

MỘT SỐ LƯU Ý KHI XÂY DỰNG KẾ HOẠCH/DỰ ÁN SỐ HÓA TÀI LIỆU

Ths. Hoàng Thị Thu Hương và Ths. Nguyễn Trung Thành

Trường Đại học FPT

Xây dựng nguồn tài nguyên số là một trong những mục tiêu quan trọng của tất cả các loại hình thư viện trên thế giới, nhằm hướng đến mục tiêu phục vụ thông tin cho người dùng ở mọi lúc, mọi nơi, không phân định không gian, thời gian. Quá trình xây dựng thư viện số, thư viện ảo, thư viện điện tử ... đã được phát triển ở các nước phương Tây từ những năm 90 của thế kỉ XX. Nhưng ở Việt Nam, thư viện số, thư viện điện tử mới chỉ được quan tâm từ 6-7 năm trở lại đây. Và rất nhiều các cơ quan thông tin, thư viện hiện nay mới quan tâm tìm hiểu xem thư viện số là gì, xây dựng thư viện số thì cần những yếu tố gì? Trong phạm vi bài tham luận này, chúng tôi đưa ra một số kinh nghiệm triển khai dự án thư viện số để thảo luận.

I. Xây dựng một kế hoạch tổng thể về phát triển bộ sưu tập số

1. Xây dựng mục tiêu

Việc xây dựng thư viện số là nhằm đáp ứng năm (5) yêu cầu sau [1,13]:

- Societies – Thoả mãn nhu cầu thông tin
- Scenarios – Cung cấp dịch vụ thông tin
- Spaces – Hiện thị đa phương tiện
- Structures – Tổ chức cấu trúc thông tin
- Streams – Truyền đạt thông tin

2. Xác định nguồn tin số hoá/cần số hoá:

2.1. Loại hình văn bản cần số hoá

- Văn bản & Tài liệu: Sách, tạp chí, bài viết, báo cáo, luận văn, ...
- Âm thanh & Video: Bài nói chuyện, bài giảng, âm nhạc, phim ảnh, ...
- Thông tin địa lý: Hình ảnh, bản đồ, ...
- Thông tin sinh học: con người, động vật, thực vật, gen, ...
- Hình ảnh & Đồ hoạ: Hình ảnh 2D, 3D, bản vẽ kỹ thuật, sơ đồ, ...

- Phần mềm & dữ liệu: chương trình, hình mẫu (model), mô phỏng (simulation), vv...

2.2. Xác định các nguồn/kho tài liệu đã được máy tính hoá

- Tài nguyên trong thư viện: tài liệu dạng truyền thống (sách, báo...), tài liệu đã được máy tính hoá (file âm thanh, file văn bản, các CSDL thư mục, các CSDL toàn văn (luận văn, luận án, hình ảnh...)...
- Các hệ thống liên thông: thư viện khác, bảo tàng, trung tâm thông tin, ...

3. Đánh giá cơ sở hạ tầng

- **Cơ sở hạ tầng CNTT:** máy tính, máy chủ, hệ thống mạng và các hệ thống lưu trữ dữ liệu. Đánh giá khả năng có kho tài nguyên tại chỗ mang tính tập trung cùng với các hệ thống thư viện liên thông mang tính phân tán. Việc tổ chức liên thông và giao diện sử dụng hiệu quả là nhu cầu chỉ có công nghệ thông tin có thể giải quyết được
- **Con người:** kế hoạch yêu cầu có bao nhiêu nguồn nhân lực tham gia, trong đó có bao nhiêu nhân viên cơ hữu có đủ trình độ có thể tham gia dự án ngay, bao nhiêu nhân viên bên ngoài có thể thuê theo sự vụ, bao nhiêu nhân viên cần phải đào tạo để có thể thực hiện được công việc...?
- **Nguồn lực tài chính:** xác định ngân sách và kế hoạch giải ngân sát với thực tế để đảm bảo tiến độ triển khai dự án/ kế hoạch đề ra.

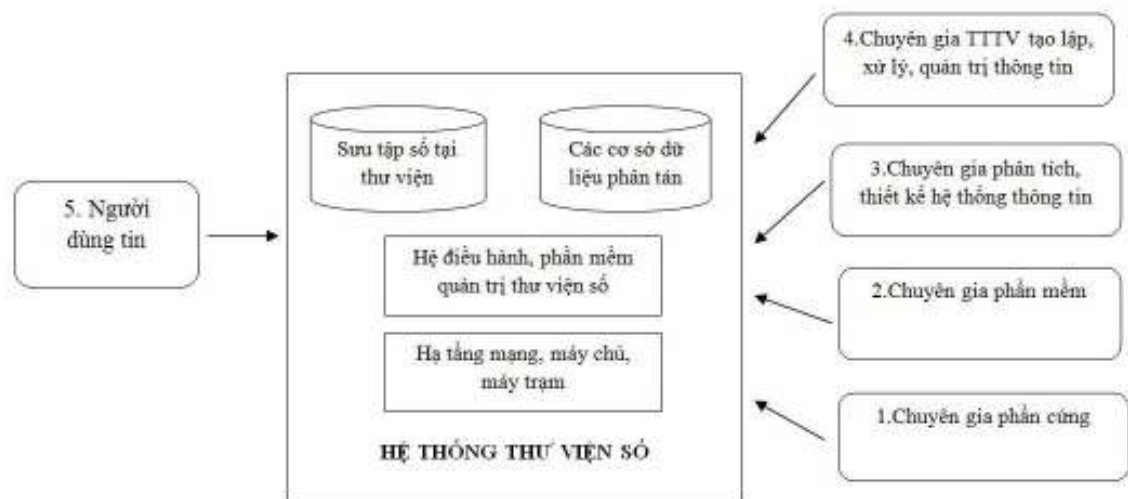
4. Những yêu cầu về quản lý tài liệu số hoá

- Đối tượng là các chuỗi dữ liệu (ví dụ các biểu ghi thư tịch, các chuỗi hình ảnh) có cấu trúc (ví dụ Dublin Core hoặc video MPEG-7) kèm theo phương thức sử dụng (không chỉ đơn thuần biểu ghi hoặc nội dung văn bản)
- Đa ngôn ngữ (Việt, Anh, Pháp, ...):
- Loại hình thông tin: toàn văn (full-text), tóm tắt, siêu dữ liệu
- Đa phương tiện (multimedia): cấu trúc, chất lượng, tìm kiếm
- Cấu trúc dựa trên chuẩn XML: DC, MODS, MARC 21 XML, ...
- Bộ sưu tập phân tán: Kleisli, CIMI, Z39.50, ...
- Tìm kiếm tổng quát: thu thập đa nguồn, chọn lọc tự động theo kinh nghiệm, tìm kiếm song song, hoà trộn kết quả
- Truy cập: Phân quyền, phân cấp, IPR, tính phí, bảo mật và cá nhân hoá
- Đối với tài liệu trên giấy:

- Đối với những tài liệu không nhận dạng quang (gọi tắt là OCR) là những tài liệu xuất bản bằng hình thức in roneo, đánh máy chữ và chất liệu giấy rất xấu thì dung lượng sản phẩm phải nhỏ nhưng vẫn giữ được độ sắc nét của chữ, hình ảnh trong tài liệu.
- Đối với tài liệu OCR (thường là những tài liệu in sau này có chất lượng tốt) thì sản phẩm phải có chức năng tìm kiếm toàn văn nội dung tài liệu (Searchable fulltext). Định dạng của sản phẩm đầu ra phong phú, bao gồm nhiều định dạng: doc, xls, txt, pdf...

5. Phần mềm quản lý thư viện số và mô hình thư viện số

Phần mềm quản lý thư viện số là một phần không thể thiếu để quản lý nguồn tài nguyên số nói riêng và quản lý khai thác thư viện số nói chung. Hiện nay, có rất nhiều phần mềm quản lý thư viện số từ những phần mềm thương mại của các công ty tin học như phần mềm thư viện số của Công ty CMC, Công ty Tinh Vân, Công ty Lạc Việt..., cho đến các phần mềm mã nguồn mở như Dspace, Greenstone... Dưới đây là một mô hình thư viện số cơ bản trên thế giới:



Mô hình thư viện số cơ bản [2]

6. Vấn đề bản quyền tài liệu số hoá

Vấn đề bản quyền là vấn đề rất quan trọng đối với các thư viện hiện nay. Theo Luật sở hữu trí tuệ Việt ả am, những tài liệu dưới đây không được bảo hộ bản quyền:

- Luật sở hữu trí tuệ Việt ả am 2005; Chương I, Mục 1, Điều 15 “Các đối tượng không thuộc phạm vi bảo hộ quyền tác giả: 1. Tin tức thời sự thuần túy đưa tin. 2. Văn bản quy phạm pháp luật, văn bản hành chính, văn bản khác thuộc lĩnh vực tư pháp và bản

dịch chính thức của văn bản đó. 3. Quy trình, hệ thống, phương pháp hoạt động, khái niệm, nguyên lý, số liệu”

- Luật sở hữu trí tuệ Việt ăm 2005: Điều 27, khoản (a), (b) “...Tác phẩm không thuộc loại hình quy định tại điểm a khoản này có thời hạn bảo hộ là suốt cuộc đời tác giả và năm mươi năm tiếp theo năm tác giả chết; trong trường hợp tác phẩm có đồng tác giả thì thời hạn bảo hộ chấm dứt vào năm thứ năm mươi sau năm đồng tác giả cuối cùng chết”

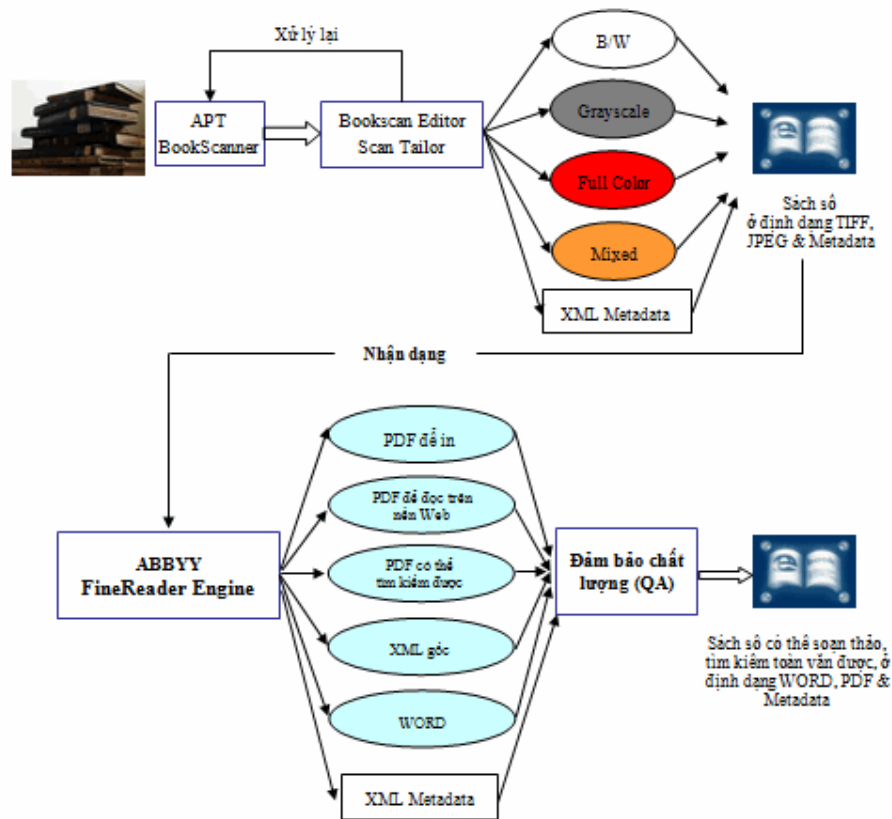
Việc số hoá tài liệu cho thư viện số là không vi phạm bản quyền nếu: Tài liệu nằm ngoài bản quyền hoặc tài liệu được bảo hộ bản quyền nhưng số hoá để sử dụng với mục đích phi thương mại trong phạm vi hạn chế của thư viện, trường học, viện nghiên cứu.

Bản thân việc số hoá tài liệu không vi phạm bản quyền, việc vi phạm hay không phụ thuộc vào mục đích sử dụng (chẳng hạn dùng với mục đích thương mại làm ảnh hưởng đến lợi ích kinh tế của người nắm giữ bản quyền là vi phạm) và phạm vi sử dụng (ví dụ nếu phổ biến rộng rãi ra công chúng, ngoài phạm vi thư viện là vi phạm).

Việc download, phát tán tài liệu số hiện nay đang rất phổ biến ở Việt ăm đang ảnh hưởng nghiêm trọng đến quyền lợi của tác giả và các nhà xuất bản mà ả hà nước cần có những quy chế quản lý và xử phạt để bảo hộ quyền lợi chính đáng của tác giả và nhà xuất bản.

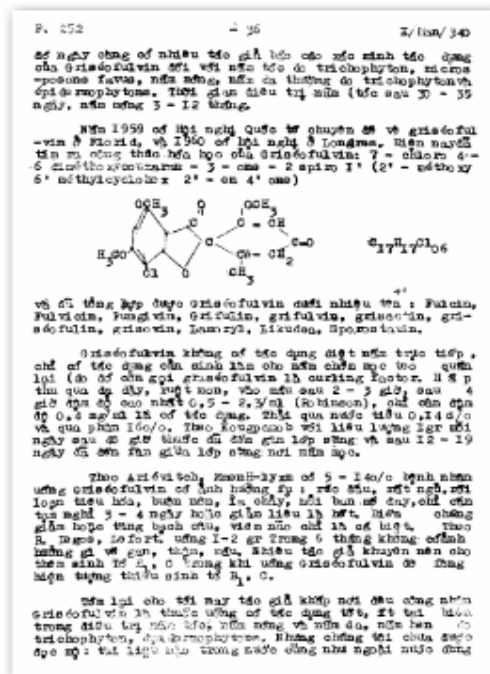
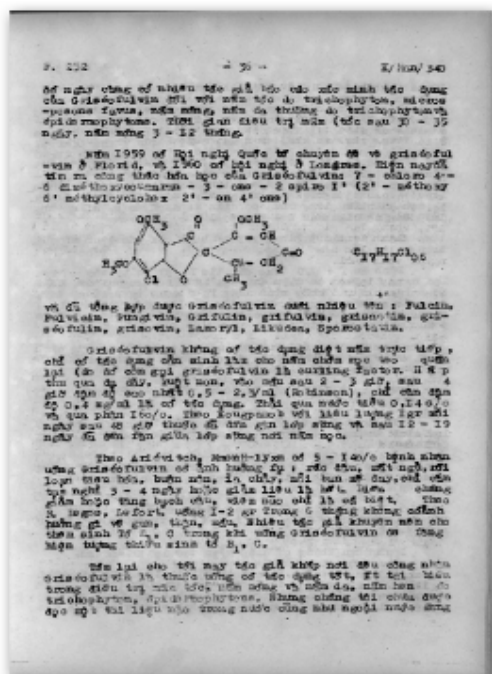
II. Giải pháp số hóa tài liệu tài liệu trên giấy

1. Quy trình số hóa tài liệu



2. Giải pháp số hóa tài liệu

* Đối với loại hình tài liệu không OCR: Tài liệu sau khi scan có thể sử dụng một số phần mềm xử lý ảnh như: Photoshop, OmiPage, ... để chỉnh sửa và chuyển đổi sang định dạng pdf. Tại Viện Công nghệ thông tin - Thư viện y học TW, Công ty Đông Kinh đề xuất sử dụng phần mềm xử lý do công ty nghiên cứu và phát triển. Sản phẩm sau khi số hóa:



Ảnh sau khi quét có dung lượng 95MB

Ảnh sau khi xử lý dung lượng 115 KB

Sản phẩm đầu ra của phần mềm do Đông Kinh phát triển đạt được các yêu cầu:

- Dung lượng nhỏ.
- Độ sắc nét của tài liệu cao.

Giải pháp này có thể áp dụng cho các đơn vị có nguồn lực kinh tế hạn chế, hạ tầng kỹ thuật không cao. Yêu cầu cho giải pháp: Máy scan văn phòng, phần mềm xử lý của công ty Đông Kinh phát triển. Tổng chi phí cho giải pháp chưa đến một ngàn đô la Mỹ.

Sản phẩm đầu ra có thể được quản lý bằng các phần mềm mã nguồn mở sẵn có như: GreenStone, DSpace ... Biên mục theo chuẩn Dublincore, XML ...

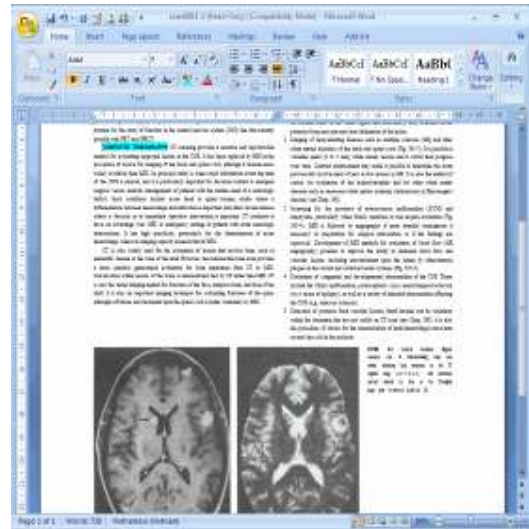
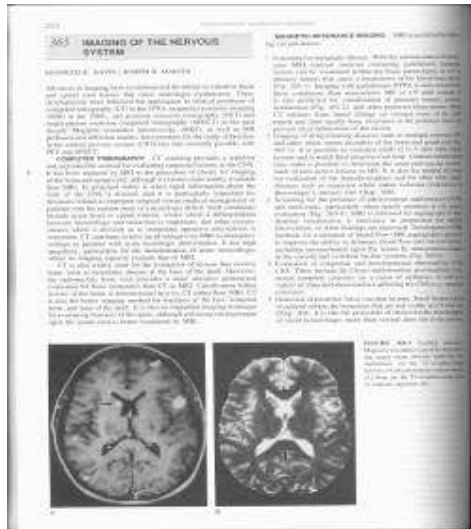
* Đối với loại hình tài liệu OCR: sử dụng phần mềm FineReader Engine của hãng ABBYY. Phần mềm này có một số ưu điểm nổi trội so với các giải pháp khác, đó là:

- **Xử lý hoàn hảo đối dữ liệu tiếng Việt:** Mặc dù Kofax, Captiva đều là những giải pháp hàng đầu về số hóa và bóc tách thông tin trên thế giới, nhưng tại thời điểm này mới chỉ duy nhất ABBYY FineReader Engine có khả năng làm việc được với các tài liệu tiếng Việt.
- **Độ chính xác:** Qua kiểm nghiệm và đánh giá, kết quả nhận dạng và bóc tách thông tin từ Báo cáo tài chính và Giấy yêu cầu mở thư tín dụng đạt trên 99%. Với tổ chức kiểm soát lỗi nhiều mức, kiểm soát lỗi do quá trình nhận dạng cho phép kiểm soát và hạn chế tối đa lỗi.

- **Tính linh hoạt:** Với kiến trúc mở và mềm dẻo, ABBYY FineReader Engine cho phép đáp ứng các mô hình triển khai khác nhau, xử lý tập trung, xử lý tại các chi nhánh tùy theo yêu cầu của khách hàng.
- **Khả năng tích hợp:** Với khả năng xử lý tất cả các định dạng ảnh phổ biến, làm việc được với tất cả các máy quét thông dụng trên thị trường, cho phép kết xuất kết quả ra nhiều định dạng file (Excel, XML, ...) cho phép ABBYY FineReader Engine có thể tích hợp với hầu hết các hệ thống Back Office đã triển khai tại khách hàng.
- **Giữ nguyên bố cục của tài liệu gốc:** ADRT (Adaptive Document Recognition Technology) là một bước tiến lớn trong công nghệ nhận dạng tài liệu; là một phần quan trọng và chỉ có trong công nghệ nhận dạng của ABBYY. ADRT được dùng để nhận dạng các cấu trúc logic, cách dàn trang cũng như các định dạng khác nhau trong tài liệu nhiều trang, ví dụ như: mục lục, đầu trang, chân trang, chú thích, chú thích bảng, chú thích ảnh, số trang v.v... Khi kết quả nhận dạng được lưu dưới dạng Microsoft Word, các định dạng trên được nhìn nhận như những đối tượng tương ứng trong Word chứ không đơn thuần chỉ là các khối văn bản. Để đạt được tính chính xác trong việc phân tích dàn trang, cấu trúc logic cũng như các định dạng trong tài liệu là do ADRT xem xét và xử lý các tài liệu nhiều trang như là một đối tượng tổng thể chứ không phải là tập hợp các trang riêng biệt. Với công nghệ ADRT, người dùng sẽ không mất, hoặc mất rất ít thời gian cho việc biên tập lại kết quả nhận dạng.
- **Kết xuất kết quả nhận dạng:** FineReader Engine có khả năng chuyển đổi khối lượng lớn tài liệu giấy sang tài liệu số dưới các định dạng có thể tìm kiếm và biên tập được như MSWord, MSEXcel, PDF, PDF/A. Đặc biệt là định dạng PDF/A, là định dạng được coi là định dạng lý tưởng để lưu trữ kết quả số hóa tài liệu hiện nay. Định dạng PDF/A gồm 2 lớp, lớp trên là ảnh quét, lớp dưới là text nhận dạng được.
 - Lớp ảnh quét là hình ảnh trung thực của tài liệu gốc nên nó mang đến cho người đọc tính chính xác của thông tin cũng như hình thức thể hiện.
 - Lớp dưới là text nhận dạng được, là cơ sở để các máy tìm kiếm có thể lập chỉ mục nội dung của tài liệu và qua đó người dùng có thể tìm kiếm nội dung trong tài liệu số.
 - Định dạng PDF/A có một điểm đặc biệt là: lớp text nhận dạng được dàn trùng khớp về vị trí với ảnh gốc. Bởi vậy, cho dù là đang đọc trên lớp ảnh nhưng người đọc có thể sao chép một cách chuẩn xác nội dung số của tài liệu gốc.

- Một số kết quả nhận dạng của tài liệu OCR:

+ ảnh dạng trang sách



+ Định dạng PDF hai lớp



Kết luận

Có rất nhiều các vấn đề mà các nhà thư viện hiện nay cần quan tâm khi xây dựng một kế hoạch/dự án phát triển thư viện số như: phát triển bộ sưu tập số, phát hiện và tổ chức nguồn tin, tìm tin/truy cập, tương tác người - máy/người sử dụng, hạ tầng thông tin-quản trị tri thức, lưu trữ số, dịch vụ số, ứng dụng mạng xã hội, sở hữu trí tuệ, an ninh mạng, các vấn đề văn hoá và luật pháp... Việc xây dựng, phát triển bộ sưu tập tài liệu số cho mỗi thư viện sẽ được tiến hành ở quy mô và mức độ khác nhau dựa trên thực lực của đơn vị cụ thể. Chúng tôi sẽ đi sâu phân tích những yêu cầu cụ thể ở những nghiên cứu tiếp theo.

Tài liệu tham khảo

1. Bản tin thư viện tháng 12/2003.
<http://gralib.hcmuns.edu.vn/fesal/bantin1203/bai3.pdf>
2. Candela , L. (2007) Setting the Foundations of Digital Libraries. The DELOS Manifesto
3. Đề xuất áp dụng công nghệ số hóa vào việc xây dựng kho dữ liệu số: Tài liệu kỹ thuật/ Công ty đầu tư và phát triển Đông Kinh.H.:2010.41tr.-27cm.
4. Số hóa tài liệu tiếng Việt/ Quốc Dũng// Tạp chí Thế giới vi tính. – ả ăm 2010, tháng 8.(nguồn tham khảo: <http://www.pcworld.com.vn/articles/cong-nghe/ung-dung/2010/08/1219965/so-hoa-tai-lieu-tieng-viet>)
5. ả guyễn Trung Thành (2011). Giới thiệu giải pháp số hoá tài liệu tại Viện Công nghệ thông tin – Thư viện y học TW, *Tạp chí Thư viện*, số 4.