QUY TRÌNH CÔNG VIỆC CHO MỘT DỰ ÁN SỐ HÓA

TOM DE MULDER

Unix System Programmer/Administrator
Dspace@Cambridge Project Team –
Cambridge University Library –
tdm27@cam.ac.uk;
www.lib.cam.ac.uk

TÓM TẮT

Ngày càng nhiều các cơ quan mong muốn chuyển đổi nội dung truyền thống của mình sang định dạng số. Trong các dự án như vậy, giai đoạn số hóa và tạo lập siêu dữ liệu thường diễn ra không đồng thời. Bài báo này nhận dạng tầm quan trọng của sự kiểm tra chéo thường xuyên cả hai giai đoạn này. Chúng tôi đề nghị một quy trình số hóa theo một quy trình thống nhất, và một cách thực hành kỹ thuật để tự động hóa nó.

1. DẪN NHẬP

Trong ngành công nghiệp giải trí, người ta đều hiểu rõ tầm quan trọng của việc đồng bộ hóa phần tiếng và hình ảnh (audio and video) của một bộ phim. Điều quan trọng rằng cả âm thanh và hình ảnh (cả phần phụ đề nếu có) cần chạy đồng bộ cùng với nhau. Nếu sự đồng bộ này không có thì kết quả là một sự trộn lẫn tín hiệu này sẽ trái ngược nhau. Tương tự như vậy, chúng ta cần giữ cho các phần siêu dữ liệu và dữ liệu trong quá trình số hóa nội dung để chúng được đồng bộ hóa, khi đó sản phẩm cuối cùng của chúng ta, một bộ sưu tập dữ liệu và siêu dữ liệu, sẽ trở lên có ý nghĩa.

Trước kia, sự thiếu đồng bộ hóa đã gây ra nhiều vấn đề cho các dự án số hóa tại thư viện Đại học Cambridge (Cambridge University Library). Trong những trường hợp như thế này, quy trình số hóa hoàn toàn tách rời với các chuyên gia tạo ra siêu dữ liệu cho những tiêu đề tài liệu được số hóa. Chỉ đến khi cả hai phần dữ liệu và siêu dữ liệu cuối cùng được kết hợp lại với nhau, và khi đó các chuyên gia thường thấy sự không thống nhất giữa hai phần dữ liêu này.

Sự không đồng bộ này đã cho thấy rằng chúng ta phải tốn rất nhiều thời gian và gây nên sự phức tạp để giải quyết vấn đề: chúng ta cần sự can thiệp của nhân viên thư viện vào rà soát toàn bộ sưu tập để phát hiện và sửa lỗi cũng như các thiếu sót đã xảy ra. Chúng ta phải tốn nhiều thời gian để triển khai nhiều công việc hơn đối với bộ phận số hóa, và kết hợp lại những kết quả cuối cùng.

Trong bài báo này, chúng tôi cố gắng nêu lên những vấn đề và định hình một quy trình nhằm phát hiện lỗi trước khi tác động đến các công đoạn khác của

quy trình hình ảnh hóa nội dung. Trong khi bài này tập trung vào việc hình ảnh hóa các bản thảo, thì chúng ta cũng có thể dễ dàng nhận thấy nó liên quan đến bất kỳ dự án nào mà sự tạo ra siêu dữ liệu và dữ liệu diễn ra tách rời nhau, như hình ảnh số hóa, đối tượng số theo kích cỡ 3 chiều, phần âm thanh hoặc hình ảnh analog (hình ảnh truyền theo công nghệ tín hiệu tương tự) số hóa.

2. ĐỒNG BỘ HÓA: NHÚNG KHÓA CHUNG.

Nói một cách rộng ra, siêu dữ liệu phục vụ hai mục đích: nhận dạng và mô tả dữ liệu. Nó sẽ được dùng để di chuyển tới hoặc xác định vị trí dữ liệu (trong trường hợp của chúng tôi, đó là các hình ảnh bản thảo) khi duyệt hoặc tìm kiếm trên một kho dữ liệu cũng như thu thập thông tin nhiều hơn về chính dữ liệu đã được tìm thấy.

Trong một quy trình số hóa, sự nhận dạng là cách sử dụng mà chúng ta quan tâm nhất - chỉ sau khi dữ liệu đã được nhận ra rõ ràng thì nhiều công việc hữu ích mới được tiến hành đối với nó, ví dụ như nhập thêm dữ liệu mô tả.

Bởi vậy, chúng ta cần phải tìm ra một cách để nhận dạng duy nhất một đối tượng được số hóa. Ví dụ chúng ta có thể sử dụng cách đánh dấu phân lớp ("classmark") dùng trong thư viện . Cách dễ dàng nhất để mã hóa dấu phân lớp này trong hình ảnh với công nghệ hiện hành là làm nó như một phần của tập tin.

Một cách tiếp cận tương tự tới sự nhận dạng ảnh là đảm bảo rằng dấu phân lớp luôn hiển thị chính bên trong hình ảnh đó. Dấu nhận dạng này có thể được in trên một mẩu giấy, hoặc viết trên một cái bảng nhỏ và đưa vào trường dữ liệu của máy chụp khi bắt lấy hình ảnh.

Ngoài ra, có nhiều cách cho phép nhúng siêu dữ liệu trực tiếp vào tập tin dữ liệu (xem Phụ lục A). Cách này sẽ gắn kết hiệu quả dữ liệu và siêu dữ liệu cùng lại với nhau, đồng thời giảm việc chia tách chúng. Thao tác này diễn ra càng sớm trong một quy trình xử lý, thì quy trình đó sẽ càng có tính đồng bộ hóa cao. Nó cũng khiến cho chúng ta dễ dàng hơn để giải quyết những khác biệt sau này.

3. QUY TRÌNH CÔNG VIỆC

Định nghĩa

Vì mục đích của quy trình này, chúng ta hãy định nghĩa "chuyên gia" ("expert") như là môt người kiểm soát siêu dữ liệu đối với tài liệu được số hóa; "Nhiếp ảnh gia" ("Photographer") là người (hoặc nhóm người) chịu trách nhiệm tạo ra tập tin ảnh của tài liệu. "Đánh dấu phân lớp" ("classmark") là đánh dấu duy nhất của một tài liệu.

4. THỰC HÀNH KỸ THUẬT

4.1 Cơ sở thực hiện

Đối với ví dụ này, chúng ta giả sử rằng một cấp độ ảnh hưởng kỹ thuật có thể xảy ra với tất cả các bước của quy trình số hóa và xử lý siêu dữ liệu. Thiếu nó, sự đồng bộ hóa quy trình sẽ trở lên khó khăn.

4.1.1 Dịch vụ tập trung

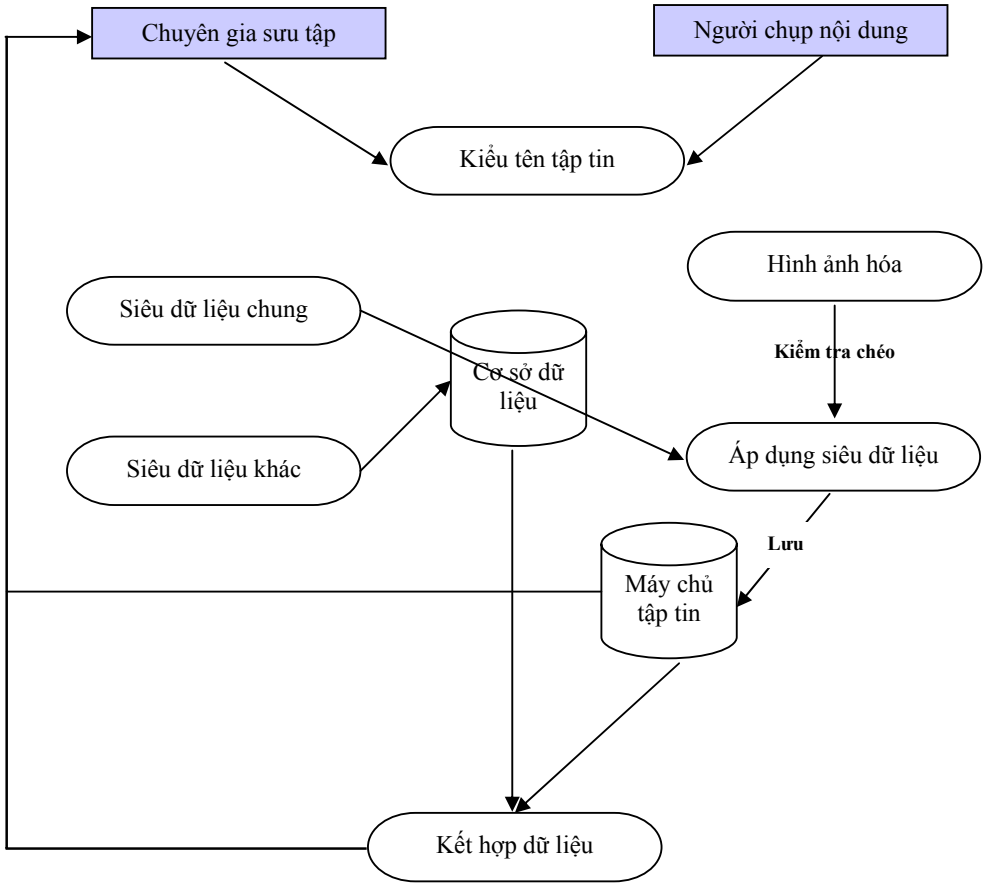
Tâm điểm của hệ thống là một bộ dịch vụ nối kết mạng tập trung. Một trong những dịch vụ này là một dịch vụ

cơ sở dữ liệu quan hệ nhằm xử lý siêu dữ liệu của dự án. Lược đồ của nó (thực tế là một bộ duyệt xem[1]) được biên tập phù hợp với nhu cầu của mỗi dự án cụ thể. Các trường siêu dữ liệu có thể được hiển thị thông qua một giao diện Web hoặc thông qua một máy khách dùng giao thức ODBC[2] (Giao thức Nối kết Cơ sở dữ liệu mở).

Hơn nữa việc cung cấp một dịch vụ trung tâm là một hệ thống tập tin nối kết mạng, có khả năng truy cập cả bằng máy chủ trung tâm và người hình ảnh hóa nội dung. Dịch vụ này giúp các hình ảnh khi được chụp sẽ được lưu trữ và tại đó chúng sẽ trải qua một loạt thao tác theo quy trình được tự động hóa. Dịch vụ trung tâm được mô tả chung nhất là sự cung cấp các "thủ tục từ xa". Nó được sử dụng bởi nhiều cấu thành khác của một hệ thống để truy xuất hoặc lưu trữ thông tin liên kết với nhiều bước khác nhau của quy trình này.

4.1.2 Dịch vụ máy khách

Chúng ta giả sử rằng người chụp hình ảnh nội dung sẽ sử dụng một máy Mac Apple cài đặt phiên bản Mac OS X. Phiên bản này cho phép sử dụng những "thao tác thư mục" ("Folder actions"). Những thao tác kiểm soát (điển hình là các chương trình nhỏ) được thực hiện bất cứ khi nào một tập tin được lưu/mở/sửa đổi.



4.2 Các bước đầu tiên.

Trước khi siêu dữ liệu hoặc quy trình số hóa bắt đầu, chúng ta cần thống nhất một vài tiêu chuẩn để tuân thủ. Những tiêu chuẩn này sẽ là những yếu tố chính để đồng bộ hóa quy trình.

- Định danh từ vựng chính xác của bộ phân lớp. Ví dụ: nn.xxx-yyy:bbb, [r/v], có trường hợp dãy ký tự cho nn, xxx, yyy và bbb cần được định nghĩa và r/v được thống nhất như là trang phải/trang trái (recto/verso).
- Lược đồ siêu dữ liệu được sử dụng. Trong hầu hết các trường hợp, sử dụng Dublin Core[3] sẽ thích hợp, có thể với những mở rộng tuỳ chọn.
- Siêu dữ liệu chung: một bộ thẻ siêu dữ liệu sẽ áp dụng tới toàn bộ bộ sưu tập, ví dụ như "tên bộ sưu tập",…

4.3 Quy trình

4.3.1 Siêu dữ liệu

Khi một chuyên gia nhập siêu dữ liệu vào cơ sở dữ liệu trung tâm, giả sử không có một trật tự cụ thể về nhập liệu thì thao tác nhập này có thể diễn ra theo khối dữ liệu nếu máy khách hỗ trợ chức năng này, sau đó nó được chuyển tới máy chủ. Cần thống nhất khi khởi đầu mỗi dự án số hóa đó là trường dữ liệu nhận dạng biểu ghi nên được kiểm tra nghiêm ngặt.

Bất kỳ khi nào một máy chủ thấy một biểu ghi siêu dữ liệu được điền vào hoàn chỉnh, nó có thể kiểm tra hệ thống tập tin nối kết mạng xem các tập tin có phù hợp không. Nếu các tập tin đã sẵn có thì siêu dữ liệu có thể được điền thêm.

Điều quan trọng rằng nếu một biểu ghi siêu dữ liệu được đánh dấu trước đó là đã "hoàn chỉnh" bị thay đổi thì siêu dữ liệu được nhúng trong một ảnh tương ứng cần được thay đổi ngay lập tức.

4.3.2. Hình ảnh hóa

Nếu nhiếp ảnh gia sử dụng Adobe Photoshop CS thì sau đó khuôn mẫu siêu dữ liệu cần được xác định để nắm giữ siêu dữ liệu chung của bộ sưu tập. Khuôn mẫu này sau đó được sử dụng trước khi hình ảnh được lưu, đồng thời đảm bảo hình ảnh đó chứa đựng siêu dữ liệu của nó sớm nhất. Cách này làm giảm đi số lượng hình ảnh không gắn kết siêu dữ liệu ("orphaned").

Khi hình ảnh được lưu trên hệ thống tập tin nối kết mạng thì chúng ta có thể dùng thao tác thư mục để kiểm tra tên tập tin để đảm bảo nó tuân thủ tiêu chuẩn được định nghĩa trong bước đầu tiên của dự án (ví dụ, nn.xxx-yyy:bbb, [r/v]). Bất kỳ nhầm lẫn nào ở giai đoạn này sẽ được phát hiện ngay lập tức và quy trình này tạm dừng lại đến khi vấn đề được giải quyết.

4.3.3 Tự động hóa trên máy chủ

Thường sẽ là một cách thực hành tồi khi chỉ dựa vào những thao tác trên thư mục để đồng bộ hóa hai quy trình công việc mà không có một cấp độ kiểm tra phụ thêm.

Máy chủ thường tiến hành kiểm tra định kỳ trên toàn bộ cơ sở dữ liệu và hệ thống tập tin nối kết mạng, đồng thời kiểm tra tên tập tin, thử nghiệm những hình ảnh đối với dữ liệu đã nhúng và xác nhận hợp lệ hoặc điền thêm siêu dữ liệu phù hợp.

4.3.4 Thông báo

Cả bên chuyên gia (người kiểm soát siêu dữ liệu) và bên làm thao tác số hóa có thể được tự động thông báo về tiến trình của bên kia. Vào cuối ngày, nhiếp ảnh gia có thể được gửi một thư điện tử tóm lược

về những biểu ghi siêu dữ liệu nào đã hoàn thành. Bên chuyên gia có thể nhận một danh mục hình ảnh, đồng thời chỉ ra biểu ghi siêu dữ liệu thích hợp nào sẵn có hoặc bị thiếu.

Một giao diện web đơn giản có thể cho thấy toàn bộ tình trạng dự án tại bất kỳ thời điểm nào, đồng thời cho thấy những khác biệt giữa hai quy trình công việc này.

4.4 Giai đoạn cuối: hình ảnh được làm giàu thông tin bằng siêu dữ liệu toàn diện

Một khi cả quy trình tạo ra siêu dữ liệu và hình ảnh hóa hoàn thành, thì một sát nhập cả hai phần này có thể diễn ra để tạo ra một dữ liệu toàn diện với siêu dữ liệu được nhúng. **Dù sao, một khối siêu dữ liệu trực tiếp tách riêng dưới định dạng XML (đối với hầu hết các ứng dụng thì điều này dễ dàng sử dụng hơn siêu dữ liệu nhúng)** là cách ưa thích hơn. Những bước thực hành cuối cùng này dường như cho thấy các bước thực hành trước đó là thừa, song những bước thực hành trước đó có tính quyết định để có được sự gắn kết của cả dự án trong trường hợp vì một lý do không thể dự báo nào đó, dự án bị huỷ bỏ hoặc trì hoãn trong một thời gian dài. Trong trường hợp như vậy, thường không có khối dữ liệu đầu ra ("output dump") cuối cùng, song ít ra dữ liệu được tạo ra thường vẫn có thể nhận diện được.

5. Kết luận

Dữ liệu nhúng dường như là một giải pháp đối với nỗi quan ngại khi số hóa nội dung. Tuy nhiên, bởi vì nhiều phần mềm (và nhiều định dạng số) đã không được thiết kế ngay trong ý tưởng với siêu dữ liệu nhúng cho nên có nhiều khó khăn mà người ta phải thận trọng để tránh vấp phải.

Một rủi ro đó là, do khó khăn của việc phân tích và chỉ mục siêu dữ liệu nhúng, cho nên chúng ta có thể sử dụng một giải pháp khác để lưu trữ siêu dữ liệu thực tế (ví dụ, cơ sở dữ liệu quan hệ), và trừ phi chúng ta tiến hành kiểm tra kỹ lưỡng nếu không rủi ro này sẽ rất cao dẫn đến hai bộ siêu dữ liệu sẽ bắt đầu khác biệt nhau.

Bất kỳ một tiện ích hoặc kho dữ liệu nào dùng để đọc, hoặc thao tác với siêu dữ liệu nhúng cần được biết về những trở ngại này. Điều quan trọng là cần phải định nghĩa nguồn chính xác cho siêu dữ liệu, và kiểm tra định kỳ bất kỳ bộ siêu dữ liệu nào khác đã lưu trữ để so sánh với nó.

Trong hầu hết các trường hợp, dữ liệu nhúng sẽ là sự thay đổi cuối cùng đối với dữ liệu trước khi nó được lưu trữ trong một kho dữ liệu, và dữ liệu và siêu dữ liệu đó sẽ không bao giờ thay đổi lại. Trong trường hợp này nó trở thành một cấu thành có giá trị của bảo quản số vì nó đảm bảo rằng trong tương lai dữ liệu và siêu sữ liệu sẽ không bị chia tách ra. Dù sao, như đề cập trong bài báo này, siêu dữ liệu nhúng có thể là một công cụ hữu ích cho quản lý quy trình công việc, và tăng sự tin cậy cũng như giá trị của tài liệu số.

PHỤ LỤC A: SIÊU DỮ LIỆU NHÚNG: MỘT TÓM TẮT VỀ KỸ THUẬT

LỊCH SỬ

Hầu hết định dạng đồ họa có một lịch sử cho phép siêu dữ liệu được nhúng vào. Ví dụ, định dạng nén tập tin ảnh TIFF và JPEG cho phép nhúng các bộ trường siêu dữ liệu EXIF và IPTC. Tuy nhiên, những trường này thường có phạm vi hẹp và có xu hướng nhằm vào mô tả các mặt kỹ thuật của một quy trình sao chụp hình ảnh hơn thay vì siêu dữ liệu mô tả mà chúng ta quan tâm.

Tuy nhiên, với sự ra đời của công nghệ web ngữ nghĩa (được biết đến nhiều nhất đó là RDF[4]), nó đã có thể mô tả nội dung mà không phải bó hẹp vào các bộ trường dữ liệu cố định. Thay vào đó, một lược đồ có thể linh hoạt và tuy biến hơn đối với mỗi nội dung trong khi vẫn có khả năng đọc máy.

Để sử dụng công nghệ này, hãng Adobe đã phát triển tiêu chuẩn XMP[5]. Nó cho phép siêu dữ liệu ở định dạng RDF/XML[6] được nhúng vào nhiều loại định dạng tập tin. Tới năm 2005, tất cả các phiên bản hiện hành của sản phẩm Adobe đều hỗ trợ tiêu chuẩn này, nhiều công cụ của bên thứ ba cũng hỗ trợ nó.

Một trong những lược đồ mô tả siêu dữ liệu mặc định được hỗ trợ bởi tiêu chuẩn XMP đó là bộ phần tử Dublin Core Giản lược (Simple Dublin Core - SDC). Đối với một số mục đích của chúng ta thì điều này dường như làm mọi người thất vọng, vì bộ phần tử Dublin Core chuẩn hóa (Qualified Dublin Core - QDC) phù hợp hơn nhiều đối với một siêu dữ liệu toàn diện (dù sao thảo luận về SDC so với QDC nằm ngoài phạm vi của bài báo này). Tuy nhiên, có vài cách sẵn có để mã hóa bộ phần tử Qualified Dublin Core dưới dạng RDF/XML để trình phân tích cú pháp mong muốn Simple Dublin Core có thể vẫn đọc được các trường SDC.

Dĩ nhiên, nếu sử dụng cách tiếp cận này nên chú trọng khi biên tập tập tin - nếu một tập tin chứa đựng QDC XMP được biên tập và lưu trữ sử dụng một công cụ dùng SDC thì những trường siêu dữ liệu phụ thêm sẽ rất có khả năng bị mất.

Tháng 5/2005

---

[1] Một trình duyệt xem cơ sở dữ liệu là một phân lớp trừu tượng trên một lược đồ cơ sở dữ liệu thực tế, đồng thời khiến nó có thể đại diện cho thông tin theo một cách có ý nghĩa đối với người dùng trong khi che dấu đi những phức tạp về mặt kỹ thuật khi thực hành. Xem thêm thông tin tại http://philip.greenspun.com/sql/view.html

[2] Một giao thức truy cập cơ sở dữ liệu từ xa, thường thông qua các máy khách như OpenOffice BASE hoặc Microsoft Access.

[3] http://dublincore.org/

[4] The Resource Description Framework (Khung mô tả tài nguyên). Như tên của tiêu chuẩn này ám chỉ thì nó là một định khuôn để mô tả và trao đổi siêu dữ liệu

[5] http://www.adobe.com/products/xmp/main.html

[6] Một cách chung để mã hóa RDF trong định dạng XML.