XÂY DỰNG HỆ THỐNG QUẢN LÝ MẪU VĂN BẢN

DEVELOPMENT OF A DOCUMENT MANAGEMENT SYSTEM

Nguyễn Đình Lầu, Phan Huy Khánh

Trường Đại học Bách khoa, Đại học Đà Nẵng

TÓM TẮT

Trong hoạt động xử lý văn bản nói chung, người ta thường phải tiến hành thẩm định một văn bản đã cho có đúng với yêu cầu sử dụng hay không. Việc thẩm định đòi hỏi phải kiểm tra nội dung và hình thức trình bày văn bản dẫn đến mất rất nhiều công sức, thời gian của người sử dụng (NSD), thậm chí xảy ra nhầm lẫn, sai sót. Trong bài báo này chúng tôi giới thiệu hệ thống xác thực văn bản cho phép thẩm định một văn bản soạn thảo trong Winword có đúng với mẫu văn bản chuẩn đã được ấn định trước hay không. Giải pháp đề xuất là sử dụng phương pháp chuyển đổi tệp văn bản Winword cần thẩm định sang tệp chiếu XML để xử lý so khớp các phần tử trong tệp chiếu XML. Công cụ này giúp xử lý tự động các loại văn bản đáp ứng được nhu cầu quản lý văn bản tại các cơ quan, doanh nghiệp. bắt buộc

ABSTRACT

In the process of document management, there is a general assessment of a document whether it is appropriate to common purposes. This requires obligatory verification and validation of its content as well as the format of a related document. Consequently, users must spend much time and effort, and they even commit mistakes and blunders. In this paper we present a tool system for identifing any Winword document with a defined document template. By transferring a Winword file into the XML match file so as to compare the unification of the coincidence of elements in the XML match file, this tool enables users to automatically process a variety of documents, meeting the needs of managing all kinds of documents in business offices and companies.

1. Đặt vấn đề

Trong cuộc sống, nhu cầu giao tiếp, trao đổi, quản lý hồ sơ, văn bản hành chính (gọi tắt là xử lý văn bản, tiếng Anh : Winword Processing)... luôn luôn phải được đáp ứng, ngày càng nhiều, đa dạng và phong phú [3], [4]. Trong lĩnh vực ứng dụng công nghệ thông tin, xử lý văn bản đã trở nên quen thuộc và phổ biến. Tuy nhiên thường xảy ra hiện tượng các văn bản không được trình bày theo mẫu nhất quán về mặt hình thức, nội dung, dễ có sai sót về sử dụng và tạo ra văn bản. Thực tế, khi một văn bản được tạo ra, chất lượng văn bản đó không chỉ phụ thuộc vào tính chuyên nghiệp, mà còn phụ thuộc vào thói quen, sở thích người tạo ra văn bản, và còn nhiều những yếu tố khác nữa. Việc soạn thảo nội dung, hay định dạng văn bản (Text Formatting) đôi khi tùy tiện như vậy đã gây ra nhiều khó khăn cho người đọc, người xử lý, làm người đọc tốn nhiều công sức, thời gian để tiếp nhận, sử dụng. Mặt khác, việc xử lý soạn thảo hồ sơ văn bản một cách thủ công, giao tiếp thuần tuý giấy tờ, trong bối cảnh hiện nay sẽ không còn phù

hợp và mang lại hiệu quả không cao, làm cho nhiều đơn vị tổ chức nhà nước và doanh nghiệp mất đi nhiều cơ hội phát triển kinh tế của đơn vị mình.

Mặc dù bản chất các hồ sơ, văn bản ở các đơn vị khác nhau thường không giống nhau, nhưng giống nhau ở quá trình xử lý văn bản. Đó là quá trình thực hiện các thao tác tạo ra văn bản, luân chuyển văn bản từ nguồn (nơi phát) đến đích (nơi nhận), sử dụng văn bản và quản lý văn bản (lưu trữ, chuyển giao). Thực tế, sai sót thường xảy ra ở bước tạo ra văn bản. Ở bước này, văn bản phải được kiểm tra xem đã soạn thảo đúng với mẫu quy định hay không, trước khi gửi đi, trước khi sử dụng. Có hai mức kiểm tra là kiểm tra hình thức và kiểm tra nội dung văn bản. Việc kiểm tra hình thức phụ thuộc vào cách inấn, tr ình bày văn bản, thực chất là kiểm tra định dạng. Việc kiểm tra nội dung thường mất nhiều thời gian, công sức hơn, nhiều khi phải qua tay nhiều người, nhiều vòng luân chuyển trước khi hoàn tất.

Để có được văn bản đúng đắn, có hiệu quả sử dụng cao, bước kiểm tra trở nên đóng vai trò quan trọng trong xử lý văn bản. Tuy nhiên, khi khối lượng hồ sơ văn bản lớn thì việc kiểm tra do con người thực hiện mất rất nhiều thời gian, nhàm chán và rất khó bảo đảm tính nhất quán, tính đúng đắn toàn cục và khả năng dùng lại những văn bản đã có.

Trong bài báo này, chúng tôi đề xuất giải pháp xây dựng hệ thống quản lý mẫu văn bản với công cụ cho phép kiểm tra một văn bản đã cho có soạn thảo đúng với mẫu văn bản chuẩn đang quản lý hay không. Chúng tôi sử dụng mẫu văn bản của Winword (Winword Document Template) để xây dựng mẫu chuẩn về cấu trúc, định dạng và nội dung văn bản. Nội dung bài báo như sau : sau phần mở đầu, chúng tôi trình bày tóm lược về các cấu trúc mẫu văn bản, phần tiếp theo trình bày giải pháp xây dựng công cụ, cuối cùng là phần đánh giá kết quả và kết luận.

2. Tìm hiểu các cấu trúc mẫu văn bản

2.1. Mẫu văn bản Winword

Trong Winword, mẫu văn bản [1] là một tập hợp các dạng thức (Style). Mỗi dạng thức thể hiện cách định dạng (Format) một đoạn văn bản (Paragraph) được định nghĩa bởi các lệnh đơn định dạng (*Format*) như thay đổi phông chữ sử dụng (*Format-Font*)…, thay đổi cách trình bày các đoạn văn bản (*Format-Paragraph*), v.v… Winword có sẵn các mẫu văn bản, trong đó có mẫu chuẩn là Normal gồm các dạng thức có cấu trúc phân cấp từ Heading1 đến Heading9, dạng thức đoạn văn bản chuẩn Normal và các dạng thức khác. Trong quá trình sử dụng Winword, người sử dụng (NSD) có thể tuỳ ý sửa đổi hay tạo mới các mẫu văn bản tuỳ theo nhu cầu.

Theo cách nhìn khác, khi soạn thảo văn bản với Winword, NSD có thể sử dụng đồng thời hay một trong ba thành phần cấu trúc của Winword là phần ứng dụng (Winword Application), phần văn bản (Document) và phần mẫu (Template). Theo đó, mỗi thành phần tác động lên văn bản WinWord một cách khác nhau. Phần ứng dụng

cung cấp các menu chuẩn, các lệnh (Command) và các thanh công cụ (Toolbar). Phần mẫu Template được dùng với hai mục đích : hoặc cung cấp các kiểu mẫu có sẵn để tạo những văn bản mới theo các chủ đề khác nhau, hoặc làm nơi lưu trữ các dạng thức mới tạo, các Macros, các dữ liệu của AutoText, AutoCorrect, các lệnh và cấu hình của Toolbar đã được "tùy chọn hóa" (Customized). Cuối cùng, phần văn bản chứa văn bản, hình ảnh, các thông tin về định dạng (bao gồm định dạng trang, định dạng ký tự, đoạn) cho ngay chính văn bản đó, v.v…

2.2. Mẫu văn bản XML

XML (eXtension Makup Language) [5] là ngôn ngữ đánh dấu mở rộng được phát triển trên cơ sở ngôn ngữ đánh dấu siêu văn bản HTML (High Text Makup Language) và nhằm khắc phục những hạn chế của HTML. HTML được sử dụng phổ biến trong các trình duyệt web. XML mô tả dữ liệu dựa trên việc dùng các thẻ tự định nghĩa hay tự mô tả nội dung và ý nghĩa của nó.

XML có ba thành phần cơ bản là thẻ, phần tử, và thuộc tính. Mỗi thẻ được đặt giữa cặp dấu ngoặc đơn bên trái (**<**) và ngoặc đơn bên phải (**>**). Có thẻ bắt đầu (như **<**name**>**) và thẻ kết thúc (như **</name>**). Mỗi phần tử được xác định bởi một thẻ bắt đầu, một thẻ kết thúc, và mọi thứ giữa chúng. Thuộc tính là một cặp giá trị tên trong thẻ bắt đầu của một phần tử có cú pháp là **<name attribute =** "value"**> content </name>**.

XML làm đơn giản hóa quá trình trao đổi dữ liệu. Sử dụng XML, NSD có thể chuyển đổi những định dạng dữ liệu bên trong các cơ sở dữ liệu trở thành XML và ngược lại. Các văn bản XML được tổ chức để nhận dạng từng thông tin quan trọng và mối quan hệ giữa các thông tin đó, có thể viết mã để xử lý văn bản XML mà không cần tác động của NSD. Mặt khác XML cho phép tìm kiếm thông minh. Mặc dù hiện nay các công cụ tìm kiếm đã được cải thiện dần, tuy nhiên nhận được những kết quả không chính xác vẫn phổ biến xảy ra. Chẳng hạn trong tìm kiếm Google, nếu cần tìm kiếm một người nào đó mang tên "Chip" trong những trang HTML, NSD sẽ tìm thấy một loạt các trang web về chip sô -cô-la, chip máy tính, chip gỗ và nhiều thứ vô dụng khác. Tìm kiếm văn bản XML cho <first-name> các yếu tố chứa từ Chip sẽ mang lại cho NSD những kết quả tốt hơn rất nhiều.

2.3. Chuyển đổi mẫu văn bản Winword sang XML

Mỗi mẫu văn bản Winword đã được định dạng chuẩn chứa các thuộc tính dạng thức (Font, Paragraph, Bullet, Margin, Page Setup...), các thông tin về chính tả (Spelling, Autocorrect...) có thể chuyển đổi thành văn bản XML. Khi chuyển đổi Winword-XML yêu cầu phải có tính tương tác giữa các tệp văn bản XML và tệp nguồn Winword DOC, nghĩa là phải làm sao cho tệp XML thừa kế được cấu trúc và các thuộc tính của văn bản trong tệp DOC. Chúng tôi quy ước như sau:

GIẤY BÁO KẾT QUẢ

Hội đồng chiêu sinh trường Cao Đẳng Giao thông Vận tải II báo cho:

- Thí sinh: Nguyễn Văn Hùng, sinh ngày: 15-12-1988

Đã được xét hồ sơ vào học tại trường Cao Đẳng Giao thông Vận tải II, ngành Sửa chữa Ô tô.

```
1. <?xml version="1.0" encoding="utf-8" ?>
2. <document xmlns="Schema.Winword">
3. <section>
4. <pageSetup margin-top="72" margin-right="72" margin-bottom="72" margin-
   left="72" pageSize="letter" pageOrientation="portrait" />
5. <body>
6. <p listType="none" align="center" leftToRight="true" firstLineIndent="0"
   leftIndent="0" rightIndent="0" spaceBefore="0" spaceAfter="0"
   lineSpacing="20">
7. <run font-size="18" font-name="Times New Roman" font-color="#000000"
   leftToRight="true" font-bold="true" font-italic="false" font-underline="false"
   font-strikeThrough="false">GIẤY BÁO KẾT QUẢ</run>
   </p>
```

Hình 1. *Chuyển đổi một tệp Winword DOC sang XML.*

Ứng với mỗi kiểu dạng thức Style gồm tên kiểu (Style Name), kiểu định dạng (Style Type), các hình thức phông chữ (Font, Size, ...) sẽ có một thẻ tương ứng trong XML. Nội dung của mỗi kiểu dạng thức được giữ nguyên và chuyển thành nội dung của các thẻ trong XML. Dữ liệu trong các tệp XML được biểu diễn bởi các cặp thẻ mở và đóng. Thẻ quản lý từng dạng thức cho các tệp XML đều được đặt là <run>. Hình 1 minh hoạ ví dụ chuyển đổi một tệp Winword DOC sang XML.

3. Giải pháp xây dựng hệ thống quản lý mẫu văn bản

3.1. Xây dựng quy trình xử lý văn bản

Chúng tôi xây dựng hệ thống quản lý mẫu văn bản dựa trên bốn quy trình như sau :

1. Tiếp nhận và chuyển giao văn bản.
2. Xử lý văn bản.
3. Ban hành văn bản.
4. Lập hồ sơ công việc.

Từ bốn quy trình đã mô tả, chúng tôi xây dựng hệ thống quản lý mẫu văn bản trên cơ sở giải pháp chuyển đổi một văn bản Winword DOC đang xét sang cấu trúc XML, được đặt tên là *tệp chiếu* (Match File). Sau đó, từ tệp chiếu XML, tiến hành bóc tách các thẻ, phần tử, lấy ra các thuộc tính và nội dung văn bản để thiết lập hay kiểm tra tính nhất quán của mẫu văn bản Template phục vụ thao tác xác thực văn bản của hệ thống.

3.2. Xác định giá trị so sánh cho mẫu

Sau bước chuyển đổi một văn bản Winword DOC, dựa trên cấu trúc XML của tệp chiếu, bước tiếp theo là thêm vào tệp chiếu phần tử thẻ <match> để xác định vùng nội dung văn bản chuẩn và thuộc tính cấu trúc của vùng. Phần tử *match* cho phép đánh dấu kiểm tra ngôn từ, phục vụ so khớp với mẫu văn bản chuẩn. Phần tử *match* được đặt trong thẻ <run> có các kiểu kiểm tra như sau :

- *Exact* : Đọan văn bản có giá trị chính xác với một giá trị nhất định.
- *Starts with* : Đọan văn bản phải bắt đầu bằng một cụm từ nhất định.
- *Ends with* : Đọan văn bản phải kết thúc bằng một cụm từ nhất định.
- *Contains* : Đọan văn bản phải chứa một cụm từ nhất định.
- *Lower case* : Viết thường.
- *Upper case* : Viết hoa.

Ví dụ : trong tệp chiếu XML của văn bản chuẩn đã cho trong hình 3, thẻ <match> xác định giá trị mẫu chuẩn để xác thực được thêm vào như sau :

<match match="Exact"> GIẤY BÁO NHẬP HỌC </match>

3.3. Xác thực văn bản

Bước xác thực văn bản theo mẫu có các chức năng chính như sau :

Tìm nội dung so khớp : – NSD đưa văn bản Winword DOC cần so với mẫu vào hệ thống. Hệ thống tự động chuyển đổi (Conversion) sang tệp chiếu XML và bắt đầu thực hiện việc kiểm tra xác thực bằng cách tìm nội dung so khớp trên văn bản chuẩn. Từng nội dung được phân tích lấy mẫu để so khớp, từ đó kết luận văn bản có hợp lệ không.

So trùng cấu trúc : – Tiến hành so trùng cấu trúc được bằng cách duyệt từng phần tử trong tệp chiếu XML nhưng bỏ qua phần tử đánh dấu match. Nếu trong mẫu Template có những phần tử nào thì văn bản so khớp cũng phải chứa những phần tử ấy, tuân thủ theo đúng trình tự xuất hiện và cấu trúc phân cấp.

Kiểm tra chính tả, nội dung : – Việc xác thực dựa theo giá trị đoạn text trong thẻ run. Hệ thống sẽ kiểm tra tuần tự cấu trúc XML của từng văn bản. Nếu văn bản XML của tài liệu là đúng thì dòng xử lý chứa con trỏ xử lý trỏ đến tiêu đề trong văn bản XML. Hệ thống tiến hành kiểm tra chính tả, nội dung tiêu đề trong Template và lần lượt duyệt đến hết văn bản. Con trỏ xử lý tìm đến vị trí từng đoạn text trong run và luôn luôn

hoạt động song song, sử dụng hàm duyệt thăm lấy ra nội dung yêu cầu của văn bản template để xác thực với văn bản.

Mỗi khi duyệt thăm rút trích thông tin đoạn text, đoạn text được chia ra thành các từ, mỗi từ cách nhau một ký tự trắng. Sau đó dò từng từ trong từ điển tiếng Việt, từ không có trong từ điển là từ sai chính tả. Tro ng giải pháp, chúng tôi sử dụng từ điển tiếng Việt là kho ngữ liệu hiện có [1], [2]. Quy trình gồm hai thao tác như sau :

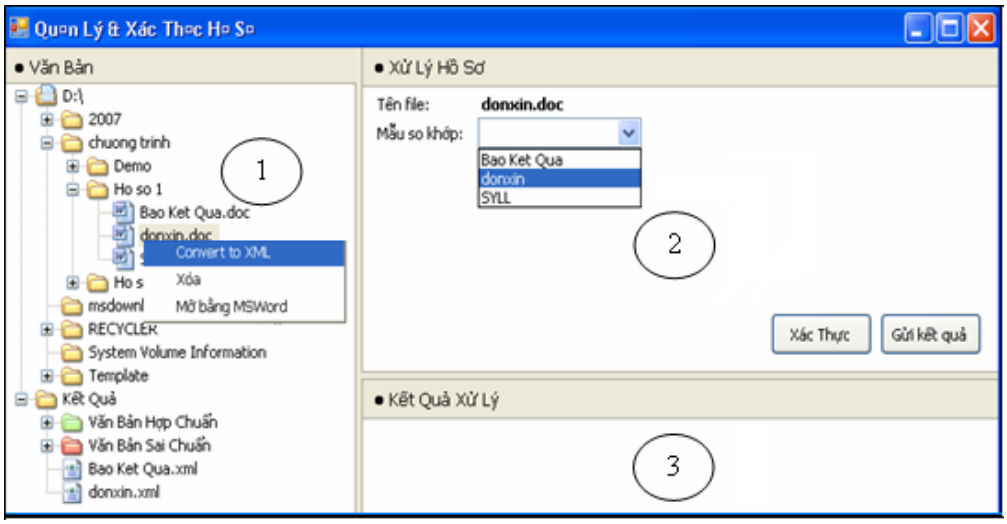
- Hệ thống duyệt thăm tại vị trí phần tử *match* và lấy đoạn text ra.

- Hệ thống duyệt từng từ của đoạn text so sánh với từ điển để kiểm tra.

Kiểm tra ngôn từ : – Kiểm tra ngôn từ là so sánh nội dung giữa hai văn bản có giống nhau hay không, nghĩa là so sánh cả cấu trúc định dạng thuộc tính và nội dung đoạn text có trùng nhau không. Lúc này ta dùng phần tử đặt biệt <match> đã được xây dựng trước đây. Hệ thống tiến hành duyệt con trỏ đến phần tử match và xác thực theo từng ràng buộc của match đã qui định trước trong văn bản mẫu.

Thu nhận kết quả : – Nếu đúng mẫu, hệ thống thông báo kết quả văn bản hợp chuẩn, nếu không, hệ thống cho phép người sử dụng cập nhật lại theo đúng mẫu.

3.4. Mô tả chức năng công cụ



Hình 5. *Giao diện công cụ xác thực mẫu văn bản.*

Văn bản mẫu Template được lưu trong thư mục Template và được hiển thị ở hộp combobox trên ửa sổ xử lý hồ sơ của giao diện. NSD chuyển một tệp văn bản Template sang XML bằng cách nhấp chuột phải (Right Click) vào tên văn bản đó, chọn lệnh chuyển (Convert), hệ thống sẽ tự động chuyển tệp văn bản đó sang dạn g tài liệu XML. Bước tiếp theo, NSD sử dụng lệnh đổi tên (Rename) để đặt lại tên tệp bằng cách thay đổi phần mở rộng, chẳng hạn .DOC thành .XML. Sau đó, NSD đánh dấu đoạn mẫu chuẩn của văn bản bằng phần tử Match.

4. Kết luận

Với giải pháp chuyển đổi văn bản Winword DOC sang cấu trúc XML, chúng tôi đã xây dựng được hệ thống quản lý mẫu văn bản. Ý nghĩa thực tiễn của giải pháp là giúp các cơ quan, đơn vị giải quyết được những vấn đề quản lý quy trình xử lý văn bản, cho phép tiết kiệm thời gian, công sức, tiết kiệm chi phí, giảm thiểu các nhầm lẫn, sai sót phát sinh. Trên cơ sở kết quả đạt được, chúng tôi sẽ tiếp tục thực hiện việc xử lý nhiều thành phần của tài liệu Winword để có được dạng XML chính xác thể hiện được đầy đủ hầu hết các mẫu hồ sơ. Xây dựng các module chuyển các tài liệu PDF, RTF, HTML, … sang dạng XML để mở rộng khả năng xác thực nhiều dạng văn bản khác, thay vì chỉ xử lý được văn bản Winword như hiện nay.

TÀI LIỆU THAM KHẢO

[1] Phan Huy Khánh (2005), Nghiên cứu xây dựng cơ sở dữ liệu từ vựng danh từ kết hợp trong tiếng Việt, *Kỷ yếu Hội thảo Khoa học Quốc gia Lần thứ 8 Hải Phòng* 08/2005.

[2] Phan Huy Khánh (2003), Xây dựng từ điển đa ngữ sử dụng dạng thức văn bản RTF Winword, *Hội thảo Khoa học Quốc gia lần thứ nhất về Nghiên cứu Phát triển và Ứng dụng CNTT và Truyền thông,* ICT.rda'2003 Hà Nội.

[3] Võ Hồng Lan (2007), *Đề xuất quy trình mẫu về quản lý văn bản tại cơ quan cấp bộ*, Phòng Nghiệp vụ hành chính, Vụ Hành chính, Văn phòng Chính phủ.

[4] Vũ Thị Phụng (2007), Nghiên cứu chuẩn hóa quy trình quản lý và xử lý văn bản, *Kỷ yếu Hội thảo Lưu trữ học và Quản trị văn phòng,* Trường Đại học Khoa học Xã hội và Nhân Văn, Đại học Quốc gia Hà Nội.

[5] Đặc tả XML 1.1 http ://www.w3.org/TR/xml11/

[6] Cấu trúc tài liệu Winword, MicrosoftOffice Help Documents 2003.