hỏi dưới đây:

1. Mục tiêu thư viện số của bạn là gì?
2. Nhóm đối tượng mà bạn quan tâm?
3. Nhóm đối tượng này đa dạng như thế nào, địa phương, tôn giáo hay toàn cầu?
4. Số lượng tài liệu bạn muốn có trong thư viện số?
5. Tổng cộng bao nhiêu trang?
6. Có bao nhiêu tài liệu loại hình ảnh đồ họa?
7. Tài liệu có cần thiết được chia thành các phần được tra cứu bởi một số ít người đọc và các phần được tham khảo một cách phổ biến?
8. Các tài liệu đã ở sẵn dạng kỹ thuật số chưa?
9. Nếu vậy, chúng ở dạng nào?
10. Quyền hạn và bản quyền của các tài liệu là gì?
11. Ai sở hữu bản quyền?
12. Có những tổ chức nào khác có cùng nhóm đối tượng không?
13. Bạn có sẵn sàng hợp tác với những tổ chức khác không?
14. Ngân quỹ bạn dành cho toàn bộ dự án thư viện số là bao nhiêu?
15. Bao nhiêu nhân lực dành cho việc biên tập tài liệu, quét tài liệu và lập trình?
16. Cần bao nhiêu máy tính cho dự án?
17. Bao nhiêu đĩa CD-ROM bạn muốn phát hành?
18. Chúng là miễn phí hay để bán?

Phần mềm Thư viện số Hòn đá xanh – Greenstone

Giải pháp tốt để xây dựng bộ sưu tập nhằm hình thành Thư viện số chính là sử dụng phần mềm nguồn mở Hòn đá xanh – Greenstone.

Greenstone đáp ứng yêu cầu của các Thư viện:

- Xây dựng các bộ sưu tập tài liệu điện tử từ Internet và các CSDL

trực tuyến dạng đa phương tiện: Sưu tập âm thanh, tranh ảnh, hình ảnh động, hoạt hình, đồ hình, toàn văn.
- Xây dựng các sưu tập về các chuyên ngành, bằng cách số hoá các tài liệu hiện có tại thư viện: Sách, tạp chí, luận văn, báo cáo khoa học, đề tài nghiên cứu khoa học, bài giảng, giáo trình, vv… với Sưu tập toàn văn.
- Xây dựng CSDL Sưu tập dạng thư tịch biên mục theo Dublin Core hay MARC 21.
- Hỗ trợ thực hành xây dựng sưu tập và biên mục Dublin Core và MARC 21 của Greenstone bằng công cụ Librarian Interface.
- Greenstone có thể tích hợp vào phần mềm quản lý thư viện có sẵn.
- Greenstone có thể được phát triển thành một phần mềm quản lý thư viện hoàn chỉnh theo yêu cầu của từng thư viện.

Greenstone chính là công nghệ giúp cán bộ tham khảo (reference librarian) chuyển câu hỏi của độc giả thành câu trả lời qua những bộ sưu tập và xuất bản thông tin trên CD-ROM hay những phương tiện truyền thông khác để phục vụ độc giả. Greenstone hỗ trợ cho cán bộ giảng dạy và nghiên cứu tự xây dựng những bộ sưu tập chuyên ngành và cùng với cán bộ thư viện đóng góp vào kho tài nguyên học tập theo phương châm ***"Thư viện số là sự tương tác giữa người sử dụng với thư viện để phục vụ chính người sử dụng".***

Phần mềm nguồn mở Thư viện số Hòn đá xanh – Greenstone là người bạn đồng hành của tất cả chúng ta, có thể hỗ trợ cho tất cả các loại hình thư viện với bất cứ quy mô nào để cùng XÂY DỰNG THƯ VIỆN SỐ.

TÀI LIỆU THAM KHẢO

1. FOX, Adward A., SULEMAN, Hussein, LUA. Ming. ***Building Digital Libraries Made Easy: Toward Open Digital Libraries.*** Proccedings. – 5th ICADL 2002. – Singapore, 11-14/12/2002.
2. LESK, MICHAEL. ***Practical Digital Libraries: Books, Bytes, and Bucks.*** – San Francisco: Morgan Kaufmann, 2000.
3. NGUYỄN MINH HIỆP và ĐOÀN HỒNG NGHĨA. ***Bài giảng Thư viện điện tử - Thư viện số (powerpoints)***. – TP. HCM : Thư viện Cao học, 2004.
4. NGUYỄN MINH HIỆP và ĐOÀN HỒNG NGHĨA. ***Thực hành xây dựng thư viện số***. – TP. HCM : Đại học Quốc gia, 2004.
5. ***Từ giấy đến bộ sưu tập.*** – Phần mềm Thư viện số Hòn đá xanh – Greenstone Digital Library Software, 2003
6. WITTEN, Ian H. và BAIBRIDGE, David. ***How to Build a Digital Library***. – New York : Morgan Kaufmann, 2003.