

**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 .....	3
1.1 Thành viên .....	3
1.2 Danh sách chức năng và phân công công việc .....	3
2. HƯỚNG DẪN SỬ DỤNG .....	4
2.1 Cấu trúc tập tin XML lưu trữ Xpath.....	4
2.2 Giao diện chương trình.....	5
2.3 Thêm mới / chỉnh sửa 1 website.....	5
2.4 Xóa trang web .....	6
2.5 Xem bài viết 1 chuyên mục của trang web .....	7
2.6 Cập nhật bài viết mới của chuyên mục .....	7
3. THAM KHẢO.....	8

# 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

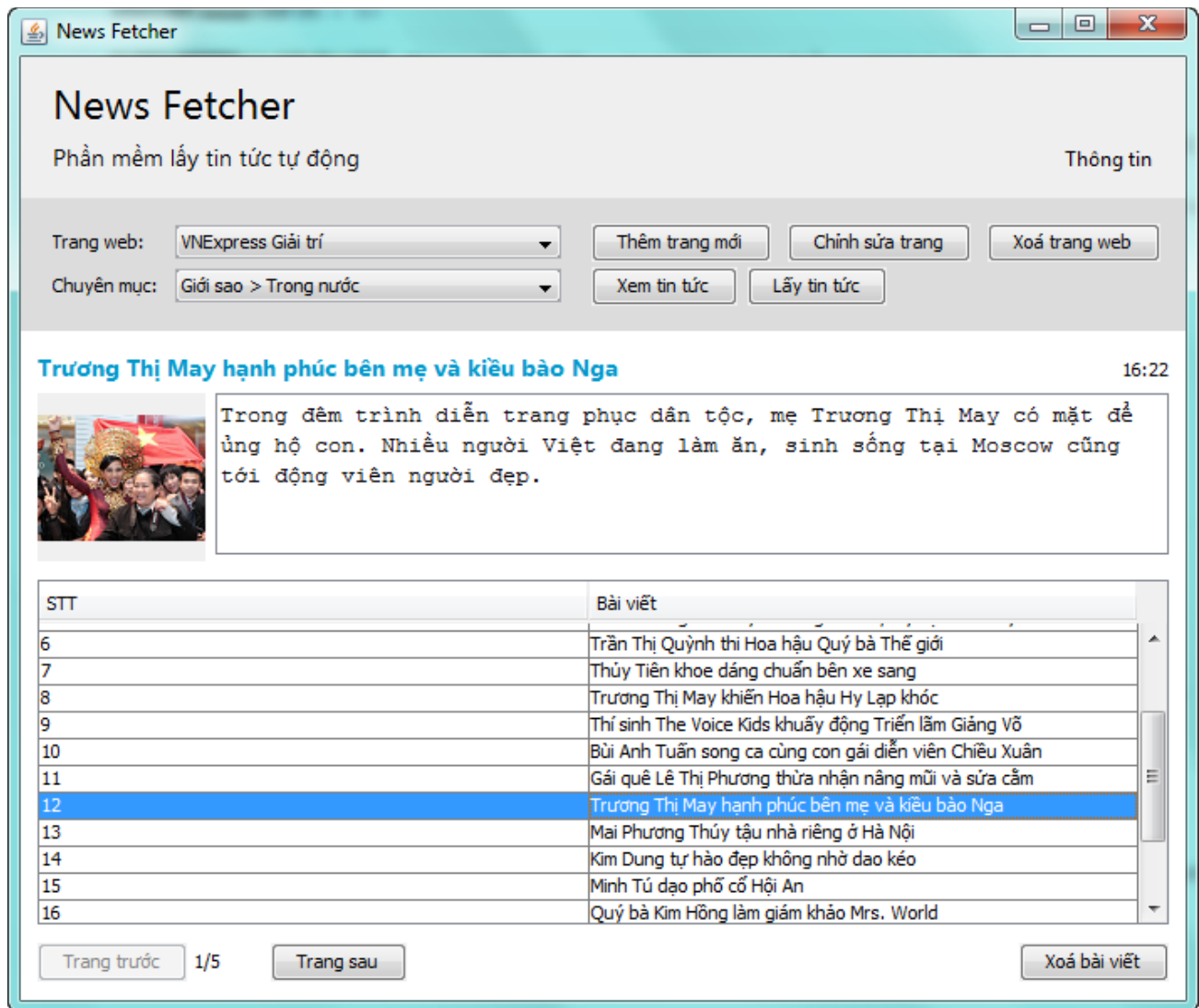
### 2.1 Cấu trúc tập tin XML lưu trữ Xpath

```
<?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
  - o <item>: Chuyên mục
    - Name: Tên chuyên mục
    - url: Địa chỉ của chuyên mục sẽ lấy tin
  - o <xpathlayout>: Xpath của thẻ bao trùm bài viết
  - o <link>: Xpath của địa chỉ bài viết chi tiết
  - o <title>: Xpath của tiêu đề bài viết
  - o <image>: Xpath của hình ảnh minh họa
  - o <date>: Xpath của ngày đăng bài viết
  - o <description>: Xpath của câu miêu tả tóm tắt bài viết
  - o <nextpage>: Xpath đến địa chỉ của trang tiếp theo trong chuyên mục
  - o <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.

Website Information

Website: Kênh 14

URL: http://kenh14.vn

Category Editor

ID: 61

Image: n1"/div[5]/div/div/div[3]/div[1]/ul/li[1]/a/img/@src

Name: Fashion

Date: div[5]/div/div/div[3]/div[1]/ul/li[1]/p[1]/span/@title

URL: http://kenh14.vn/fashion/trang-2.chn

Description: n1"/div[5]/div/div/div[3]/div[1]/ul/li[1]/p[2]/text()

XPath Layout: //\*[@id="form1"]/div[5]/div/div/div[3]/div[1]/ul/li

Next page: //\*[@id="form1"]/div[5]/div/div/div[3]/div[2]/a[3]

Link: 11"/div[5]/div/div/div[3]/div[2]/ul/li[1]/h2/a/@href

XPath Next Page: //\*[@id="form1"]/div[5]/div/div/div[3]/div[2]

Title: t="form1"/div[5]/div/div/div[3]/div[1]/ul/li[1]/h2/a

Thêm

Cập nhật

Xóa

Nhập lại

ID	Website ID	Name	Url	XPath Layout	Link	Title	Image	Date	Descript
61	20	Fashion	http://kenh...	//*[@id="fo...	//*[@id="fo...	//*[@id="fo...	//*[@id="fo...	//*[@id="fo...	//*[@id="fo...

Xuất XML

Lưu

Đóng

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 Xóa trang web

Trang web: VNExpress Giải trí

Thêm trang mới

Chỉnh sửa trang

Xóa trang web

Chuyên mục: Giới sao > Trong nước

Xem tin tức

Lấy tin tức

Chọn trang web cần xóa, sau đó nhấn nút **Xóa trang web**.

Lưu ý: Chức năng này sẽ xóa 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

The screenshot shows a web application interface. At the top, there are two dropdown menus: 'Trang web:' with 'VNExpress Giải trí' selected, and 'Chuyên mục:' with 'Giới sao > Trong nước' selected. To the right of these are buttons: 'Thêm trang mới', 'Chỉnh sửa trang', 'Xem tin tức', and 'Lấy tin tức'. Below the 'Chuyên mục:' dropdown, a list of categories is visible: 'Giới sao > Trong nước', 'Giới sao > Quốc tế', 'Chuyện màn ảnh > Góc nhìn', 'Thời trang > Sao đẹp - Sao xấu', and 'Thời trang > Làng mốt'. Below this, there is a section titled 'Trương Thị...' and a snippet of text: 'g phục dân tộc, mẹ Trương T... ung họ con. Nhiều người Việt đang làm ăn, sinh sống'.

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.

11	Gái quê Lê Thị Phương thừa nhận nâng mũi và sửa cằm
12	Trương Thị May hạnh phúc bên mẹ và kiều bào Nga
13	Mai Phương Thúy tậu nhà riêng ở Hà Nội
14	Kim Dung tự hào đẹp không nhờ dao kéo
15	Minh Tú dạo phố cổ Hội An
16	Quý bà Kim Hồng làm giám khảo Mrs. World

Trang trước 1/5 Trang sau Xoá bài viết

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.

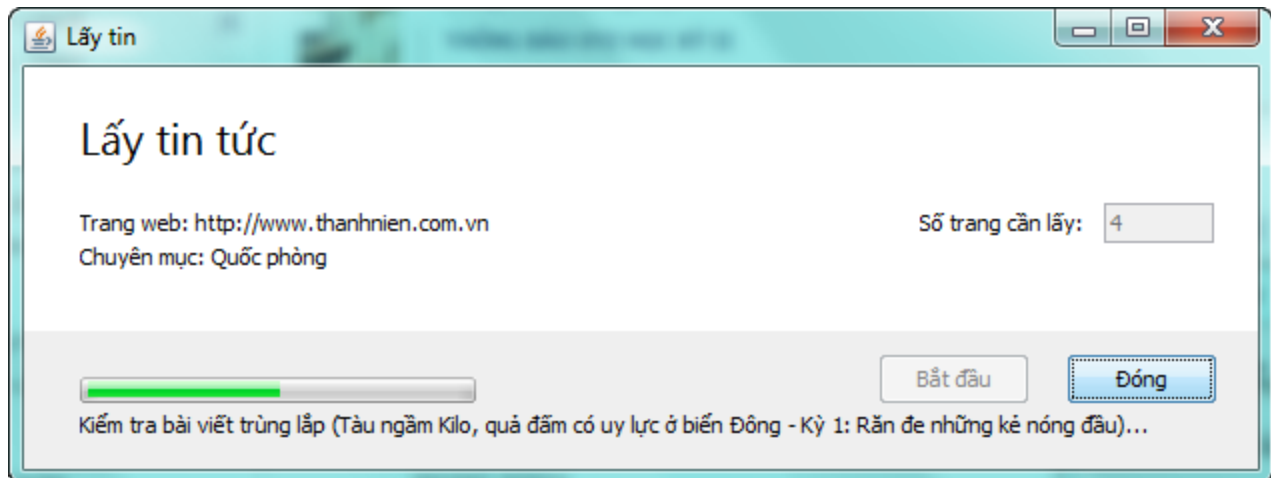
Bạn cũng có thể loại bỏ bài viết không thích bằng cách nhấn nút **Xoá bài viết**.

## 2.6 Cập nhật bài viết mới của chuyên mục

The screenshot shows the 'News Fetcher' application. It has a title 'News Fetcher' and a subtitle 'Phần mềm lấy tin tức tự động'. There is a 'Thông tin' link on the right. Below, there are two dropdown menus: 'Trang web:' with 'Thanh niên' selected, and 'Chuyên mục:' with 'Chính trị xã hội' selected. To the right are buttons: 'Thêm trang mới', 'Chỉnh sửa trang', 'Xoá trang web', 'Xem tin tức', and 'Lấy tin tức'. Below the 'Chuyên mục:' dropdown, a list of categories is visible: 'Chính trị xã hội' and 'Quốc phòng'. Below this, there is a section titled 'Kim Dung tự...' and a snippet of text: 'Kinh tế'.

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 tùy 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](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>