TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN
BỘ MÔN CÔNG NGHỆ PHẦN MỀM

[News Fetcher]
ỨNG DỤNG LẤY TIN TỰ ĐỘNG TỪ CÁC BÁO ĐIỆN TỬ

Nhóm: Nguyễn Ngọc Khánh – 1241350
Bùi Bá Lộc – 1241363
Dương Diệu Pháp – 1241378
Nguyễn Quốc Tuấn – 1241431

Tp HCM – 2013.

MỤC LỤC

1. THÔNG TIN NHÓM & DANH SÁCH CHỨC NĂNG

1.1 Thành viên

Họ tên	MSSV	Đánh giá
Nguyễn Ngọc Khánh	1241350	100%
Bùi Bá Lộc	1241363	100%
Dương Diệu Pháp	1241378	100%
Nguyễn Quốc Tuấn	1241431	100%

1.2 Danh sách chức năng và phân công công việc

Danh sách chức năng		Thành viên
	Thiết kế giao diện	Pháp
	Thiết kế CSDL	Lộc
Chức năng 1	Cấu trúc tập tin XML lưu trữ Xpath	Lộc, Khánh
	Lấy Xpath mẫu	Lộc, Khánh
	Tạo lớp đối tượng CSDL trong java	Khánh
	Đọc, ghi tập tin XML	Khánh
	Thêm, xoá, sửa thông tin website	Tuấn
Chức năng 2	Lấy nội dung bài viết dựa trên Xpath	Pháp
	Lấy tin của trang kế tiếp	Pháp
	Lưu bài viết xuống CSDL	Tuấn
Chức năng 3	Xem danh sách bài viết của chuyên mục	Tuấn
	Xoá bài viết không phù hợp	Tuấn
	Xây dựng file build.xml	Lộc
	Viết báo cáo	Pháp

2. HƯỚNG DẪN SỬ DỤNG

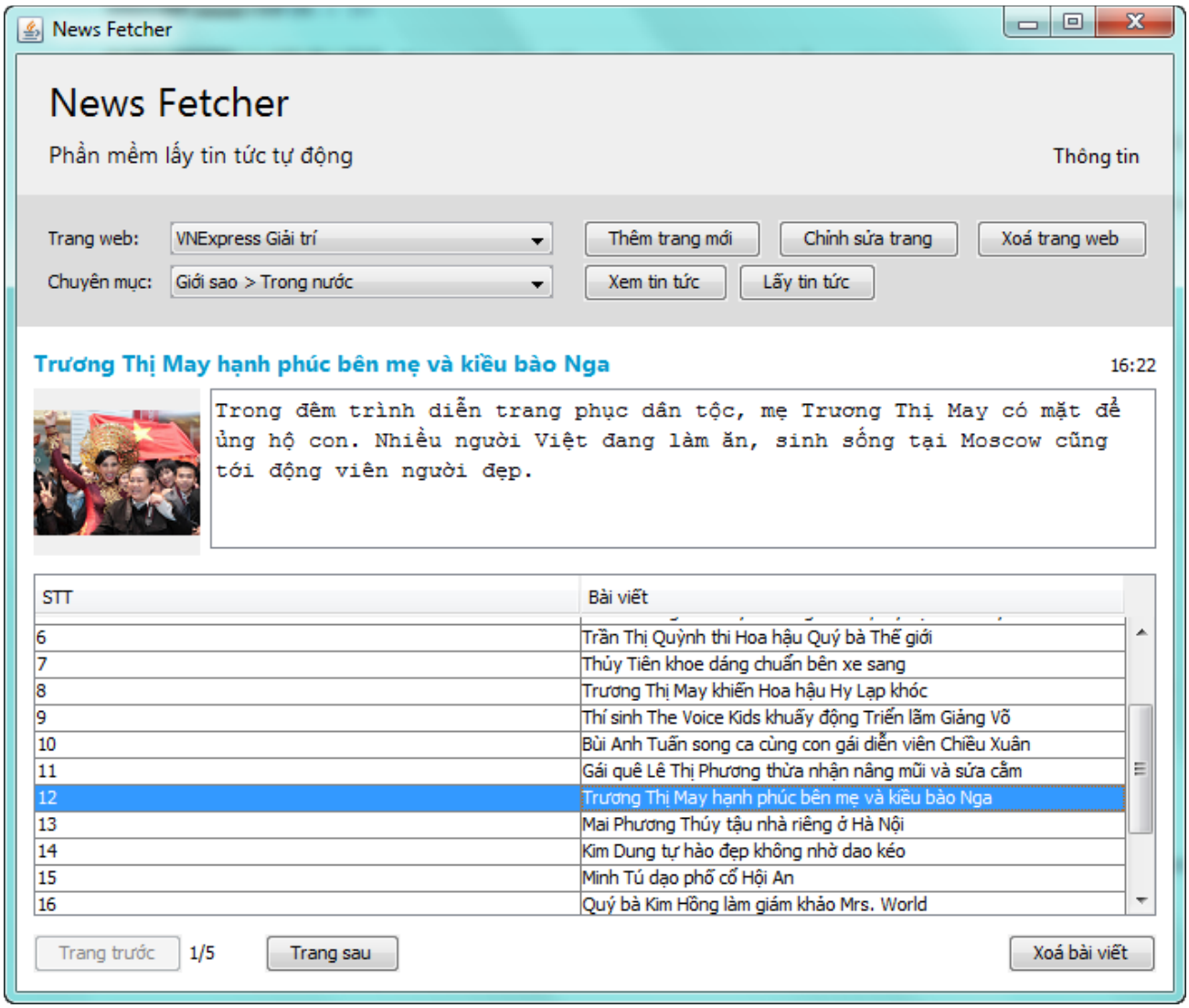
2.1 Cấu trúc tập tin XML lưu trữ Xpath

```
<?xml version="1.0" encoding="UTF-8"?>
<_1241350_1241363_1241378_1241431>
    <website>VNExpress Giải trí</website>
    <url>http://giaitri.vnexpress.net</url>
    <category count="5">
        <item id="1" name="Giới sao > Trong nước" url="
        http://giaitri.vnexpress.net/tin-tuc/gioi-sao/trong-nuoc">
            <xpathlayout>//*[@id="box_list_news_sub"]/div/ul/li</xpathlayout>
            <link>//*[@id="box_list_news_sub"]/div/ul/li[1]/h2/p/b/a/@href</link>
            <title>//*[@id="box_list_news_sub"]/div/ul/li[1]/h2/p/b/a</title>
            <image>//*[@id="box_list_news_sub"]/div/ul/li[1]/a/img/@src</image>
            <date>//*[@id="box_list_news_sub"]/div/ul/li[1]/p[1]/span</date>
            <description>//*[@id="box_list_news_sub"]/div/ul/li[1]/p[2]/text()
            </description>
            <nextpage>//*[@id="pagination"]/ul/li[2]/a</nextpage>
            <xpathnextpage>//*[@id="pagination"]/ul</xpathnextpage>
        </item>
```

Các thẻ chính:

- <website>: Tên trang web
- <url>: Địa chỉ trang chủ
- <category>: Chứa danh sách các chuyên mục
    - <item>: Chuyên mục
        - Name: Tên chuyên mục
        - url: Địa chỉ của chuyên mục sẽ lấy tin
    - <xpathlayout>: Xpath của thẻ bao trùm bài viết
    - <link>: Xpath của địa chỉ bài viết chi tiết
    - <title>: Xpath của tiêu đề bài viết
    - <image>: Xpath của hình ảnh minh hoạ
    - <date>: Xpath của ngày đăng bài viết
    - <description>: Xpath cảu miêu tả tóm tắt bài viết
    - <nextpage>: Xpath đến địa chỉ của trang tiếp theo trong chuyên mục
    - <xpathnextpage>: Xpath của thẻ bao trùm phần phân trang

2.2 Giao diện chương trình

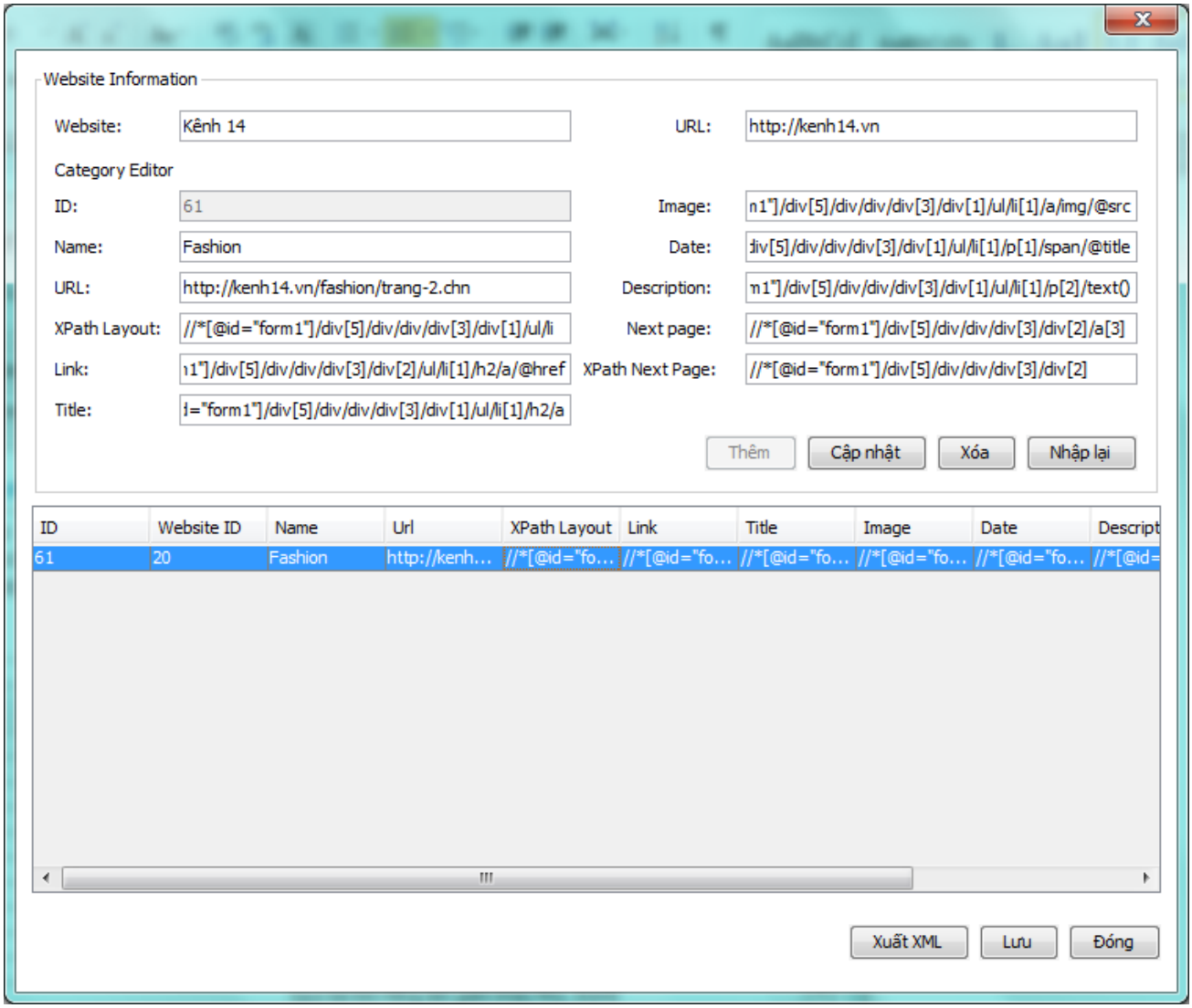


2.3 Thêm mới / chỉnh sửa 1 website

Nhấn vào nút **Thêm trang mới** và tìm chọn tập tin XML để nhập vào CSDL.

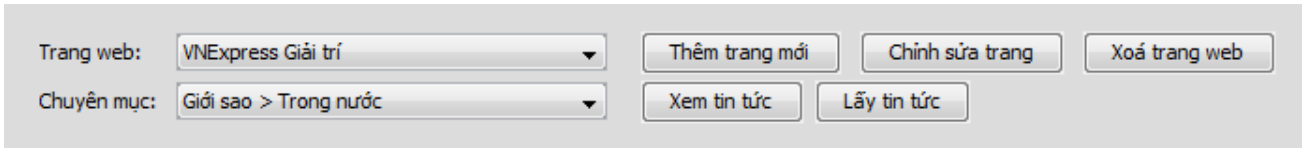
Sau khi thêm thành công, thì danh sách, thông tin chuyên mục của trang web này sẽ được liệt kê chi tiết tại hộp thoại như hình bên dưới.

Bạn có thể thay đổi thông tin của trang web bằng cách chỉnh sửa các trường dữ liệu tương ứng. Sau đó nhấn **Lưu** để các thay đổi được cập nhật xuống CSDL.



Đi kèm là chức năng xuất ra định dạng XML giúp ích cho việc lưu trữ dưới dạng tập tin, thuận tiện nhu cầu sao chép, chia sẻ dữ liệu.

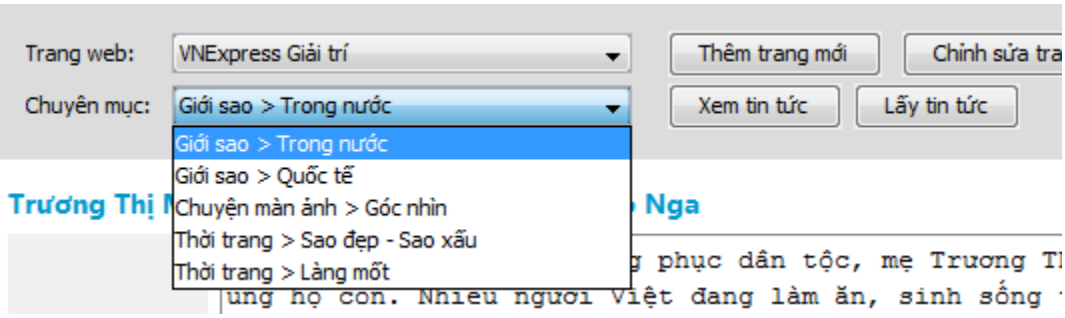
2.4 Xoá trang web



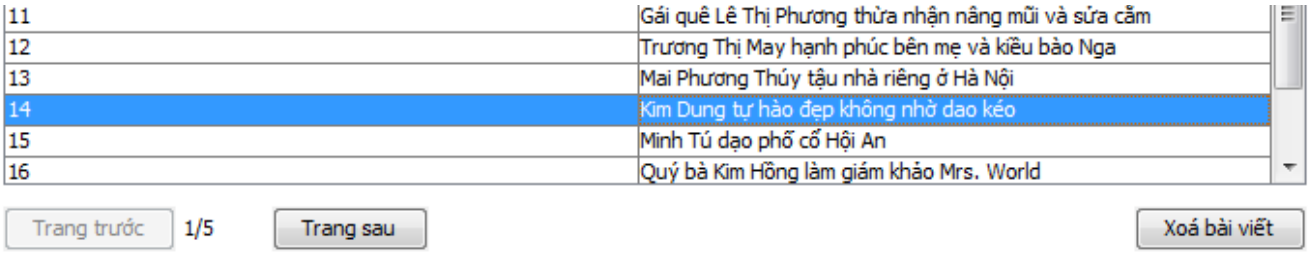
Chọn trang web cần xoá, sau đó nhất nút **Xoá trang web**.

Lưu ý: Chức năng này sẽ xoá tất cả thông tin của trang web, bao gồm những chuyên mục, bài viết có liên quan.

2.5 Xem bài viết 1 chuyên mục của trang web



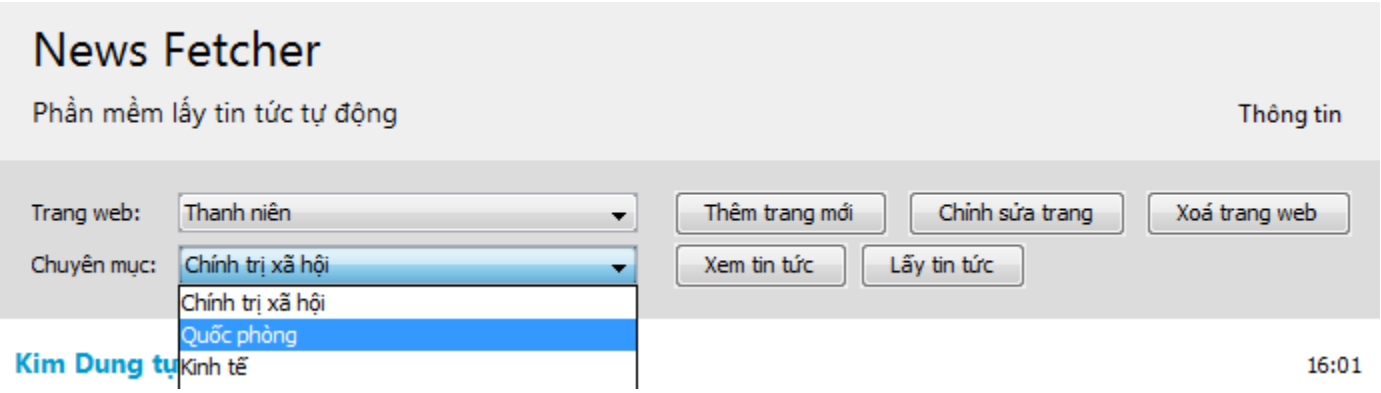
Đầu tiên chọn trang web cần xem, danh sách chuyên mục hỗ trợ lấy tin sẽ được liệt kê tại combo box bên dưới, bạn chỉ việc chọn chuyên mục cần xem. Tiếp theo nhấn nút **Xem tin tức** để chương trình tải danh sách bài viết đã được lưu trữ trong CSDL.



Để chuyển qua trang tiếp theo hoặc xem lại bài viết của trang trước, bạn chỉ cần nhấn nút **Trang sau** / **Trang trước** tương ứng.

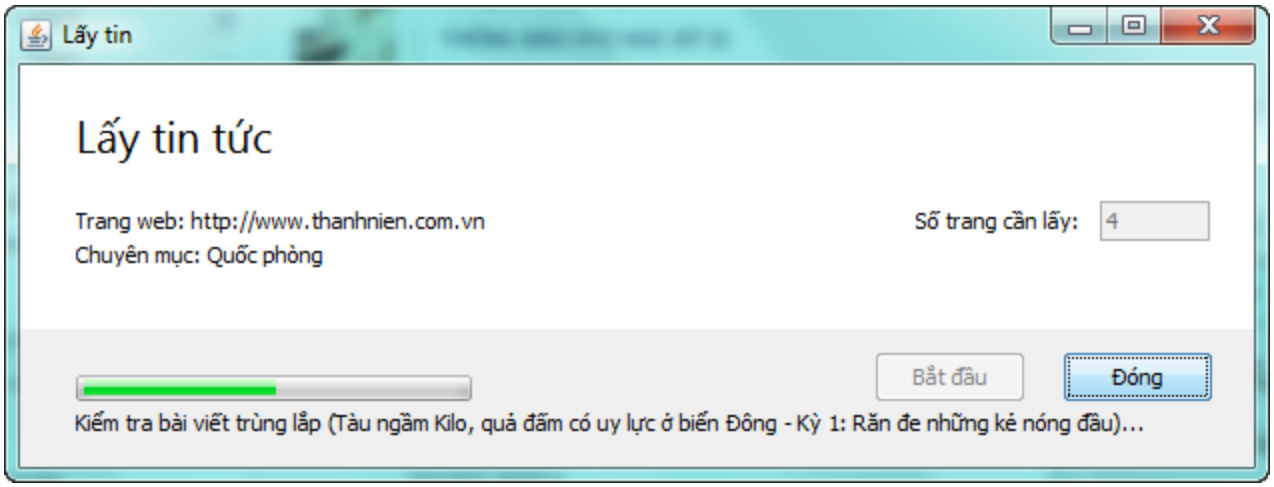
Bạn cũng có thể loại bỏ bài viết không thích bằng cách nhất nút **Xoá bài viết**.

2.6 Cập nhật bài viết mới của chuyên mục



Chọn chuyên mục của trang web cần lấy, sau đó nhấn nút **Lấy tin tức**.

Hộp thoại cấu hình Lấy tin hiện ra. Mặc định chương trình sẽ lấy bài viết trong 4 trang mới nhất của chuyên mục, bạn có thể tuỳ chỉnh lại nếu muốn lấy nhiều hơn.



Nhấn nút **Bắt đầu** để chương trình thực hiện. Tốc độ lấy bài viết khá nhanh, ví dụ như trang *Thanh niên*, chuyên mục *Quốc phòng*, chương trình lấy 4 trang mới nhất chỉ tốn vài giây ít ỏi.

Sau khi lấy hoàn tất, chương trình sẽ tự động kiểm tra bài viết lấy được và sẽ loại bỏ nó nếu bị trùng vì thế bạn không phải lo lắng gì cả!

3. THAM KHẢO

HttpClient 4.3.1

http://hc.apache.org/downloads.cgi

JDOM 2.0.5

http://jdom.org/downloads/source.html

HTML Cleaner 2.6.1

http://htmlcleaner.sourceforge.net/download.php

Apache Common Lang 3.1

http://commons.apache.org/proper/commons-lang/download_lang.cgi

Microsoft JDBC Driver 4.0 for SQL Server

http://www.microsoft.com/en-us/download/details.aspx?displaylang=en&id=11774