THƯ VIỆN SỐ GREENSTONE

TỪ GIẤY ĐẾN BỘ SƯU TẬP

Giáo sư Michel Loots, Dan Camarzan and Ian H.Witten

Human Info NGO, Belgium
Simple Words, Romania
Trường Đại học Waikato, New Zealand

Greenstone là một bộ phần mềm giúp xây dựng và phân loại các tập hợp thư viện số. Nó đưa ra một cách tiếp cận mới trong việc tổ chức và xuất bản thông tin trên Internet hoặc trên CD-ROM. Greenstone là kết quả của dự án thư viện số tại trường đại học Waikato, NewZealand (New Zealand Digital Library Project), đã được triển khai và phân phối với sự hợp tác của hai tổ chức UNESCO và Human Info NGO. Greenstone là một phần mềm nguồn mở có sẵn tại địa chỉ http://greenstone.com , trong mục GNU General Public License.

Chúng tôi đảm bảo rằng phần mềm này đáp ứng tốt nhu cầu của bạn. Nếu có bất kì vấn đề nào liên quan đến phần mềm này xin trình bày tại greenstone@cs.waikato.ac.nz

Greenstone gsdl-2.39 **Tháng 3 năm 2003**

Nội dung tập tài liệu

Tài liệu này mô tả cách tạo bộ sưu tập CD-ROM từ các tài liệu giấy. Nó miêu tả đầy đủ các thủ tục và nhu cầu tài chính cần thiết liên quan đến việc quét và quá trình nhận dạng ký tự, vì vậy phần nội dung phải được định dạng đúng để ứng dụng được phần mềm Greenstone. Nó cũng miêu tả cách sử dụng chức năng tổ chức bộ sưu tập, nói đơn giản là "Organizer", để tạo ra và chỉnh sửa nguyên liệu liên quan đến bộ sưu tập. Đây là phần mềm sẵn có, được phân phối dưới tên gọi Greenstone chạy trên hệ điều hành Windows. Chúng tôi cố gắng làm cho đơn giản đi nhằm giúp bạn đọc dễ hiểu và khi dụng phần mềm này. Khi nhắc đến một thương hiệu hay sản phẩm nào hoàn toàn là nhằm mục đích minh họa và không cũng phải chúng tôi khẳng định sản phẩm đó tốt hơn hoặc quan tâm nhiều hơn một sản phẩm nào khác.

Các tập tài liệu trong Bộ phần mềm Greenstone

Bộ phần mềm này bao gồm 4 tập tài liệu:

- Hướng dẫn cài đặt
- Hướng dẫn sử dụng
- Hướng dẫn phát triển
- Từ tài liệu bằng giấy đưa lên mạng.

Những thành viên tham gia dự án phần mềm Greenstone

Quá trình scanning, Organizer và các quá trình khác có liên quan đến việc tạo ra các bộ sưu tập từ sự cộng tác phi lợi nhuận, được phát triển bởi Giáo sư Michel Loots, MD, Human Info NGO và HumanityCD, Dan Camarzan of Simple Words, và các nhóm cộng tác viên ở Brasov, Romania.

Phần mềm này là sự đóng góp của nhiều người trong đó Rodger McNab và Stefan Boddie là hai người đóng góp chính trong việc xây dựng và phát triển phần mềm này. Ngoài ra còn có sự đóng góp của các tác giả sau: David Bainbridge, George Buchanan, Hong chen, Elke Duncker, Carl Gutwin, Geoff Holmes, John McPherson, Craig Nevill-Manning, Gordon Paynter, Bernhard Pfahringe, Todd Reed, Bill Rogers và Stuart Yeates. Những thành viên khác trong dự án Thư viện số tham gia phần Thiết kế hệ thống là: Mark Apperley, Sally Jo Cunningham, Steve Jones, Te Taka Keegan, Michel Loots, Malika Mahoui và Lloyd Smith.

Chúng tôi cũng chân thành cảm ơn những đơn vị đã tham gia khâu đóng gói cũng như phân phối bộ phần mềm này: MG, GDBM, WGET, WV, PDF2HTML, PERL.

MỤC LỤC

3.3 Các hình thức khác trong tiến trình nhận dạng kí tự:

Tự đánh máy
Các tập tin hình ảnh

3.4 Kết hợp giữa việc quét và nhận dạng kí tự

4 BA VÍ DỤ: TỪ 1000 ĐẾN 100,0000 TRANG

4.1 Tập hợp nhỏ: 500-1000 trang

4.2 Toàn bộ tài liệu từ một tổ chức: 5000 trang

4.3 Thư viện nhỏ: 100,000 trang

5 TẠO RA MỘT BỘ SƯU TẬP ĐIỆN TỬ

5.1 Các phương pháp xây dựng tập hợp

5.2 Công cụ tổ chức

Cài đặt và sử dụng Organizer
Mô hình tài liệu
Tìm hiểu chức năng tổ chức

5.3 Các file tài liệu đính kèm

1. Giới thiệu

Mục tiêu của phần mềm thư viện số Greenstone là nhằm giúp cho các tổ chức như các trường đại học, các tổ chức Liên hiệp quốc, các tổ chức phi chính phủ, phi lợi nhuận và các chính phủ trong việc tạo ra các loại thông tin có thể được phân phối trực tuyến hoặc trên các CD-ROM.

Các bước cài đặt cơ bản:
i. Chọn các tài liệu muốn thêm vào
ii. Thiết đặt quyền hạn, bản quyền cho việc sử dụng các tài liệu này trong thư viện số.
iii. Dùng máy quét và ORC để chuyển thể các tài liệu giấy tờ thành dạng kỹ thuật số
iv. Chuyển đổi các tài liệu này thành một định dạng (có thể tích hợp giữa văn bản và hình) mà phần mềm Greenstone 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 plug-in nhưng với mức độ chính xác khác nhau (xem phần hướng dẫn người sử dụng của Greenstone để biết thêm thông tin)
v. Đặt tên cho các chương, các đoạn và hình ảnh cho tài liệu.
vi. Sắp xếp các bộ sưu tập này thành thư viện số có cấu trúc tối ưu hóa.
vii. Xây dựng thư viện số bằng phần mềm Greenstone.
viii. Xuất bản tập hợp này thành CD-ROM và/hay phân phối trên Internet.

Để tạo ra một thư viện số, các văn bản phải ở dạng kỹ thuật số. Nếu tài liệu là sách, 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 iii). Thông thường công việc này được thực hiện nhờ vào bộ nhận dạng kí tự ORC, nhưng thỉnh thoảng vẫn dùng đánh máy. Tiến trình này được trình bày trong các chương 2 đến 4 của phần hướng dẫn sử dụng.
Bước v. cho phép người đọc chọn và xem các phần khác nhau trong văn bản một cách độc lập trong thư viện số. Còn bước vi. gán các thuộc tính cho các tài liệu chẳng hạn như loại chủ đề, các từ khóa, các dữ liệu thư mục giúp sắp thứ tự và tìm kiếm trong thư viện. Những bước này được mô tả trong chương 5 với những hướng dẫn chi tiết về chương trình Organizer đi kèm trong bộ phần mềm Greenstone.

Tài liệu hướng dẫn này giới thiệu nhiều vấn đề ảnh hưởng đến quá trình biên tập tạo ra thư viện số từ tài liệu, văn bản giấy. Trước bắt đầu, bạn nên quan tâm đến những câu hỏi dưới đây:

- Mục tiêu thư viện số của bạn là gì?
- Nhóm đối tượng mà bạn quan tâm?
- Nhóm đối tượng này có qui mô như thế nào: địa phương, khu vực hay toàn cầu?
- Số lượng tài liệu bạn muốn có trong thư viện số ?
- Tổng cộng bao nhiêu trang?
- Có bao nhiêu tài liệu là hình ảnh đồ họa?
- 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?
- Các tài liệu đã ở sẵn dạng kỹ thuật số chưa?
- Nếu vậy, chúng ở dạng nào ? (Xin lưu ý các tập tin dạng PDF sẽ không được xem chuyển đổi tự động sang dạng văn bản kỹ thuật số, vì các trang trong tập tin thường chỉ là hình ảnh.)
- Bản quyền của tài liệu là gì?
- Ai sở hữu bản quyền?
- Có những tổ chức nào khác có cùng nhóm đối tượng không?
- Bạn có sẵn sàng hợp tác với những tổ chức khác không?
- Ngân quỹ bạn dành cho toàn bộ dự án thư viện số là bao nhiêu?
- Bao nhiêu nhân lực bạn dành cho việc biên tập tài liệu, quét tài liệu và lập trình ?
- Cần bao nhiêu máy tính cho dự án?
- Bao nhiêu đĩa CD-ROM bạn muốn phát hành?
- Chúng miễn phí hay để bán?

2. MÁY QUÉT VÀ QUÉT TÀI LIỆU

Bước đầu tiên khi chuyển các tài liệu giấy tờ thành một tập hợp thư viện số là có hình ảnh các trang tài liệu ở dạng kỹ thuật số. Khâu kế tiếp là nhận dạng kí tự bằng quang học (OCR) và khâu này cần các hình ảnh tài liệu rõ ràng và có chất lượng cao. Giai đoạn số hóa đòi hỏi máy quét phải làm việc được ở độ phân giải 300 dpi. Hầu hết các công việc quét được thực hiện theo dạng trắng đen, nhưng đối với các tài liệu dùng màu sắc thì cần phải quét với một máy quét màu. Thông thường bìa sách sử dụng màu và sẽ được quét như là một hình ảnh màu.

2.1 Máy quét

Các máy quét rất đa dạng về giá cả, hình dạng và kích thước. Chúng có giá từ 100USD cho các máy quét hình phẳng cho đến 50000USD cho các máy quét công nghiệp cở lớn của các nhà sản xuất như Bell & Howell. Rất nhiều website cung cấp đa dạng máy quét. Để tìm những website này, bạn chỉ cần dùng từ khóa "scanners" vào Google, Altavista hoặc Yahoo.

Kết quả của một trang tài liệu được quét là một tập tin máy tính mà thông thường ở định dạng TIFF hoặc Bitmap. Định dạng nén TIFF phiên bản 4 là dạng tốt nhất. Trung bình một trang được nén và được chuyển thành định dạng này chỉ chiếm khoảng 50Kb, trong khi ở định dạng Bitmap không nén sẽ là 2Mb.

Các máy quét hình phẳng giá thấp

Các loại máy quét hình phẳng là rẻ nhất và được sử dụng nhiều nhất. Thuộc nhiều hãng khác nhau: HP, Agfa, Acer v.v.., giá cả từ 100USD đến 300USD. Chúng đều có thể quét hình trắng đen hay màu. Do chi phí thấp nên có thể trang bị cho mỗi máy tính một máy quét riêng.
Điểm bất lợi của những máy in này là cho ra những hình ảnh của trang tài liệu ở mức trung bình, tỉ lệ quét thấp, không bền trong những môi trường ẩm thấp và khá dễ hư. Chúng ta phải quét từng trang một. Mỗi trang phải được định vị cẩn thận theo lề bảng quét. Hiệu suất của những máy in này kém. Mặc dù các nhà sản xuất khẳng định rằng mỗi trang tài liệu có thể được quét trong vòng chưa tới một phút nhưng thực tế cho thấy rằng khó có thể thể đạt tới mức 12 trang mỗi tiếng. Tiến trình quét thường làm ì ạch máy tính gắn kết. Do vậy những máy in này chỉ hữu dụng cho các công việc nhỏ (số lượng trang cần quét ít- từ 200 đến 400 trang một tháng một cách thương xuyên) hoặc các công việc xảy ra một lần từ 1000 đến 2000 trang

Máy quét cấp thấp có ngăn để giấy

Các máy quét này thường có giá từ 500USD cho đến 1200USD. Có thể quét từ 10 đến 50 trang tài liệu một lần. Vì vậy người điều khiển không cần có mặt liên tục tại máy quét. Điều này sẽ làm gia tăng số lượng trang đến 150-200 trang/ngày. Những loại máy in này có tuổi thọ cao hơn, thường thì khoảng từ 30000 đến 50000 trang.

Điểm bất lợi của chúng là tại một thời điểm chỉ quét một mặt tài liệu – ngăn để các trang tài liệu phải được đảo lại để quét mặt sau của tài liệu. Và điều này có thể gây ra vấn đề bởi vì ngăn để giấy rất thường gặp trục trặc và đôi lúc làm kẹt giấy.

Những loại này hữu ích cho công việc quét từ 1500 đến 3000 trang/tháng.

Các máy quét màu

Để quét hình màu thì nhất thiết ta phải có máy quét màu. Nhưng nói chung, chưa đến 5% các ấn phẩm chứa màu cộng với bìa tài liệu. Vì vậy một máy quét hình phẳng giá thành thấp như kể trên là thường là đáp ứng được nhu cầu. Chúng ta nên chọn máy quét có độ phân giải lên đến 600dpi.

Các máy quét 2 mặt chuyên nghiệp

Các máy quét chuyên nghiệp là các máy tốt và đáng tin cậy, có khả năng xử lý một số lượng lớn trang tài liệu- từ 2000 đến 10000 trang/ngày. Chúng có hệ thống khay để giấy tự động, xử lý các nhóm gồm từ 50 đến 200 trang. Các máy quét tốt nhất và nhanh nhất thuộc dạng này có thể quét cả 2 mặt tài liệu cùng lúc.
Các máy quét này yêu cầu máy tính kết nối với nó phải mạnh và có dung lượng ổ cứng ít nhất là 10 -20Gb, giá từ 5000 – 50000USD. Chẳng hạn như: máy quét Cannon DR-6020 giá khoảng 5000USD, có thể quét 2 mặt tài liệu, 2000 trang/ngày và tuổi thọ từ 600000 – 800000 trang. Các máy quét nhãn hiệu Bell&Howell và Fujitsu, giá từ 10000 – 50000USD, có tuổi thọ đến hàng triệu trang.
Các máy quét phích nhỏ có giá từ 15000USD đối với loại bán tự động cho đến 80000USD đối với loại tự động hoàn toàn.

Các chương trình quét

Mỗi máy quét đều có phần mềm riêng được cài đặt trên máy tính để điều khiển máy quét. Một số máy quét có card được cài đặt vào máy tính để tăng tốc độ quét.

2.2 Chuẩn bị các tài liệu

Trước khi được quét, tài liệu phải được chuẩn bị tốt. Tài liệu phải sạch, khô ráo, các ghim kẹp tài liệu được tháo rời, và các trang được xếp thẳng
Gáy sách nên được gở bỏ. Các cuốn sách của thư viện thông thường được đóng lại, khi đó bạn nên cẩn thận khi gở bỏ gáy sách để dể dàng khi đóng sách lại.
Nếu như chỉ có ít tài liệu thì việc cắt gáy sách có thể được thực hiện bằng tay thông qua một cây thước và bộ cắt. Còn nếu có nhiều tài liệu thì nên dùng các máy cắt bằng tay đặc biệt.
Đối với số lượng lớn – hơn 20 tài liệu thì chúng tôi khuyến cáo nên yêu cầu thợ in hoặc chủ tiệm photo sử dụng máy cắt chuyên dụng của họ, đừng quên gở bỏ các ghim kẹp kim loại vì chúng có thể gây hư hại máy cắt.

2.3 Tiến trình quét

Nhờ vào phần mềm đi kém với máy quét, một bức ảnh tài liệu kĩ thuật số sẽ được quét và chuyển thể thành hình ảnh định dạng Bitmap hoặc TIFF. Những tập tin hình này sẽ được lưu trữ trên ổ cứng với các tên chuẩn, và tiến trình nhận dạng kí tự sẽ được kích hoạt ngay khi một số tài liệu được quét. Công việc này có thể thực hiện bởi ngươi quét tài liệu hoặc người khác.

Thông thường ta cần độ phân giải khi quét vào khoảng 300dpi , mặc dù đôi lúc 200dpi cũng chấp nhận được.

Quản lý chất lượng

Mục tiêu cuối cùng của giai đoạn quét hoặc là nhận dạng kí tự trong trang để có được các bản tài liệu ở dạng văn bản hoặc HTML, hoặc là để tạo ra các tập tin ảnh tốt, chẳng hạn như: các tập tin ảnh PDF. Trong cả 2 trường hợp thì chất lượng của các ảnh là rất quan trọng. Nếu như chất lượng ảnh thấp thì các tập tin ảnh không đẹp và tốn nhiều bộ nhớ hơn. Chất lượng ảnh đặc biệt ảnh hưởng đến tiến trình nhận dạng kí tự: với chất lượng thấp, hiệu suất giảm đến 40%. Thông thường quá trình nhận dạng kí tự chiếm hơn 90% tổng chi phí, vì vậy chất lượng quét có thể ảnh hưởng đến chi phí.
Chất lượng của tập tin TIFF có thể được nâng cao bằng cách điều chỉnh tiến trình quét cho mỗi loại tài liệu thông qua việc sử dụng các tuỳ chọn được cung cấp bởi phần mềm quét. Loại tài liệu khá rõ ràng sẽ cần các tuỳ chọn

sáng hơn, nghĩa là độ tương phản phải được điều chỉnh phụ thuộc vào chất lượng bản in và …
Đầu tiên chia thành từng nhóm có chất lượng in và giấy tương tự nhau. Tiến hành kiểm tra OCR trên một trang đơn giản của nhóm đầu tiên để lựa chọn các chuẩn tốt nhất. Sau đó quét tất cả các trang còn lại trong nhóm này trước khi xử lý đến nhóm khác.

Qui định tên tập tin

Mỗi cuốn sách hay mỗi tài liệu có một số hoặc mã duy nhất, con số này sẽ trở thành tên của thư mục chứa tất cả các hình ảnh TIFF trong tài liệu. Tuỳ theo hệ điều hành máy tính (DOS, Windows, UNIX, LINUX, …), các tên này dài từ 8 – 128 kí tự. Chúng ta chỉ giới thiệu đến tên tài liệu từ 8 -16 kí tự. 5 kí tự đầu tiên xác định tài liệu, 3 kí tự còn lại xác định các loại tài liệu. Ví dụ: u1748e12.tif xác định tập tin TIFF trong trang 12 của cuốn sách được viết bằng tiếng Anh có mã số là u7548.
Chỉ định một thư mục trên ổ cứng cho các công việc quét, sau đó tạo thư mục con cho mỗi công việc này. Bên trong thư mục con này tạo thư mục con tương ứng cho mỗi phần. Ví dụ: thư mục u7548e sẽ chứa toàn bộ các tập tin hình ảnh TIFF, bao gồm các ảnh màu.

2.4 Hiệu suất và các tài nguyên

Bạn không nên đánh giá thấp tầm quan trọng của công việc quét tài liệu và đặc biệt là tiến trình nhận dạng kí tự. Chúng ta nên xem tiến trình quét và nhận dạng kí tự là 2 tiến trình riêng biệt. Chúng ta nên căn cứ trên cả 2 phương diện kinh tế và thực tế để đưa ra sự lựa chọn tối ưu.
Một số quan điểm cần xem xét là việc đầu từ vào các máy quét và máy tính; không gian và tài nguyên con người; huấn luyện nhân lực; chi phí lương; số lượng trang khởi đầu và tổng số trang quét; thời hạn; và tài liệu có được xuất khẩu sang các đối tác khác không.

Chi phí quét

Việc đầu tư vào các trang thiết bị quét và tự thực hiện công đoạn quét tài liệu hay để đối tác khác thực hiện quét tài liệu là một quyết định quan trọng.
- Áp lực thời gian của công việc quét tài liệu
- Tổng số trang cần quét
- Chi phí lương phải trả cho người thực hiện công đoạn quét.

Những người thực hiện công việc quét phải năng nỗ, lành nghề và có tinh thần trách nhiệm cao.
Thông thường chi phí quét tài liệu của một công ty chuyên nghiệp là 0.06USD/page. Chi phí này có thể phải được cộng thêm vào chi phí vận chuyển 0.03USD/page từ quốc gia đang phát triển đến quốc gia phát triển hay 0.015USD/page cho chi phí vận chuyển trong nước.

Bảng 1 thẩm định chi phí quét tài liệu ứng các loại máy quét khác nhau. Ba cột đầu liên quan đến chi phí lao động. Cột đầu tiên liên quan đến năng suất tính theo số trang/tháng, giả định đây là công việc toàn thời gian. Cột thứ 2 là tài nguyên tính theo số giờ trong tháng của mỗi người trên mỗi trang được tính bằng cách lấy số giờ làm việc trong một tháng chia cho số lượng trang trong, giả định có 180 giờ làm việc / tháng.

BẢNG SCANNER và SCANNING

	Khả năng (Trang/tháng)	Số Giờ/trang (180-giờ/tháng)	C.phí/trang (tối đa $4/giờ)	Scanner acquisition	Tuổi thọ của máy Scanner (trang)	Số trang đưa dịch vụ quét ($.06/trang)
Flat bed scanner	2,500	0.072	$0.288	$300	7,000	5,000
Scanner with sheet-feeder	8,000	0.0225	$0.09	$800	30,000	13,000
Professional: low-end duplex	40,000	0.0045	$0.018	$6,000	600,000	100,000
Professional: high-end duplex	150,000	0.0012	$0.0048	$50,000	8,000,000	833,000

Để tính chi phí cho mỗi trang, chúng ta nhân tổng chi phí lương theo giờ với cột thứ 2 trong bảng 1. Ví dụ, cột thứ 3 cho ta giá của một trang màtự chúng ta quét lấy ở tỉ lệ lương 4USD/giờ – không kể chi phí đầu tư.

Những phép tính này giả định rằng máy in được sử dụng vừa phải để điều chỉnh chi phí đầu tư. Ba cột cuối trong bảng 1 cho biết thêm thông tin về máy quét. Cột đầu tiên cho biết thông tin về chi phí máy quét. Cột kế tiếp cho biết tuổi thọ quét của máy quét. Cột cuối thể hiện số trang được quét cho mục đích thương mại, với chi phí 0.06USD/page tính trên mỗi máy quét.

Có nhiều nhân tố ảnh hưởng đến việc lựa chọn máy in: ngân sách, giảm thiểu sự lệ thuộc vào các đối tác, mong muốn tạo dựng nền tảng riêng, điều bắt buộc phải quét tự tài liệu, không vận chuyển v.v..

Các yếu tố trên đưa ra khối lượng trang cần thiết để điều chỉnh các mức đầu tư khác nhau. Rất ít khi một cơ quan hay một tổ chức cần quét 800.000 trang. Nếu ở mức quét như vậy thì sẽ có rất nhiều vấn đề nảy sinh, chẳng hạn như chi phí bảo trì, khả năng làm tăng chi phí gấp đôi vì phải nhờ dịch vụ khác thực hiện công việc quét tài liệu.

Người ta hay nghĩ rằng việc phát triển khả năng quét văn bản là một công việc kinh doanh, đặc biệt là ở những quốc gia đang phát triển. Nhưng chúng ta nên nhớ rằng công việc này không lặp lại; nghĩa là một khi tài liệu được quét xong thì khách hàng sẽ không bao giờ đặt những đơn đặt hàng mới để quét lại những tài liệu đó, bất kể họ có mối quan hệ thân thiện như thế nào với công ty quét. Từ quan điểm thương mại, quảng cáo mạnh mẽ là rất cần thiết. Chúng tôi không khuyên các tổ chức NGOs hay các tổ chức phi lợi nhuận khai thác vào chặng đường này mà không qua các thử nghiệm ban đầu hay một chiến lược kinh doanh được hoạch định cẩn thận.

Nói chung nếu chúng ta muốn quét từ 10.000 đến 50.000 trang thì nên nhờ đối tác thực hiện. Chi phí cho máy quét chuyên nghiệp low-end khoảng 6000USD chỉ có thể được điều chỉnh nếu như cần quét hơn 100.000 trang. Bạn có thể hợp tác với một vài tổ chức khác- có thể là NGOs hay các thư viện để mua một máy in như thế.

3. OCR: Nhận dạng kí tự

Nhận dạng kí tự hay còn gọi là hệ thống OCR làm công việc chuyển thể các hình ảnh được quét thành văn bản. Đầu vào là một hình ảnh kỹ thuật số ở định dạng TIFF hoặc Bitmap, tốt nhất là ảnh có chất lượng cao. Đầu ra là văn bản hoặc trang web, cơ bản là các định dạng RTF, Word hoặc HTML. Sau đây là các bước cơ bản để chuyển thể tài liệu giấy tờ thành dạng kỹ thuật số:

1. Quét tài liệu
2. Phân tích lề trang
3. Nhận dạng
4. Quét ảnh và các bảng

Tuân theo những bước này, bạn kiểm tra chất lượng các tập tin kết quả và lưu chúng ở định dạng thích hợp.

Trên thị trường có rất nhiều chương trình nhận dạng kí tự tốt với giá cả từ 100USD đến 400USD, chẳng hạn như:
- Read-Iris (http://www.readiris.com)
- Omnipage (http://www.omnipage.com)
- Fine-Reader (http://www.finereader.com)

Tất cả thông tin bao gồm cả nhà phân phối địa phương đều có thể được tìm thấy trên các website của các nhà sản xuất. Trong số này, theo kinh nghiệm của tác giả, hai phần mềm có giao diện người dùng thân thiện nhất là Fine-Reader và Omnipage. Fine-Reader là rẻ nhất, 100USD, linh hoạt và hỗ trợ nhiều ngôn ngữ nhất.

Để tự thực hiện công việc quét tài liệu cần có máy quét, phần mềm nhận dạng kí tự và sự đảm bảo về chất lượng. Do cần phải quyết định sự lựa chọn giữa tự quét hay ban giao cho đối tác thực hiện việc quét.

3.1 Tiến trình nhận dạng kí tự

Với mỗi chương trình nhận dạng kí tự thì tiến trình nhận dạng kí tự cũng khác nhau và cũng yêu cầu việc nghiên cứu sử dụng. Có bốn điểm cần đặc biệt chú ý trong tiến trình này là: quản lý chất lượng, các bảng, các hình ảnh và các tài liệu chuyên ngành chẳng hạn như các công thức, các kí tự nước khác…

Quản lý chất lượng

Chúng ta phải luôn kiểm tra chất lượng, Thông thường có 4 loại kiểm tra chất lượng.
Loại thứ nhất được thực hiện cùng lúc với giai đoạn nhận dạng kí tự. Mỗi chương trình nhận dạng thường sẵn có một bộ kiểm tra ngữ vựng, sẽ làm nổi bật những từ bi nghi ngờ có sai sót. Cùng thời điểm có ảnh của từ cũng xuất hiện vì vậy cũng sẽ làm cho việc kiểm tra và sửa lỗi dễ dàng hơn.

Loại thứ hai là kiểm tra tổng thể văn bản sau khi việc quét hoàn tất. Các lỗi thông thường bắt gặp trong giai đoạn này là mất trang, mất đoạn, các tiêu đề chương v.v..

Loại thứ ba là kiểm tra ngữ vựng sử dụng chương trình Microsoft Word. Chương trình này có một tự điển phong phú hơn và vì thế tốt hơn phần cài sẵn trong các chương trình nhận dạng kí tự. Tài liệu sau khi quét sẽ được mở bằng Word để kiểm tra ngữ vựng, sẽ có nhiều lỗi được phát hiện và chỉnh sửa hơn. Nên thêm những từ ngữ phức tạp cho bộ kiểm tra ngữ vựng.

Loại cuối cùng là tài liệu sau khi hoàn tất 3 khâu kiểm tra trên sẽ được kiểm tra lần cuối bởi một người khác. Người này sẽ lấy mẫu tài liệu và kiểm tra lỗi, các vấn đề với cấu trúc bảng, hình ảnh, các thẻ và kiểm tra tổng quát tài liệu văn bản kết quả. Chỉ sau lần kiểm tra này thì tài liệu mới được xem là sẵn sàng cho giai đoạn sau.

Bảng

Các chương trình nhận dạng kí tự thường không xử lý tốt các bảng. Hơn nữa các bảng rất khó kiểm tra. Chúng có nhiều kí tự, đôi lúc có cả các dấu chấm, dấu phẩy và các mục dễ nằm sai hàng, sai cột. Điều này đòi hỏi việc kiểm tra phải thật cẩn thận và đảm bảo thật tốt chất lượng. Có 3 cách để kiểm tra:

Cách thứ nhất, xem bảng như hình ảnh, nghĩa là quét chúng dưới dạng các hình ảnh trắng đen và đặt chúng vào vị trí thích hợp trong tài liệu. Đây là giải pháp đơn giản nhất. Sẽ không có lỗi và thời gian duy nhất cần thiết là thời gian để tạo ra hình ảnh của bảng. Tuy nhiên giải pháp này tiêu tốn nhiều bộ nhớ máy tính hơn và độ phân giải cũng thường không đủ khi các bảng lớn được hiển thị trên màn hình máy tính. Nếu như bạn tạo ra hình ảnh các bảng vừa vặn thì độ phân giải lại quá nhỏ. Nếu như bạn cố làm bảng lớn hơn thì người sử dụng phải cuộn lên xuống để xem các cột, các hàng và do đó không có được cái nhìn tổng quát nội dung bảng.

Cách thứ hai, các bảng có thể được tái tạo lại bằng cách tạo ra một bảng mới có cùng số hàng số cột và đánh nội dung bảng vào.

Cách thứ ba là bảng có thể được nhận dạng như là các kí tự. Cách này sẽ tiết kiệm thời gian so với tái tạo lại bảng nhưng dễ gây ra lỗi sau này. Đôi lúc

các cột có thể được nối lại và các dấu phẩy, dấu chấm không được nhận dạng.

Hình ảnh

Việc xuất bản bao gồm 3 loại hình ảnh sau:
- Các đường trang trí trắng đen
- Các hình ảnh trắng đen
- Các hình ảnh màu

Các đường trang trí trắng đen nên được quét ở chế độ thích hợp và được lưu ở dạng tập tin GIF hoặc PNG. Hình ảnh trắng đen nên được quét trong chế độ xám và được lưu ở dạng tập tin GIF hoặc JPEG. Các hình ảnh màu được quét ở chế độ màu và lưu ở dạng tập tin JPEG. Thông thường các hình ảnh JPEG chất lượng trung bình sẽ cho độ phân giải phù hợp.

Hình ảnh thường chiếm nhiều bộ nhớ của đĩa cứng hoặc CD. Vì vậy cần phải cân bằng ba đặc tính: độ rõ, tầm nhìn và kích thước của ảnh. Để giảm bộ nhớ, bạn có thể bỏ đi các hình ảnh không phù hợp lắm với văn bản.

Hình ảnh nên được quét riêng biệt từng cái một. Chúng tôi khuyến khích tên tập tin ảnh tuân theo dạng sau: 5 hay 6 kí tự đầu chỉ định tài liệu nào, theo sau là vị trí trang chứa ảnh. Cách khác, giả định như mỗi tài liệu nằm trong thư mục riêng, chỉ đơn giản dùng kí tự p theo sau là số thứ tự trang chứa ảnh. Nếu như có nhiều ảnh trên cùng một trang thì hãy nối thêm các kí tự a,b,c,… vào tên tập tin ảnh. Ví dụ, nếu như một ảnh JPEG nằm trên trang 36 của tài liệu u7548e thì nó sẽ có tên tập tin là u7548e36.jpg hay p36.jpg. Một khi các bức ảnh đã được quét xong bạn có thể sử dụng chương trình để định dạng kích cỡ hoặc làm chúng đẹp hơn cùng một lúc

Các tài liệu chuyên ngành

Nhiều tài liệu chuyên ngành chứa các kí tự đặc biệt, các công thức, và các trang phức tạp. Các kí tự đặc biệt thường là từ các ngôn ngữ khác hoặc các dấu đặc biệt. Cần phải thiết lập tùy ngôn ngữ thích hợp cho chương trình nhận dạng kí tự. Các công thức sẽ phải được tái tạo lại. Các tài liệu phức tạp hoặc bị hư phải được đánh lại.

3.2 Hiệu năng và các tài nguyên

Như đã bàn từ trước bạn không nên đánh giá thấp những khó khăn gặp phải trong tiến trình nhận dạng kí tự. Mặc dù cần phải xem xét tính kinh tế và thực tế cho tiến trình nhận dạng kí tự một cách riêng biệt với tiến trình quét tài liệu nhưng một số điểm tương tự cũng nảy sinh: sự đầu tư cần thiết vào máy tính, tài nguyên con người và các kỹ năng quản lý; đào tạo nhân lực, chi phí lương, tổng số trang cần xử lý và tài liệu có được bàn giao cho các đối tác khác không.

Trong phần này chúng tôi sẽ chia sẻ kinh nghiệm trong tiến trình nhận dạng kí tự ở Belgium, Romania và Aán Độ. Hầu hết các trường hợp, các tính toán và các phỏng đoán chỉ dựa trên số tình huống chung, các tài liệu có độ phức tạp gồm bảng và hình ảnh chẳng hạn như các tài liệu ở thư viện. Tiến trình nhận dạng kí tự rất khó được thực hiện hoàn hảo. Các kết quả tốt thường đạt được trong những giờ làm việc đầu tiên của mỗi ngày. Sau ba giờ làm công việc kiểm tra trong tiến trình nhận dạng kí tự thì hiệu năng giảm rõ rệt, giảm đến 50% so với mức trong những giờ đầu. Cũng vậy, các tuần đầu thường cho kết quả tốt hơn những tuần sau đó. Việc kiểm tra chất lượng sẽ được người bản xứ thực hiện tốt hơn và người trẻ cũng tập trung cao hơn người lớn tuổi hơn (thường là từ 18 đến 23 tuổi tốt hơn trên 25 tuổi).

Và cuối cùng là công việc trong tiến trình nhận dạng kí tự là một công việc nhàn chán, vì thế sự tập trung là điều hết sức quan trọng.

Một số hướng dẫn trong việc tổ chức tiến trình nhận dạng kí tự:

- Chọn người trẻ từ 18-25 tuổi
- Do hiệu năng tốt chỉ đạt được trong những giờ đầu, nên hoặc là tổ chức làm bán thời gian hoặc là chỉ làm toàn thời gian đối với những người có mức tập trung cao và lâu dài.
- Hai phần ba nhân lực có khuynh hướng bỏ cuộc sau khoảng 3-5 tuần. Điều này giải thích vì sao chất lượng và hiệu năng kém hẳn vào những tuần sau đó.
- Cần phải đào tạo và duy trì số lượng nhân viên đều đặn để đảm bảo chất lượng.

	Giờ làm việc/ngày	Số trang/ngày	Số trang/tháng
Thời gian huấn luyện ban đầu (6 tuần)	3	6	120
Hiệu năng của quá trình	3	9	150 đến 200
	7	28	500 đến 600

Bảng 2: OCR productivity

Bảng 2 mô tả hiệu năng của quá trình nhận dạng kí tự. Các tài liệu thuộc đủ loại kích thước chất lượng. Các tính toán này giả định rằng các tài liệu có số lượng trung bình các hình ảnh và bảng, chẳng hạn như có một hình và một bảng 5 hàng 5 cột trên mỗi 8 trang tài liệu và các ảnh tài liệu có chất lượng trunh bình; điều này phụ thuộc vào chất lượng quét cũng như khả năng ngôn ngữ của những người tham gia vào tiến trình nhận dạng kí tự.

Tuy nhiên tỉ lệ của những trang phức tạp có chất lượng thấp gồm nhiều cột hoặc nhiều bảng là rất ít, khoảng 300-400 trang/tháng nếu làm toàn thời gian.

Giả sử chi phí lương cho nhân viên làm toàn thời gian trong tiến trình quét tài liệu là 400USD/tháng và các chi phí ngoài dự tính, máy tính, phòng làm việc, các công cụ sẽ thêm khoảng 300USD/tháng. Vì vậy chi phí cho một trang tài liệu được trong tiến trình nhận dạng kí tự là 1.2USD-1.6USD/trang. Nếu quan tâm đến chi phí huấn luyện, lượng thời gian, khoảng thời gian dự trữ, chi phí tìm nhân viên mới khi thiếu hụt nhân lực thì chi phí cho một trang sẽ gia tăng lên từ 1.5USD-2.5USD/trang.

Chi phí của việc tự quét tài liệu với việc tài liệu được quét bởi đối tác cũng nên được so sánh. Thông thường những công ty này sẽ ra giá từ 1.5USD-4USD/trang, bao gồm cả các trang có hình ảnh và bảng. Công ty Human Info NGO/Sinple World có một chi nhánh như thế ở Rumani và tính với chi phí ưu đãi cho các tổ chức phi lợi nhuận từ 1.2USD-2USD/trang. Để biết thêm thông tin xin liên hệ tại scanning@humnaninfo.org.

3.3 Các hình thức khác trong tiến trình nhận dạng kí tự

Tự đánh máy

Hình thức này sử dụng một bộ xử lý văn bản để đánh lại các tài liệu. Tuy vậy việc này vẫn cần phải quét các hình ảnh và trang bìa nhưng các trang còn lại thì không cần quét.

Những người làm dưới hình thức này không cần hiểu tài liệu văn bản. Họ chỉ cần đánh máy lại một cách chính xác những gì họ thấy. Hình thức này cần thiết có 2 người làm việc độc lập trên cùng các trang tài liệu để sau đó đối chiếu các trang.

Ý tưởng giả định từ sự đối chiếu này là nếu một từ được 2 người đánh độc lập mà giống nhau thì từ đó được đánh đúng. Tuy nhiên điều này không phải luôn đúng; sẽ cực kỳ chính xác nếu như có ba người cùng đánh các tài liệu một cách độc lập.

Thuận lợi của việc đánh máy lại tài liệu là không cần, chi phí cho các chương trính nhận dạng kí tự và các máy tính thì không cần phải mạnh. Ngược lại nếu trong trường hợp sử dụng chương trình nhận dạng kí tự thì cần phải có máy tính mạnh. Và hình thức này không cần nhân viên có kỹ năng cao. Tuy nhiên điều bất lợi là cần phải có một khóa huấn luyện ít nhất 2 tháng. Chi phí phụ thuộc hoàn toàn vào mức phát lương.

Các tập tin hình ảnh

Một hình thức khác có chi phí rất thấp trong tiến trình nhận dạng kí tự là tạo ra các trang hình ảnh ở định dạng tài liệu PDF. Chi phí khoảng 0.1USD/trang.

Sau khi quét tài liệu ta sẽ có các tập tin dạng TIFF, sử dụng phần mềm chuyển đổi tự động để chuyển tất cả các tập tin TIFF này thành các tập tin PDF. Bất lợi là các tập tin này không thể tìm kiếm được. Và chúng là khá nặng, khoảng 50Kb/trang nên download rất lâu từ đường truyền Internet tốc độ chậm và 20% chất lượng phụ thuộc vào các tập tin TIFF ban đầu. Các tập tin PDF thì lớn và không hổ trợ các thao tác văn bản trên tài liệu chẳng hạn như “cắt và dán”. Hình thức này chỉ nên được sử dụng nếu ngân sách dành

cho tiến trình nhận dạng kí tự eo hẹp và chỉ có một số ít đối tượng sử dụng có truy cập Internet ở tốc độ thấp.

3.4 Kết hợp giữa việc quét và nhận dạng kí tự

Nếu như việc máy quét được kết nối trực tiếp và máy tính có cài chương trình nhận dạng kí tự thì hầu hết các chương trình này đều có khả năng quét và nhận dạng một trang ngay lập tức, nhưng sẽ mất nhiều thời gian nếu như số lượng trang lớn. Khoảng từ 100-150 trang/tháng thì giải pháp này là không khả thi. Đối với số lượng tài liệu lớn thì nên quét xong tất cả các tài liệu rồi mới thực hiện công đoạn nhận dạng kí tự.

4. BA VÍ DỤ: TỪ 1000 ĐẾN 100,0000 TRANG

4.1 Tập hợp nhỏ: 500-1000 trang

Hầu hết các tổ chức NGOs có từ 500-1000 trang để quét. Số lượng này có thể được tự chúng ta thực hiện công việc nhận dạng kí tự nếu như có nhân lực thích hợp.

Việc quét

Bước đầu tiên là quét tài liệu để tạo ra các tập tin dạng TIFF chất lượng cao và các hình ảnh Bitmap màu, thang độ xám cho những hình minh họa. Giả sử cần phải quét 1000 trang thì điều này cần công việc bán thời gian khoảng một tháng chí cho công việc quét. Các tập tin ảnh TIFF tốn khoảng 60Mb-80Mb ổ đĩa cứng và giải pháp tốt là tạo ra các đĩa CD-ROM chứa chúng. Một máy quét phẳng giá khoảng 100USD-300USD là đủ.

Nhận dạng kí tự

Bước thứ hai là nhận dạng kí tự. Thông thường mất khoảng 5 hay 6 tháng cho nếu làm bán thời gian (20 giờ/tuần) để chuyển khoảng 1000 trang thành các tài liệu Word hay HTML.
Xuất sang đối tác
Một hình thức khác là nhờ đối tác thực hiện công đoạn quét và nhận dạng kí tự. Chi phí khoảng 1500-2000USD để chuyển đổi toàn bộ số lượng tài liệu trên thành tài liệu Word hoặc HTML.

4.2 Toàn bộ tài liệu từ một tổ chức: 5000 trang

Nhiều tổ chức lớn có khoảng 5000 trang tài liệu, các bài báo…

Việc quét

Sử dụng máy quét ở đây là không thích hợp. Công việc quét có thể nhờ bên thứ ba (khoảng 400USD/5000 trang) hoặc sử dụng máy in có ngăn (khoảng 900USD). Một cách khác là có thể chung tiền với một số tổ chức khác để mua một máy quét tốt hơn (6000USD được chia cho số đơn vị tham gia). Toàn bộ 5000 trang ở dạng TIFF sẽ tốn khoảng 30Mb-40Mb ổ đĩa cứng. Và giải pháp tốt là sử dụng đĩa CD-ROM.

Nhận dạng kí tự

Bước thứ hai là nhận dạng kí tự. Mất khoảng 25-30 tháng cho công việc bán thời gian để chuyển 5000 trang tài liệu thành dạng tài liệu Word hoặc HTML. Tổ chức phải trả cho những người quét và những người giám sát hiệu suất và chất lượng.
Xuất sang đối tác
Hình thức khác là thuê đối tác thực hiện việc quét và nhận dạng kí tự. Tốn khoảng 7500-10000USD để chuyển đổi toàn bộ sang tài liệu Word hoặc HTMl.

4.3 Thư viện nhỏ: 100,000 trang

Các tổ chức lớn, trường đại học, chính phủ và các thư viện có thể có khoảng 100,000 trang tài liệu. Vấn đề cần xem xét trước tiên là bản quyền tài liệu. Nếu chúng không phải tài liệu được phép phổ dụng thì cần phải xin phép bản quyền từ những người giữ bản quyền của chúng. Bạn cũng nên kiểm tra xem các tài liệu đã sẵn có ở bản điện tử chưa.

Việc quét

Máy quét có ngăn cũng không thích hợp cho số lượng tài liệu này. Việc quét có thể được đối tác làm (8000USD/100,0000 trang). 100,000 trang tài liệu tốn khoảng 6Gb-8Gb ổ đĩa cứng, và giải pháp là sử dụng đĩa CDROM để lưu trữ bản điện tử.

Nhận dạng kí tư

Mất khoảng 500-700 tháng để chuyển 100,000 trang tài liệu thành tài liệu WORD hoặc HTML. Điều này là không khả thi.
Để tiết kiệm chi phí, các trang ít phổ biến (80%, 80000 trang) có thể được chuyển thành dạng tập tin PDF. Tài liệu ở dạng PDF có thể chiếm khoảng 4-6Gb và khó khăn cho việc download từ Internet, nhưng có thể tiết kiệm 0.2USD/trang. Mất khoảng 10-20 tháng cho công việc bán thời gian để chuyển 80,000 trang tài liệu thành dạng PDF trên một máy tính mạnh.

Xuất sang đối tác

Nếu có 80% tài liệu PDF và 20% tài liệu HTML thì tài liệu PDF sẽ tốn khoảng 16,000USD và tài liệu HTML khoảng 30,000-40,000USD, tổng chi phí khoảng 50,000USD. Nếu như toàn bộ tài liệu được đem đi nhận dạng thì tốn khoảng 150,000-200,000USD để chuyển chúng thành các tài liệu Word hay HTMl.

5. TẠO RA MỘT BỘ SƯU TẬP ĐIỆN TỬ

Có ba khía cạnh quan trọng trước khi tạo ra tập hợp tài liệu điện tử . Thứ nhất tập hợp này phải được tổ chức. Nội dung càng nhiều thì nhu cầu lập chỉ mục và tìm kiếm càng cao. Đối với các bộ sưu tập gồm khoảng 3000-5000 trang tài liệu thì việc lập chỉ mục và hệ thống tìm kiếm là rất cần thiết. Thứ hai là nhu cầu của người dùng đầu cuối là không ngừng thay đổi. Vì vậy cần phải xác định các nhóm đối tượng sử dụng tài liệu và cần có tiến trình cải tạo thường xuyên. Thứ ba là chi phí cho dự án là bao nhiêu.

5.1 Các phương pháp xây dựng tập hợp

Có nhiều CD-ROM rất đẹp dùng trang Web để trình bày nội dung. Các tài liệu HTML, PDF hoặc Word được gắn kết vào các mục liên kết. Việc đọc tài liệu trở nên đơn giản, lôi cuốn nhờ các kết nối, các frame, từ khoá, chỉ mục v.v.. Những hệ thống như thể chỉ phù hợp cho số lượng tài liệu khoảng vài ngàn trang; nhưng nếu số lượng là từ 3000-5000 trang hoặc hơn thì tập hợp tài liệu cần phải được tổ chức tốt và có công cụ hỗ trợ tìm kiếm thông tin. Phần mềm Greenstone có thể làm được việc này.

Phần mềm Thư viện số Greenstone sẽ tạo ra Thư viện số có cấu trúc bao gồm công cụ tra cứu tài liệu. 150,000 trang có thể được lập chỉ mục và được lưu trữ trên đĩa CD-ROM. Greenstone là phần mềm mã nguồn mở và sẵn có trong mục GNU.

Các tài liệu hướng dẫn đi kèm giới thiệu các thức xây dựng tập hợp tài liệu cho Greenstone. Các tập hợp tài liệu nhỏ có thể được xây dựng tương tác bằng cách sử dụng hệ thống con gọi là "Bộ thu thập" được mô tả trong tài liệu hướng dẫn người dùng của phần mềm, hướng dẫn bạn thông qua một loạt các trang tương tác để yêu cầu thông tin cần thiết. Riêng đối tập hợp tài liệu lớn và phức tạp thì chúng tôi khuyến khích bạn nên sử dụng tiến trình xây dựng dòng lệnh được mô tả trong tài liệu hướng dẫn đi kèm. Bạn cần phải đọc tài liệu hướng dẫn trong chương 2 để vận dụng phần mềm xây dựng các tập hợp tài liệu cao cấp, phức tạp.

5.2 Công cụ tổ chức

Greenstone được sử dụng để xây dựng nhiều tập hợp tài liệu cho nhiều mục đích, các tài liệu có cùng cấu trúc và tổ chức nhưng có nội dung khác nhau. Các tập hợp mẫu như "Development Library Subset-DLS" hay "Demo" là thuộc dạng này.

Công cụ tổ chức (Organizer) là một chương trình đi kèm với Greenstone nhằm tạo ra các tập hợp có cấu trúc và tổ chức giống tập hợp mẫu DLS. Nhờ những tập hợp như thế sẽ làm đơn giản hóa tiến trình xây dựng dòng lệnh của Greenstone. Organizer là một ứng dụng được tạo ra bởi Microsoft Visual C++ và vì vậy chỉ giới hạn sử dụng trong Windows.

Organizer được thiết kế để giúp quản lý các khía cạnh tổ chức một tập hợp Thư viện số: nhập tiêu đề tài liệu, phân loại chủ đề và các siêu dữ liệu, chỉnh sửa chúng v.v.. Nó làm việc với phần mềm Thư viện số Greenstone khi xây dựng các tập hợp: collect.cfg, metadata.xml, sub.txt, org.txt, Keyword.txt và AZList.txt. Cấu trúc và nội dung của những tập tin này được giải thích trong tài liệu hướng dẫn người dùng đi kèm.

Metadata.xml được đọc bởi bộ plug-in đi kèm Rec-Plug được thảo luận tại phần cuối 2.1, tài liệu hướng dẫn. Các tập tin sub.txt, org.txt và AZList.txt định nghĩa các cấu trúc phân cấp được sử dụng bởi bộ phân cấp. 5 tập tin

này gắn kết với tập hợp mẫu DLS và các tập hợp tương tự khác thông qua tập tin cấu hình tập hợp Collect.cfg.
Metadata.xml được đọc bởi RedPlug được chỉ định trong tập tin cấu hình.

Nếu chúng ta muốn xây dựng tập hợp hình ảnh và các chức năng phân loại thông thường gặp khó khăn trong việc làm việc với những tập tin này. Và mục tiêu của Organizer là giúp định nghĩa nội dụng của các tập tin này- thực ra nó tạo ra các tập tin này cho bạn. Tuy nhiên hiện tại Organizer chỉ được dùng giới hạn cho mô hình tài liệu được mô tả dưới đây; vì vậy nếu người dùng muốn thêm các thông tin tổ chức không có trong mô hình này sẽ phải tự tay chỉnh sửa trong các tập tin kết quả của Organizer.

Cài đặt và sử dụng Organizer

Vào cuối phần cài đặt bộ phần mềm Greenstone, bạn sẽ được nhắc là có cài Organizer hay không. Nếu trả lời "yes" thì tiến trình cài đặt Organizer sẽ bắt đầu. Nếu bạn trả lời "No" hoặc bạn muốn cài đặt lại chúng sau này thì bạn đến thư mục windows_utilities trên đĩa CD-ROM Greenstone và thực thi chương trình Organizer.exe.

Khi chạy chương trình bằng cách chọn Organizer dưới mục Greenstone Digital Library trong menu Programs của menu Start của Windows. Sau đó bạn được yêu cầu chỉ định cơ sở dữ liệu (CSDL); hãy sử dụng CSDL DLS bằng cách chọn tập tin dls.mdb. Bạn sẽ được yêu cầu nhập username và password: nhập vào admin cho cả hai.

Mỗi CSDL trong Organizer chứa đựng một hoặc nhiều tập hợp tài liệu. Khi lần đầu tiên bạn chạy Organizer bạn sẽ được trình bày một danh sách các tập hợp chứa đựng một mục duy nhất: tập hợp DLS. Nói chung rất dễ dàng sử dụng Organizer để thêm vào các tập hợp mới cho CSDL mặc định, kể cả việc chỉnh sửacác tập hợp con trong các tập hợp trước đó.
Mỗi CSDl và mỗi tập hợp có 3 danh sách toàn cục chính: (Global List):

- Tài liệu: tài liệu được nạp vào tập hợp
- Các hình thức tổ chức: tên các hình thức tổ chức gắn kết với các tài liệu.
- Các chủ đề: Các cụm từ phân loại chủ đề sẵn có cho việc xây dựng các cây phân cấp tài liệu theo chủ đề.

Một số thuộc tính khác có thể được gán vào mô hình tài liệu. Xin chú ý rằng những thay đổi trong CSDL sẽ được tự động lưu lại trước khi thoát chương

trình Organizer, mà không nhắc người dùng có đồng ý lưu lại hay không. Vì vậy nếu bạn muốn lưu lại bản lưu trữ thì bạn cần phải sao chép nội dung CSDL cần lưu trữ ngay khi bạn vừa mở chương trình Organizer. Chú ý không có chức năng backup cho từng tập hợp tài liệu riêng rẽ.

Mô hình tài liệu

Trong Organizer, mỗi tài liệu được gán thông tin siêu dữ liệu (metadata) từ một tập hợp các thuộc tính sau:

i. Tiêu đề
ii. Hình thức tổ chức (thuộc tính này có thể lặp lại trình bày thông tin nhà sản xuất nhưng cũng có thể là các tác giả hoặc tổ chức liên quan đến tài liệu)
iii. Chủ đề (gán mục phân cấp)
iv. Các từ khóa : các từ ngữ được định nghĩa bởi người dùng để giúp lập chỉ mục tài liệu – trong tập hợp DLS nó được sử dụng để phân loại tài liệu ứng với các câu hỏi phổ biến được định nghĩa trước
v. Nhan đề: tiêu đề của tài liệu
vi. Ngày xuất bản
vii. Số trang
viii. Mã tác vụ: định danh tài liệu, được sử dụng bởi Greenstone để liên kết các siêu dữ liệu vào các văn bản, hình ảnh của tài liệu tại tập hợp đang xây dựng
ix. Ngôn ngữ
x. Khối: định danh của nhóm tài liệu được xử lý chung với nhau
xi. Tập hợp được yêu cầu: tập hợp tài liệu đại diện cho việc gán
xii. Thông tin bản quyền
xiii. Mã bản quyền

5 thông tin đặc biệt ở chỗ Greenstone sẽ sử dụng chúng để xây dựng và hiển thị các chỉ mục để truy xuất tài liệu. Các thuộc tình từ vi đến viii không được lập chỉ mục nhưng được đính kèm như các định danh phụ của tài liệu. Các thuộc tình từ ix đến xiii chỉ được sử dụng cho mục tiêu quản lý tài liệu bên trong của phần mềm Organizer.

Siêu thông tin về chủ đề tài liệu đặc biệt ở chỗ chúng là tính chất của tài liệu trong tập hợp chứ không phải tài liệu riêng rẽ. Khi một tài liệu được di chuyển từ tập hợp này đến tập hợp khác toàn bộ siêu dữ liệu liên quan cũng di chuyển theo ngoại trừ thông tin siêu dữ liệu về chủ đề sẽ được gán lại.

Không giống như việc người lập trình Thư viện số muốn sử dụng thông tin siêu dữ liệu trong ý nghĩa tài liệu thông thường; vì vậy ngoài việc phân loại theo chủ đề tất cả các tài liệu đều có thể được truy xuất đến trong Greenstone thông qua công cụ tìm kiếm toàn văn. Thuộc tính này có thể được xem như thuộc tính thêm vào cho việc lập chỉ mục. Ví dụ đối với danh mục tác giả (Không biết trước rõ ràng ở mẫu hiện tại) hoặc dữ liệu nguồn bằng cách thay thế dữ liệu liên quan trong từ khoá metadata. Sau đó nút dùng để phục hồi dữ liệu thông qua việc đặt lại thuộc tính mới (Bằng nhãn "how to" ở chế độ mặc định trong giao diện thư viện). Cũng có thể chọn lại bằng cách thay đổi từ "How to" thành "Author" hay "Country" ở dòng 5 của tập tin *collect.cfg* được tạo ra ở phần Organizer.

Greenstone cũng có thể tạo ra Bảng mục lục để truy cập vào các phần lớn và phần nhỏ trong mỗi tài liệu. Việc này không được điều khiển trực tiếp bằng Organizer nhưng buộc phải hoàn tất thông qua việc thêm vào phần nội dung những mục lớn và những mục nhỏ, được miêu tả trong phần 5.3 của chương này.

Nó được nhắc rằng tự bản thân Greenstone cũng rất linh hoạt trong các mẫu tài liệu sẵn có trong các thư mục, nhưng phải được tạo ra chính thức trong thư mục của các hàm metadata và được mã hoá trong các file xây dựng bộ sưu tập được miêu tả trong phần 2.2 của tài liệu *Greenstone Developer's Guide*. Nếu mẫu được yêu cầu có liên hệ gần với mẫu DLS, ta có thể bắt đầu với các file xây dựng bộ sưu tập được tạo ra bởi Organizer, và chỉnh sữa chúng theo các quy tắt của Greenstone.

Tìm hiểu chức năng tổ chức (Exploring the Organizer)

Đây là chi tiết mỗi đặc tính của chức năng Organizer

i. Cửa sổ chính của Organizer
   Cửa sổ này được trình bày khi cơ sở dữ liệu đã được chọn. Nó gồm ba phần:
   - Thanh **Horizontal menu** ngay trên đỉnh, bao gồm phần phía dưới shortcut horizontal
   - Vertical toolbar ở phía bên trái
   - Phần chính giữa (central area) là nơi trình bày phần nội dung và chức năng được chọn trên thanh **vertical**.

   a. Thanh **Horizontal menu** chứa tám menu, ứng với mỗi menu chứa một hay nhiều dòng lệnh:

- File Menu chỉ cung cấp một lệnh: Exit để thoát khỏi chương trình Organizer.

- New Menu cung cấp những lệnh để thêm hoặc thay đổi bộ sưu tập, Tài liệu, Cách thức tổ chức, các chủ đề và để thêm hoặc thay đổi danh mục các thuộc tính khác trong tài liệu mẫu. Các hộp hội thoại đối với dòng lệnh này phải tường minh, với phần bổ sung mô tả chi tiết sau:

  Lệnh *New/Organization* trình bày hộp thoại gọi là *Edit Organization Name* mà cung cấp những vùng để nhập tên đầy đủ tổ chức và tên tóm tắt (Nếu chức năng *Auto complete* được bật, nó sẽ tự động đặt tên bằng ký tự đầu tiên của mỗi từ đầu viết hoa trong tài liệu). **Nếu chức năng *Auto complete* không thể trình bày tên viết tắt** (Bởi vì không có ký tự nào trong tiêu đề hoặc là đã tồn tại một tên viết tắt trong cơ sở dữ liệu), **hộp hội thoại *Auto complete* phải không được chọn và tên viết tắt phải được nhập vào bằng tay.**

  Lệnh *New/Subject* trình bày một hộp hội thoại *Add New Subject* để thêm vào một danh sách các chủ đề toàn cầu (không phân lớp) trong cơ sở dữ liệu mà các thành phần luôn theo sau nhau liên tiếp trong cấu trúc phân lớp các chủ đề ứng với bộ sưu tập đưa ra (các đề tài được trình bày trong của sổ thuộc tính bộ sưu tập).
  Hộp hội thoại này cũng có thể được dùng để tạo ra các từ khóa trong danh mục các chủ đề toàn cầu mà được đánh dấu liên tiếp với i) thuộc tính của các loại đề tài khác nhau sẽ tạo điều kiện cho chức năng phục hồi dữ liệu được thêm vào trong bộ sưu tập (Chức năng *Add subject* các đề tài được nhìn thấy trong cửa sổ *Collection properties,* xem dưới đây, and/or với ii) một thuộc tính đưa vào bộ tài liệu chuyên đề (*Keywords* tab trong hộp hội thoại *Document properties* của những tài liệu được nhìn thấy trong cửa sổ *Collection properties*, xem phần dưới đây).

- Menu *delete* cung cấp lệnh xoá bộ sưu tập, bộ tài liệu, các tổ chức, các đề tài, và các mục trong danh mục các giá trị có thể của các thuộc tính khác trong mẫu tài liệu.

- Menu *Edit* cung cấp lệnh *copy* (Giống lệnh Ctl C) khi copy khung chính giữa và các dòng được chọn vào clipboard để dán vào các file hoặc các mục Organizer khác.

- Menu *View* cung cấp các lệnh để mở các bộ sưu tập, bộ tài liệu, các tổ chức hoặc các đề tài trong các danh mục ở phần trung tâm (Lần lượt sử dụng 4 nút đầu tiên trên thanh toolbar đứng). Nó cũng chứa các hộp *bật*/*tắt* để trình bày thanh công cụ short-cut (Được miêu tả chi tiết dưới đây) trong thanh menu và thanh status ở khung chính giữa.

- Menu *Tools* chứa các phần sau:
  - Lệnh chỉnh sửa tài liệu chứa hai công cụ: thứ nhất thay đổi tên của một tổ chức này thành tổ chức khác trong phần thuộc tính tổ chức của tất cả các tài liệu trong cơ sở dữ liệu (Tên mới phải được thêm trước tiên vào trong danh mục các tổ chức), hoặc thay đổi/thêm vào thuộc tính mới trong nhóm tài liệu được chọn trong danh mục các tài liệu (Công cụ kế tiếp là hoạt động nếu các tài liệu được thêm vào liên quan đến những tài liệu đã được chọn trong phần trình bày trang tài liệu và ngay cả những phần khác đã được kích hoạt.

  – Lệnh thứ hai cũng chứa hai công cụ. Một là công cụ *Statistics* cho phép thống kê thông tin về các bộ sưu tập và một công cụ Export lists để truy xuất ra ngoài thành dạng file text, một danh mục các tên tài liệu hoặc một trong ba thuộc tính chính của bộ sưu tập (Các danh mục này được lưu với một tên chuẩn trong mục \Program Files\Human Info\Organizer\DataBase.

  – Lệnh backup được dùng để tạo ra một file backup của toàn bộ cơ sở dữ liệu mỗi khi đươc chọn (cơ sở dữ liệu được lưu với một tên chuẩn), có ghi lại ngày, giờ backup trong mục \Program Files\Human Info\Organizer\DataBase. Phần xác nhận không trình bày những tài liệu không được yêu cầu.

- Menu *Administration tools* cung cấp các cửa sổ để thêm vào user mới và xoá một user hay thay đổi password. Chỉ một quản trị viên (e.g. "admin") mới có thể cấp quyền cho một user có thể truy cập vào cơ sở dữ liệu dưới tên là "Administrator" hay "Guest" và thay đổi password. Các user và administrator đều có quyền như nhau trong bộ sưu tập.

- Menu *Help* chứa một chuẩn Microsoft để trợ giúp trong thư mục *Help Topics* và thông tin về phiên bản và người phát triển thông qua lệnh *About Organizer.*

  Một thanh công cụ short-cut dưới thanh thực đơn horizontal cung cấp một tập các biểu tượng có thể thay đổi một các linh hoạt tùy thuộc vào phần trình bày được mở ra trong khu trung tâm (Được trình bày bởi menu *View* – xem phần trên – hoặc trên thanh công cụ vertical – xem phần dưới).

  Click vào biểu tượng thứ hai từ trái sang hay double click vào đề mục được chọn trong danh sách được chọn để chỉnh sửa các đề mục đó (xem phần dưới), cũng click như thế lên dòng đầu của cột đầu tiên trong danh sách được trình bày để sắp xếp. Các biểu tượng thư nhất, thứ ba, và cuối cùng có cùng chức năng với ba lệnh của thanh thực đơn được mô tả bên trên:

  Edit/Copy; tạo thêm một thành phần mới trong danh mục được trình bày: New/Collection/Empty,New/Document, New/Organisation hay New/Subject; và Help/About Organizer.
  Biểu tượng thứ hai trình bày liên kết với một biểu tượng thứ tư: cho phép thay đổi tên của bộ sưu tập được chọn trong danh sách bộ sưu tập (Không có trong menu chính) hoặc xuất ra danh sách các tài liệu (cũng giống như *Tools/Collections/Export Lists* với lựa chọn danh mục tài liệu, ngoại trừ việc chọn thư mục gốc và tên file).

b. Thanh **công cụ đứng** chứa 5 nút để chọn lựa phần nội dung và chức năng ở phần chính giữa của màn hình.
Nút thứ tư của toolbar phía trên cùng dùng để trình bày phần chỉnh sửa một trong trong số các phần sau ở vùng giữa khung:

Collections, *Documents*, *Organisations* or *Subjects* (các danh mục này cũng có thể được sử dụng bằng cách kích hoạt dòng lệnh tương ứng được trong menu *view* ở thanh công cụ nằm ngang):

- Danh mục các bộ sưu tập: chọn nút *Collections* (Được để ở chế độ mặc định trước khi nhập dữ liệu) trình bày tất cả các bộ sưu tập hiện hành trong cơ sở dữ liệu. Nhấp đúp vào tên của bất kỳ bộ sưu tập để xem phần thuộc tính trong cửa sổ *Collection Properties* (Được miêu tả dưới đây) trong đó bạn có thể thêm vào tài liệu mới cho bộ sưu tập hoặc thêm/chỉnh sửa thuộc tính tài liệu của bộ sưu tập. Để tạo ra bộ sưu tập mới, sử dụng lệnh *New/Collection* ở thanh menu ở đỉnh hoặc biểu tượng thứ ba từ trái sang trong thanh công cụ tắt (Xem phần trên).

- Danh mục các tài liệu: chọn nút *Documents* trình bày danh sách tất cả các tài liệu có chủ đề toàn cầu trong cơ sở dữ liệu (Các tài liệu này lấy từ bất kỳ các tài liệu nào trong số các tài liệu có liên kết, hoặc chưa được kết nối với bộ sưu tập). Nhấp đúp lên tên của bất kỳ tài liệu nào để xem/thay đổi thuộc tính của tài liệu đó nhanh hơn là phân chia chủ đề trong cửa sổ thuộc tính bộ sưu tập (Được miêu tả phía trên). Để tạo ra tài liệu mới với thuộc tính mới, sử dụng lệnh *New/ Document* ở thanh menu ở đỉnh hoặc biểu tượng thứ ba từ trái sang trong thanh công cụ tắt (Xem phần trên).

  Tìm kiếm chuỗi ký tự: có thể tìm kiếm trong danh mục có chứa một từ hay một chuỗi ký tự bằng cách nhập vào một từ trong hộp hội thoại nhỏ phía trên mục Tên tài liệu ở đầu của danh sách để nhận được dữ kiện đầu tiên, sau đó click vào biểu tượng "ống nhòm". Click vào biểu tượng kế tiếp từ phải sang (Biểu tượng "ống nhòm" và "mũi tên") để nhảy sang nhanh sang dữ kiện tiếp theo

  Lọc tài liệu: đôi khi để thuận tiện cho việc trình bày và chỉnh sửa các tài liệu cùng một ngôn ngữ, một tổ chức hay cùng một đề tài. Để làm điều này, chọn nút *lọc dữ liệu* (Hình "cái phiểu"), ở góc trên bên phải cửa sổ trình bày hộp hội thoại

Search documents, nhập vào tên tài liệu cần tìm và nhấn nút *Apply filter* để xác nhận yêu cầu.

Bạn cũng có thể thay đổi và kiểm tra việc tìm kiếm mà không cần phải rời khỏi hộp hội thoại với nút *Search* and *Reset search*. Bạn cũng có thể kích hoạt hoặc dừng chức năng *lọc* lại bất cứ khi nào với check box *Apply filter*. kỹ thuật lọc này tương tự như một Boolean "và" tìm kiếm:

- Danh mục tổ chức: chọn nút *Organisations* trình bày danh sách tất cả các tài liệu có chủ đề toàn cầu trong cơ sở dữ liệu (Các tài liệu này lấy từ bất kỳ các tài liệu nào trong số các tài liệu có liên kết, hoặc chưa được kết nối với bộ sưu tập). Nhấp đúp lên tên của bất kỳ tổ chức nào để mở hội hội thoại *Edit Organisation Name* để xem/thay đổi tên tổ chức và tên viết tắt nhanh (Giống như hộp hội thoại được trình bày với lệnh New/Organisation). Để tạo ra một tổ chức mới, sử dụng lệnh New/Organisation ở thanh menu trên cùng hoặc biểu tượng thứ ba từ trái sang trong thanh công cụ tắt (Xem phần trên).

- Danh mục chủ đề: các chủ đề là các thành phần của một phân lớp chuẩn để được tiếp cận với các tài liệu của bộ sưu tập. Có thể chọn thêm nút *Subjects*, xoá và chỉnh sửa các danh mục toàn cầu của tất cả các chủ đề trong cơ sở dữ liệu, ngay cả khi chúng chưa được chỉ định tài liệu nào. Một số không giới hạn của các thư mục đề tài được tạo ra để sử dụng tại phân lớp bộ sưu tập nhằm xây dựng và chỉ định cho tài liệu của bộ sưu tập theo một cấu trúc phân lớp đề tài. Nhấp đúp lên lên bất cứ chủ đề nào trong danh mục để mở cửa sổ *Edit subject* để xem/chỉnh sửa thuộc tính đề tài. Để tạo ra một chủ đề mới, sử dụng lệnh *Edit subject* trên thanh menu hoặc biểu tượng thứ ba từ trái sang trên thanh công cụ tắt (xem bên trên).

  Trong cửa sổ *Edit subject* hoặc *Add new subject*, từ khoá (keywords) có thể được thêm vào trong danh mục chủ đề toàn cầu (Hộp hội thoại *New keyword name*, cũng giống như việc sử dụng lệnh *New/Add-Modify keywords* từ thanh thực đơn ở trên cùng), và có thể được chỉ định với thuộc tính của

các chủ đề riêng biệt (hộp hội thoại chọn lựa chủ đề) sử dụng trong việc tìm kiếm các chủ đề để thêm vào bộ sưu tập. Từ khoá (Nếu hoặc không đánh dấu các loại đề tài) có thể được dùng để truy xuất tài liệu bằng thuộc tính của tài liệu đó (Dùng "How to" trong bộ sưu tập DLS). **Chú ý rằng có hai loại từ khoá được sắp xếp theo thứ tự alphabet trong danh mục, tạo thuận lợi cho việc sử dụng từ khoá để tìm tài liệu; cũng có thể bỏ qua bằng cách thêm vào mã "z-" trước từ khoá được chỉ định duy nhất ứng với các mục đề tài (Các mục này không được định rõ trong tài liệu nếu bạn không muốn các mã xuất hiện trong danh mục tìm kiếm từ khoá của chương trình ứng dụng phát sinh).**

Lưu ý: nếu cơ sở dữ liệu chính rất lớn, sẽ mất thời gian khá dài (một phút hoặc hơn) để upload các thành phần của bộ sưu tập. Sau khi đã chọn bộ sưu tập, hãy chờ cho đến khi tất cả các thành phần đã được upload lên hết trước khi bắt đầu công việc (Biểu tượng nhỏ có hình bóng đèn tròn xuất hiện trên dòng tab cho đến khi quá trình loading hoàn thành).

Thông báo Failure trong quá trình upload có thể làm cho chương trình bị bỏ qua. Nút *Export Files* mở ra cửa sổ *Export Settings* (Miêu tả bên dưới) có thể được lưu vào trong metadata của bộ sưu tập để truy xuất cấu trúc thư viện số vào thư viện Greenstone. Nó cũng có thể lưu hoàn toàn cơ sở dữ liệu.

ii. Cửa sổ thuộc tính bộ sưu tập

Cửa sổ này giúp cho ngừơi dùng có thể xây dựng hoặc thay đổi một bộ sưu tập riêng biệt, nó xuất hiện khi bộ sưu tập đã được chọn trong danh mục bộ sưu tập của cửa sổ Organizer Main. Nó cũng cho phép người dùng chọn môt trong trong 4 cách trình bày bộ sưu tập bằng cách click vào các tab trên cùng. Mỗi cách trình bày nó cung cấp một chữ số của các hàm chọn lựa và chỉnh sửa dữ liệu đã được miêu tả dưới đây:

a. Trình bày các đề tài: một số không giới hạn của các mục đề tài có thể tạo ra ở đây, phân cấp bộ sưu tập được chọn lên 6 cấp độ (Mặc dù các bộ sưu tập không cần quá 3

cấp). Để thêm một đề tài vào bộ sưu tập (có thể dùng nó như một thuộc tính của một hay nhiều tài liệu), i) chọn đề tài đầu tiên trong phần mà bạn muốn thêm mới thư mục vào, ii)chọn nút *Add subject*,iii) chọn *Add Subjects* từ danh mục toàn cầu , iv) sử dụng các nút để trình bày toàn bộ danh mục toàn cầu hoặc (Thủ tục thông thường hay mặc định) những nút chưa được sử dụng trong hệ thống phân cấp bộ sưu tập,v).

Chọn đề tài yêu cầu từ danh mục, và vi) chọn *OK* nếu đề tài mong muốn chưa có trong danh mục toàn cầu, sau đó sử dụng tùy chọn *Add New Subject* được cung cấp bởi nút *Add subject* (Giống như quay trở lại menu chính và thêm đề tài mới vào bằng lệnh *New/Subjects*). Các đề tài cũng có thể được thay đổi với biểu tượng *Edit* ở giữa các đề tài và tài liệu trình bày; Phần sử dụng chính của chức năng này là thêm vào tùy ý một biểu đồ số cho các cấp đề tài trong mục được sử dụng trong bộ sưu tập (Chú ý việc thay đổi các mục chỉ được dùng đối với bộ sưu tập hiện hành và không đem qua danh mục toàn cầu của các đề tài).

Để thêm vào một hay nhiều tài liệu vào trong một chủ đề (i.e. chỉ định đề tài cho một hoặc nhiều tài liệu), trước tiên bạn chọn mục đề tài hoặc mục con mà bạn muốn thêm vào các tài liệu, trong hộp danh mục cấp trên. Sau đó chọn một hay nhiều tài liệu trong hộp danh mục cấp thấp hơn, và click vào biểu tượng nhỏ giữa hai cách trình bày với một mũi tên hướng "lên' và cuốn sách màu đỏ (hoặc nhấp đúp lên từng tài liệu một).

Do vậy các tài liệu được chỉ định với đề tài được chọn sẽ được trình bày với dấu √ trước mỗi trường trong dòng tương ứng; để thuận lợi trong việc trình bày bạn có thể di chuyển lên xuống trang trình bày với các biểu tượng "√ up" và "√ down" ở giữa các phần trình bày. Bạn sẽ thấy rằng các tài liệu được thêm vào nơi tương ứng trong *cấu trúc phân lớp các đề tài*. Bây giờ lặp lại hành động này cho đến khi tất cả các tài liệu được phân lớp.

Một tài liệu có thể được chỉ định với nhiều đề tài theo ý muốn. Bạn có thể di chuyển (nhưng không phải là copy) một tài liệu được phân lớp dưới một đề tài này hay một đề tài khác bằng cách kéo và thả với nút trái chuột. Để di chuyển một tài liệu ra khỏi một đề tài, chọn danh mục tên tài liệu trong mục đề tài, nhấn phím delete và xác nhận yêu cầu.

Trình bày *các tổ chức*: Phần trình bày này được dùng như là phương tiện thuận lợi cho việc chọn hay không chọn các tài liệu đối với bộ sưu tập theo tên của tổ chức có liên quan mà cũng sẽ trở thành một phần của danh mục trong thư viện số cho việc phục hồi tài liệu thông qua tổ chức có liên quan (Lưu ý các tài liệu có thể được thêm và thay đổi thuộc tính tổ chức, từ phần trình bày các tài liệu).

Danh mục mặc định trong hộp phía tay trái chứa các tổ chức có liên hệ với với các tài liệu trong bộ sưu tập; còn hộp phía tay phải chứa các tên của các tài liệu mà tổ chức được liên hệ - Các tên này được đánh dấu với √ trong bộ sưu tập với cả những tên mà không được đánh dấu. click vào tên nào đó để chọn hoặc không chọn nó trong bộ sưu tập. Sử dụng biểu tượng ở góc dưới bên phải (ô trắng đánh dấu √ và ô tròn trắng dùng để chọn hoặc không chọn tất cả các tiêu đề).

Để chọn những tài liệu từ các tổ chức chưa đặt thuộc tính cho những tài liệu trong bộ sưu tập, chọn *Add Organisations* từ tùy chọn danh mục toàn cầu của nút *Add organisation* để thêm các tổ chức vào danh mục tổ chức bộ sưu tập, sau đó là quá trình chọn các tài liệu như trên. Tương tự, bạn có thể sử dụng nút *remove organisation* để di chuyển tất cả các tài liệu có liên quan với một tổ chức được chọn từ bộ sưu tập (Nhưng không phải từ danh mục tài liệu toàn cầu). Để làm việc chỉ với một tài liệu trong bộ sưu tập, chỉ việc bật tắt hộp checkbox để chọn danh mục các tài liệu.

Từ phần trình bày này, bạn có thể thêm các tổ chức mới, tài liệu mới vào trong danh mục toàn cầu (hộp hội thoại

Add new organisation của nút *Add Organisation* hoặc hộp hội thoại *Add new document* của nút *Add Document,* thực hiện theo thứ tự lần lượt như trên khi quay trở lại menu chính và thêm mới tài liệu với dòng lệnh *New/Organisations* hoặc *New/Subjects*.

c. Trình bày tài liệu: Danh mục tất cả các tài liệu được chọn gồm cả các tài liệu trong bộ sưu tập được kích hoạt. Điều này cũng giống như trong danh mục các tài liệu khi một tập tài liệu xuất hiện ở cửa sổ phía dưới trong phần trình bày *Các đề tài*. Điểm khác nhau chính là trong danh mục tài liệu này, khi nhấp đúp vào tài liệu cần trình bày sẽ mở ra hộp hội thoại Các thuộc tính của tài liệu ứng với tài liệu.
Sau đó bạn có thể thêm/thayđổi các thuộc tính của tài liệu đó (Hoặc khác nữa là các mục đề tài liên kết được thay đổi, được miêu tả ở trên trong mục *Subjects* view) bằng cách chọn một trong các tab đối với các thuộc tính của các phân lớp khác nhau ở trên cùng của hộp hội thoại:

- *Tab General*: trong cửa sổ này bạn có thể nhập tên tài liệu, số công việc, số trang và số hình ảnh, năm xuất bản, và tên cùng loại. Bạn cũng có thể thêm vào số lượng hình ảnh một các tự động bằng cách click vào nút *Find images* và chọn thư mục có chứa các hình ảnh của thư mục, sau đó chỉ ra kiểu định dạng phần mở rộng trong trường *Extensions.*

- *Tab Advanced:* nếu một tài liệu được xuất bản định kỳ, hoặc là một phần của bộ tài liệu, bạn có thể chỉ ra tiêu đề của tài liệu cần phát hành hay cả bộ tài liệu như là một thuộc tính, tự động tạo ra một thực thể trong danh mục các tài liệu phát hành định kỳ, nó được nhìn thấy trong cửa sổ tìm kiếm tiêu đề của chương trình ứng dụng cần sử dụng. Trong cửa sổ trình bày cấp cao mà bạn có thể chỉ định trong phần *Organisations* và *Languages* của tài liệu, Cả hai trường này đều có thể được lặp lại. Nếu có nhiều hơn một tổ chức có liên hệ với tài liệu (Nhà xuất bản, đồng tác giả) hoặc nếu được viết bằng nhiều ngôn ngữ, hoặc song ngữ Anh/Pháp, bạn nên chỉ ra tất cả các tài liệu tương ứng với từng danh mục riêng.

- *Tab Copyright*: Việc biết rõ trạng thái bản quyền tác giả khi xuất bản tài liệu rất quan trọng. Cửa sổ này gồm hai phần: Thứ nhất là nơi bản gốc được trình bày và mức độ bản quyền có thể được xác định. Thông tin này liên quan đến việc quản lý với Organizer, nó không ảnh hưởng gì đến trình ứng dụng thư viện số Greenstone.

- *Tab Suggested collections*: Phần này trình bày một danh mục các bộ sưu tập trong đó bao gồm tài liệu được đề nghị sau đó. Số lượng bộ sưu tập được đề nghị không giới hạn. Thông tin này dùng trong việc quản lý tài liệu với Organizer, và nó không ảnh hưởng gì đến trình ứng dụng thư viện số Greenstone.

- *Tab Keywords: Keyword* được dùng trong việc xuất bản tài liệu. thuộc tính này là biểu đồ phân cấp được thêm vào để bổ sung phân lớp đề tải và có thể được sử dụng để hoạt động trình ứng dụng Thư viện số để chọn và trình bày các tập tài liệu trong Thư viện. Trong DLS, nó được dùng như là một tham số "How to", nhưng nó cũng có thể được dùng cho bất cứ metadata nào khác được thêm vào, với ví dụ là tác giả hay đất nước của tài liệu nguồn.

Chú ý; Trong phần trình bày các *Đề tài, Tổ chức* và *Tài liệu*, nút *Add documents* cho phép người dùng thêm mới tài liệu trực tiếp vào bộ sưu tập từ hộp hội thoại hoặc chọn từ danh mục các dữ liệu toàn cầu. Tài liệu được thêm vào trong cửa sổ thuộc tính trình bày một cách tự động danh mục các dữ liệu toàn cầu được nhập vào cho việc sử dụng trong tương lai.

Khi một tài liệu mới được thêm vào bộ sưu tập từ danh mục các dữ liệu toàn cầu, hộp hội thoại *Search documents* được xuất hiện để người dùng có thể dễ dàng xác định nhu cầu tài liệu theo nhiều nhiều chuẩn chọn lựa (Điều này giống như chức năng lọc dữ liệu được mô tả như trên trong danh mục các tài liệu thảo luận được sử dụng bằng cách chọn nút *Documents* của thanh công cụ đứng trong cửa sổ Organizer Main)

d. Các phần trình bày khác: Phần trình bày các phân cấp khác cho thấy thứ tự các tài liệu theo các tiêu đề kế tiếp nhau (Không được phép chỉnh sửa)
và phân cấp các tiêu đề theo mẫu tự alphabet theo mỗi ngôn ngữ. Người dung có thể thay đổi nhóm phân cấp theo mẫu tự alphabet (Ví dụ: A-C, E-G or A-L, M-Z v.v…) theo cở tốt nhất để trình bày những tài liệu trong Thư viện hoàn chỉnh. Để làm điều này, click vào một ngôn ngữ và sử dụng chứa nút chia ký tự (*Split letters)*. Khi đã cảm thấy hài long với kết quả, click vào nút *Save Splitters* (Cho đến khi bạn có thể quay trở lại vị trí ban đầu hay các ký tự đã được lưu trước đó bằng cách click vào Load/Refresh để phân chia hoặc loại bớt ký tự bằng việc click vào *Eliminate Splittings*).

iii. The *Export Settings* window

Cửa sổ này được trình bày khi biểu tượng ở dưới thanh công cụ đứng của cửa sổ *Organizer Main* được chọn, cho phép bạn lấy ra kết quả công việc và thông thường là giai đoạn cuối liên quan đến việc tạo ra bộ sưu tập mới hoặc là bộ sưu tập con. Chọn *Export Files* để liên kết với cửa sổ *Export Settings* và chọn một bộ sưu tập để truy xuất ra ngoài và một thư mục để nhận thông tin được truy xuất. sau đó click *Export file*s.

Việc này sẽ làm thay đổi 5 file *collect.cfg*, *metadata.xml*, *sub.txt,org.txt, Keywords.txt* và *AZList.txt* trong thư mục được chọn. Để xây dựng bộ sưu tập với thông tin này, bạn cần di chuyển các file đến nơi liên kết. Nơi có file *metadata.xml* ở trong thư mục *import* của bộ sưu tập và các thư mục khác của bộ sưu tập .v.v..

Bắt đầu 10 bước trong 15 phút

a. Cài đặt thư mục Greenstone (xem tài liệu the *Greenstone Installer's Guide*) bao gồm luôn cả Thư viện Demo dạng DLS và các file nguồn. **Lưu ý nếu bạn muốn có thể thêm vào bộ sưu tập của mình 140 tài liệu bất kỳ trong bộ sưu tập DLS ở cơ sở dữ liệu Organizer để ở chế độ mặc định (Thay vì chỉ là 14 tài liệu như trong bộ sưu tập của chương trình Demo trong Thư viện Greenstone), bạn nên cài DLS như là một mẫu Thư viện Greenstone và thay thế "Demo củ" bằng "dls" theo cấu trúc dưới đây**. Bộ sưu tập Demo và DLS sẽ

được cài đặt theo thứ tự sau trong *c:\program files\gsdl\collect\demo and c:\program files\gsdl\collect\dls*.

Nếu bạn cài đặt Greenstone trước mà không có DLS và muốn cài thêm DLS, thì bạn có thể hủy việc cài đặt hay cài lại Greenstone chỉ với bộ sưu tập này.

b. Thiết lập cấu trúc cho bộ sưu tập mới (Chúng ta sẽ thích để nó dưới dạng là "newcol") bằng cách điều khiển dòng lệnh sau: *run* trong menu *Start* trong windows:
"c:\program files\gsdl\bin\windows\build" newcol

c. Thay thế file *collect.cfg* mặc định được tạo ra từ bước trước được sử dụng bằng chương trình Demo. Lưu lại đường dẫn *c:\program files\gsdl\collect\demo\etc\collect.cfg* thành *c:\program files\gsdl\collect\newcol\etc\collect.cfg*. Điều này cần thiết là vì Demo sử dụng (và tất cả bộ sưu tập dạng DLS) sử dụng một số tùy chọn đặt biệt mà bộ sưu tập mặc định không có (xem TL *GreenstoneDeveloper's Guide* để biết thêm chi tiết)

Bạn in những chỉ dẫn dưới đây và làm theo từng bước dưới đây:

1. Mở Collection Organizer, chọn cơ sở dữ liệu dls và nhập từ "admin" cho cả user name và password (Nút Collections của thanh công cụ đứng sẽ được tô sang mặc định; nêu không sang thì click vào nút đó)

2. Chọn lệnh *New/Collection/Empty* trong thanh thực đơn nằm ngang ở trên cùng của cửa sổ Organizer Main để tạo ra một bộ sưu tập mới trống. Đặt tên bộ sưu tập và phiên bản mà bạn chọn, ví dụ như đặt tên là "My First Collection" và phiên bản "1.0"

3. Với một số thuộc tính của tài liệu, bạn sẽ phải tạo ra một danh mục các giá trị có thể trước tiên. Vì vậy nếu bạn biết nhiều ngôn ngữ và/hoặc các tổ chức xuất bản trong các tài liệu của bạn, dùng lệnh *New/Add-Modify languages* và *New/Organisation* để them vào tất cả các ngôn ngữ mà bạn sẽ sử dụng trong tài liệu này hoặc cho bộ sưu tập trong tương lai cũng như vai trò các nhà xuất bản tài liệu của bạn.

Bạn cũng có thể dùng một dòng lệnh để thêm/Thay đổi ngôn ngữ và các tổ chức khi nào muốn, nhưng không phải là chỉnh sửa bản thân bộ sưu tập trong chỉ dẫn sau đây.

4. Nhấp đúp lên dòng tên bộ sưu tập mà bạn tạo

5. Click vào tab *Subjects* ở trên đỉnh (Nếu chưa chọn có thể để ở dạng mặc định); sau đó click vào nút *Add subject* lệnh *Add new subject*, sau đó nhập tên đề tài mới vào trường *Subject title*, nhấn phím "enter" sau mỗi lần thực hiện. Click vào dấu + trước từ *Subjects* trong danh mục liệt kê phân cấp để xem đề tài mà bạn yêu cầu.

6. Click lên tab *Documents* để mở trang trình bày tài liệu, sau đó thêm các tài liệu vào bộ sưu tập như sau:

a. Để thêm một tài liệu vào bộ sưu tập Demo (Hoặc bộ sưu tập DLS nếu đã được cài đặt trên Greenstone) vào bộ sưu tập mới của bạn, click nút *Add documents* và chọn *Add document* từ danh mục toàn cầu. Định vị trí tài liệu bạn yêu cầu (Sử dụng chức năng lọc dữ liệu được miêu tả ở trên) và thêm nó vào bộ sưu tập của bạn trong Organizer. Sau khi thêm tài liệu, định vị file nguồn của bộ sưu tập trong Thư mục Demo import (*c:\program files\gsdl\collect\demo\import*) và copy chúng vào thư mục import của bộ sưu tập mới của bạn. Ví dụ, để thêm tài liệu "Butterfly Farming in Papua New Guinea" vào số công việc khi bạn xác định nó trong Organizer. Số công việc của tài liệu này là "b22bue", vì vậy bạn nên copy thư mục "b22bue" từ *c:\program files\gsdl\collect\dls\import\ac01ne* sang *c:\program files\gsdl\collect\newcol\import\ac01ne*.

   Để thêm một tài liệu mới (Nghĩa là có một tài liệu không có trong bộ sưu tập Demo) vào bộ sưu tập mới của bạn, click nút *Add documents* và chọn *Add new document.* Nhập tên, số công việc của tài liệu (Lựa chọn của bạn), số trang, tổ chức xuất bản, ngôn ngữ và thông tin khác. Bạn phải tạo ra một thư mục mới trong *c:\program files\gsdl\collect\newcol\import* để liên hệ với số công việc

của tài liệu mới. Trong thư mục mới này bạn nên để file nguồn của tài liệu và bất cứ file hình ảnh nào có liên quan (Trong HTML hay bất cứ định dạng nào khác được chấp nhận bởi Greenstone (xem trong tài liệu *Greenstone User's Manual*).

7. Quay trở lại tab đề tài bạn sẽ nhìn thấy tài liệu của bạn được trình bày trong hộp danh mục liệt kê phía dưới. Chọn một tài liệu, sau đó chọn môt chủ đề ở danh mục nhánh trên mà bạn muốn phân lớp tài liệu này trong đó và click vào biểu tượng nhỏ giữa hai phần trình bày với một mũi tên chỉ lên ("up")và cuốn sách màu đỏ. Khi tài liệu đã được phân lớp, bạn vẫn có thể di chuyển nó từ đề tài này sang đề tài khác bằng cách kéo – thả với nút trái chuột. Bạn cũng có thể di chuyển các tài liệu hoặc các đề tài lên xuống giữa các cấp độ tương tự nhau của biểu đồ phân lớp bằng cách chọn nút lên, xuống màu xanh vào phía bên phải của danh mục phân cấp đề tài. Cố gắng phân lớp trung bình từ 6 – 30 tài liệu trong một đề tài. Một tài liệu có thể được chỉ định trong nhiều đề tài mà bạn muốn.

8. lặp lại các bước trên bằng cách thêm đề tài mới, và thêm nhiều tài liệu hơn. Khi Thư viện được hoàn thành, bạn sẽ phải xem lại danh mục các đề tài và các tài liệu, để chắc rằng tất cả đều được nhập vào và phân lớp, sắp thứ tự chính xác.

9. Cuối cùng, đóng cửa sổ thuộc tính bộ sưu tập và nhấn nút *Export Files* của thanh công cụ đứng. Phần này sẽ mở ra một cửa sổ *Export Settings*. Click vào nút *Display collection list* và chọn bộ sưu tập của bạn, sau đó click vào nút *Browse for folder* và chọn thư mục mà bạn muốn truy xuất file metadata, nhấn nút *Export files* để truy xuất metadata của bộ sưu tập cho quá trình xây dựng với Greenstone.

10. Copy file được truy xuất đến những nơi có liên quan trong cấu trúc thư mục của mới bộ sưu tập của bạn.

a. File *metadata.xml* được truy xuất, nên được copy vào thư mục *c:\program files\gsdl\collect\newcol\impor*.

b. Các file *AZList.txt*, *Keyword.txt*, *sub.txt*, and *org.txt* được truy xuất, nên được copy vào thư mục *c:\program files\gsdl\collect\newcol\etc*.

Lưu ý file collect.cfg được sinh ra bởi Organizer không được yêu cầu bởi các dòng phân lớp đã chứa file collect.cfg rồi cho bộ sưu tập Demo và DLS. Bộ sưu tập The newcol đã sẳn sàng được xây dựng. Xây dựng nó từ dòng lệnh import.pl và buildcol.pl (xem chi tiết trong tài liệu the G*reenstone Developer's Guide*).

5.3 Đính kèm các file tài liệu

Tài liệu nguồn thường cần để xây dựng những phần lớn và những phần nhỏ của bộ sưu tập, và thông tin này cần để liên kết với Greenstone để nó có thể bảo toàn cấu trúc phân cấp. Cũng như thế metadata – đề tài điển hình – có thể được liên kết với mỗi phần lớn và phần nhỏ. Các tài liệu nguồn từ quá trình OCR là điển hình cho một tập hợp các từ xử lý file, bao gồm các file hình ảnh. Nếu các file này thuộc dạng file MicrosoftWord, họ có thể input vào Greenstone bằng cách sử dụng plugin dạng Word. Có thể vừa chuyển thành file HTML vừa dùng plugin HTML để input. Trong trường hợp khác cấu trúc phân cấp của một tài liệu có thể được chỉ định bằng cách thêm đuôi dạng text như sau:

```
<!--
<Section>
<Description>
<Metadata name="Title">Realizing human rights for
poor
people: Strategies for achieving the international
development targets</Metadata>
</Description>
-->
(text of section goes here)
<!--
</Section>
-->
```

Cách ghi như trên được dùng bởi vì chúng chỉ ra các dòng lệnh ở dạng HTML; và vì thế các đuôi được thêm vào trong phần này sẽ không ảnh hưởng đến định dạng tài liệu. Bạn phải ghi những dòng trên vào phần đuôi của các phần, ngay cả khi tài liệu mà bạn đang sử dụng không phải là file HTML (e.g. nếu nó là file dạng Microsoft Word). Trong phần miêu tả chi tiết (between the <Description> and </Description> tags) các loại metadata khác có thể được chỉ định, nhưng lại không làm đối với các tài liệu mà chúng ta đang miêu tả ở đây.

Điều quan trọng là phải nhớ rằng bạn đang tạo bảng mục lục phân cấp khi chèn vào phần đuôi trong tài liệu của bạn. Điều này có nghĩa các phần này có thể được để lòng vào các phần khác. Trên thực tế, tất cả

các phần đều phải được lòng vào các phần đơn khép kín bao quanh toàn bộ tài liệu.

Ví dụ sau chứng minh được rằng một tài liệu với hai chương, chương hai chứa hai phần nhỏ. Ví dụ thực tế của các tài liệu gốc được thêm vào phần đuôi bằng cách này, hãy nhìn các tài liệu nguồn trong bộ sưu tập Demo và DLS.

```
<!--
<Section>
<Description>
<Metadata name="Title">My Document</Metadata>
</Description>
<Section>
<Description>
<Metadata name="Title">Chapter 1</Metadata>
</Description>
-->
(text of chapter 1 goes here)
<!--
</Section>
<Section>
<Description>
<Metadata name="Title">Chapter 2</Metadata>
</Description>
<Section>
<Description>
<Metadata name="Title">Subsection 1</Metadata>
</Description>
-->
(text of sub-section 1 goes here)
<!--
</Section>
<Section>
<Description>
<Metadata name="Title">Subsection 2</Metadata>
</Description>
-->
(text of sub-section 2 goes here)
<!--
</Section>
</Section>
</Section>
-->
```

Lưu ý metadata được chỉ định từ phần đuôi trong tài liệu nguồn theo thứ tự ưu tiên đã được chỉ định từ file metadata.xml (Giống như được tạo ra bởi Organizer). Điều này có nghĩa là bạn không nên chỉ định rõ metadata chủ đề cho cấp cao nhất của tài liệu nguồn trừ khi bạn muốn

bỏ qua tiêu đề mà bạn đã cài vào từ Organizer. Trong ví dụ dưới đây, nếu bạn muốn lấy tên của tài liệu mà bạn cài trong Organizer bạn nên bỏ qua dòng sau:

```
<Metadata name="Title">My Document</Metadata>.
```