# Thông tin về tác giả

|  |  |  |  |
| --- | --- | --- | --- |
| Tên | MSSV | Điện Thoại | Email |
| Đỗ Văn Tiến | 06520476 | 0987215699 | [tiendo.vn@gmail.com](mailto:tiendo.vn@gmail.com) |
| Nguyễn Phước Cường | 06520060 | 0933520007 | cuongnp.uit.se@gmail.com |

# Hướng dẫn cài đặt chương trình

**Bước 1**: Cài đặt môi trường phát triển ứng dụng.

* Cài đặt JDK tại: Do Van Tien - Nguyen Phuoc Cuong\Tools\Lib
* Cài đặt Eclipse tại: Do Van Tien - Nguyen Phuoc Cuong\Tools\eclipse
* Cài đặt Server MySQL tại: Do Van Tien - Nguyen Phuoc Cuong\Tools\MySQL.

*Lưu ý*: chương trình có sử dụng hibernate để kết nối với cơ sở dữ liệu sử dụng tên và mật khẩu mặc định là :

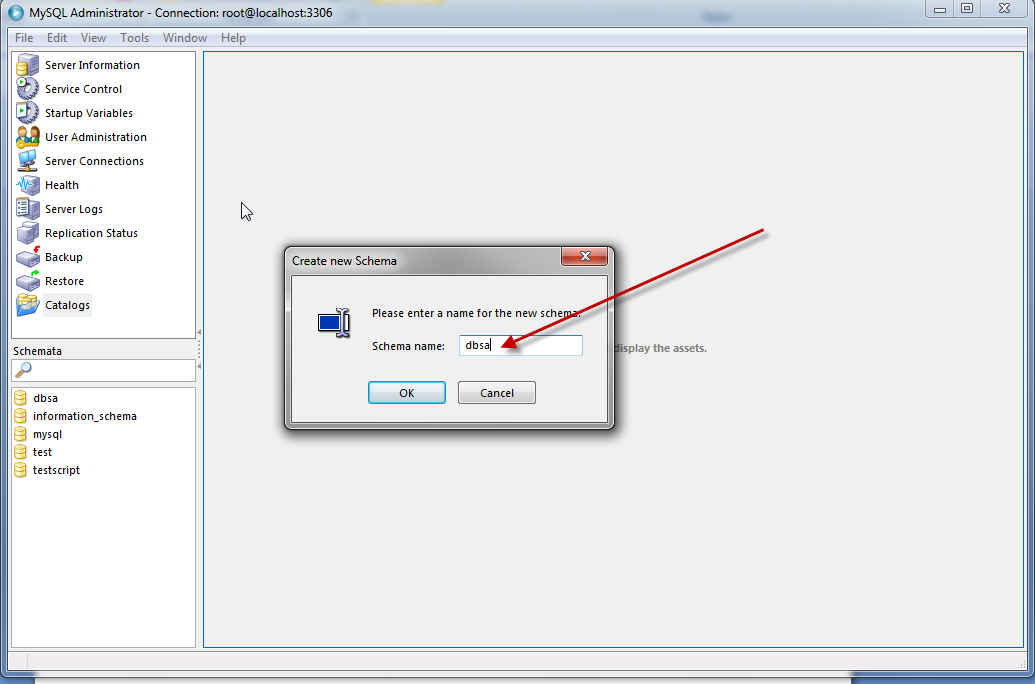
User:root

Pass: root

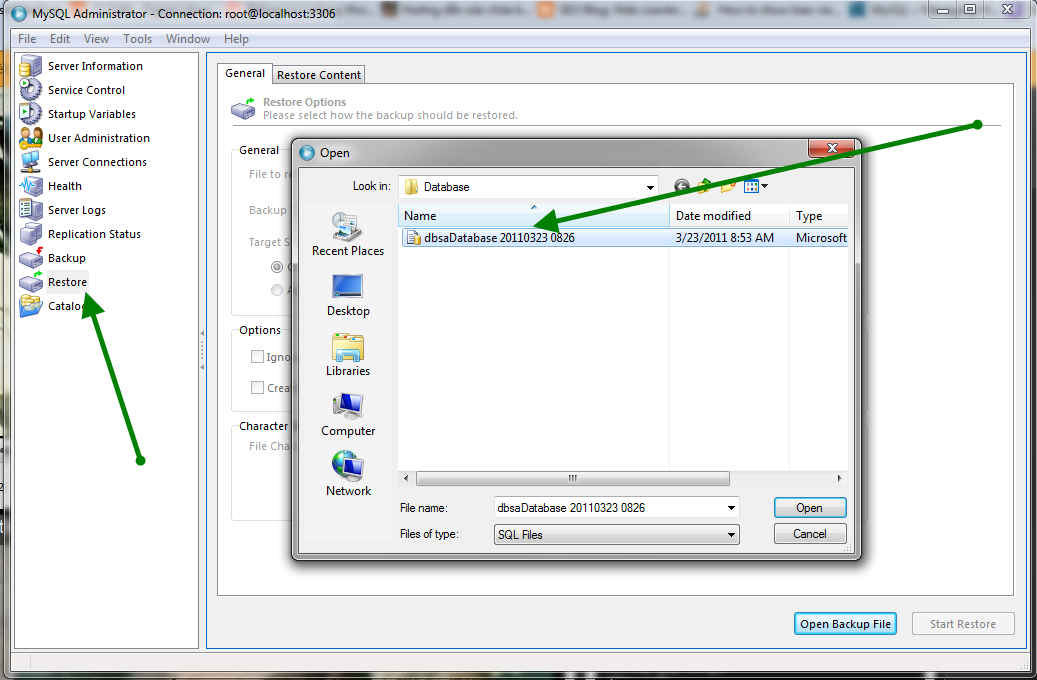
Do đó trong quá trình cài đặt nếu hai thông tin chưa phù hợp với những cài đặt trên, người dùng phải config lại thông tin cho phù hợp tránh phải vào code chỉnh sửa.

**Bước 2: Restore Database**

Trong dvd có database sẵn của chương trình, khi đã cài đặt server người dùng thực hiện các bước sau để lấy



*Hình 1: Tạo mới một database tên dbsa*



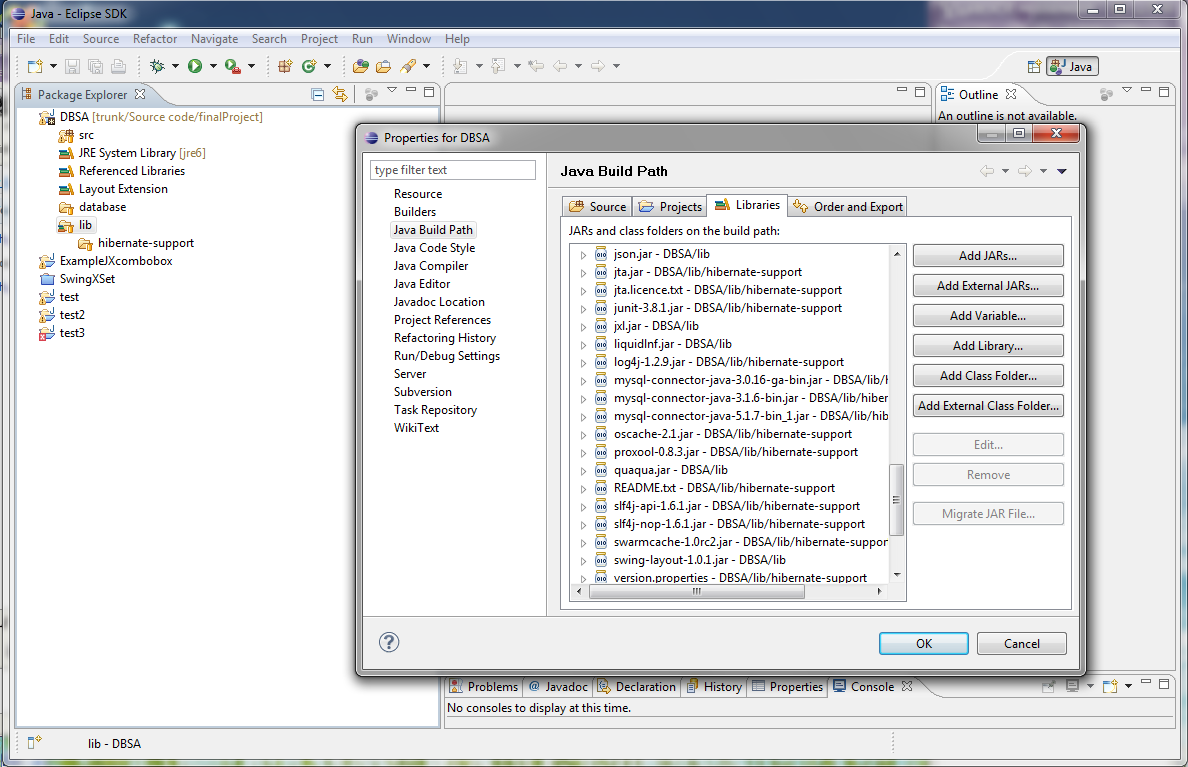
*Hình 2 : Restore dbsaDatabase dbsa có trong dvd vào database dbsa vừa tạo*

*Lưu ý:* vì file script chứ một số lượng lớn dữ liệu do đó thời gian restore khoảng 15 đến 20 máy tùy thuộc vào cấu hình máy.

**Bước 3:** Compile source code chương trình.

Sử dụng Eclipse mở project trong source.

*Lưu ý*: Cần add các thư viện cần thiết nếu như compiler báo thiếu thư viện, các thư viện có trong thư mục lib của source code/

*Hình 3: Add thư viện cần thiết cho chương trình.*

**Bước 4:** Chạy chương trình

Thực hiên chạy chương trình, trước khi chạy nên check kết nối của hệ thống với Internet và kết nối với database (đã có chức năng trong chương trình).

# Hướng dẫn tạo Database cho hệ thống.

## Các bước tạo database cho chương trình:

Đối với hệ thống chạy lần đầu chưa có CSDL, việc cài đặt cơ sở dữ liệu bao gồm việc import dữ liệu DBLP bằng tay và thêm các bảng của hệ thống bằng script SQL đi theo của chương trình.Sau đây chúng tôi xin giới thiệu cách cài đặt cơ sở dữ liệu trong trường hợp này, đối với trường hợp đã có CSDL sẵn thì việc import CSDL đơn giản là việc restore CSDL vào database.

**Thông số hệ thống:**

* Các phần mềm: Mysql Server
* Tên database : dbsa
* Hệ quản trị cơ sở dữ liệu: MySQL
* Tên truy cập : root
* Mật khẩu : root

**Bước 1:**

* Tải cơ sở dữ liệu mới của DBLP tại địa chỉ: <http://dblp.l3s.de/dblp++.php>
* Cài đặt đầy đủ các phần mềm môi trường và tương tác: bộ MySQL, Java…

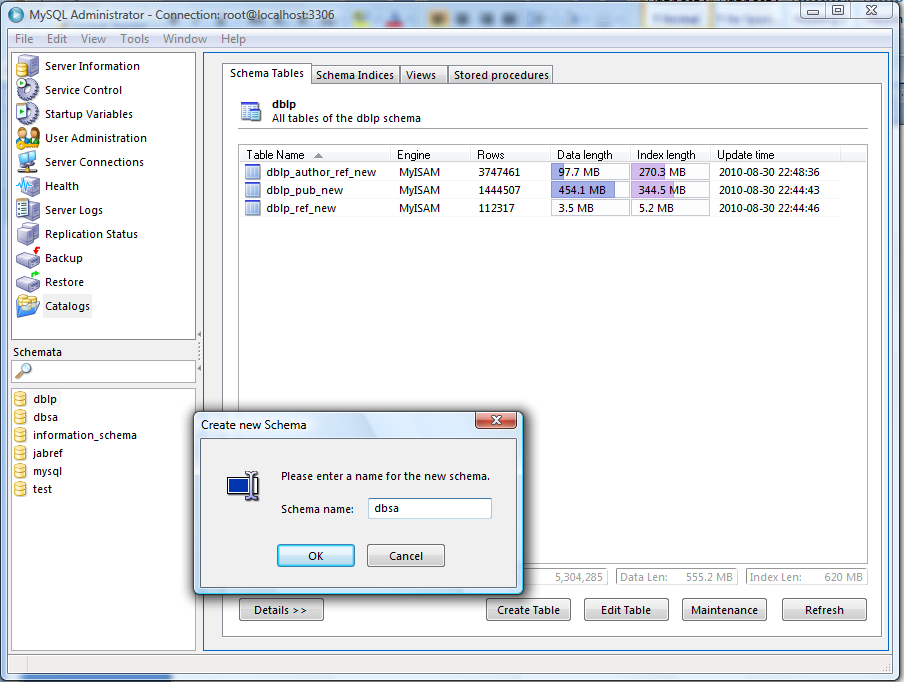
**Bước 2:**

* Kiểm tra cơ sở dữ liệu dbsa đã tồn tại hay chưa.

+Nếu đã tồn tại thì xóa bảng cũ sau đó tạo lại database mới.

+Nếu chưa có thì tạo database mới có tên là ‘dbsa’.

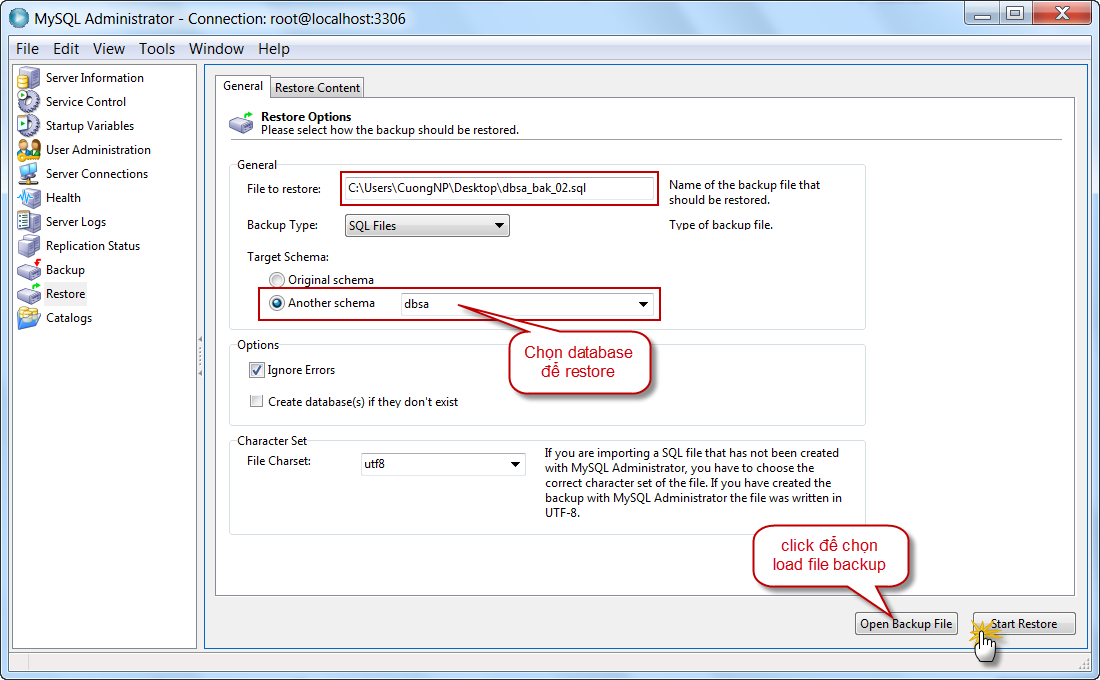
Tạo database mới: Mở MySQL Administrator 🡪 Catalogs🡪 Create New Schema (hoặc Ctrl+N) 🡪 Nhập tên database, như hình 1.



*Hình 1- Tạo database “DBSA” trong MySQL*

**Bước 3**: Restore lại database dblp từ file script vào database dbsa.

Mở MySQL Administrator 🡪 Open Backup file (Chọn file cơ sở dữ liệu DBLP mới vừa tải về). 🡪 Nhập và chọn các thông số như hình 2 🡪 Start restore.



*Hình 2 – Restore database ‘dbsa’ từ tập tin dblp\_databse.sql*

**Bước 4:** Thêm các bảng mới vào cơ sở dữ liệu “dbsa”

Sau khi đã tạo được database “dbsa” trong cơ sở dữ liệu MySQL, tiếp theo chúng ta tiến hành chỉnh sửa database cho phù hợp với hệ thống. Thêm 3 bảng mới vào cơ sở dữ liệu: **dbsa\_pub, dbsa\_pub\_in\_dblp**, **dbsa\_sbj**. Sửa tên cột **id** trong bảng **dblp\_author\_ref\_new** thành **pub\_id.**

* Thêm bảng **dbsa\_pub:**

*CREATE TABLE `dbsa\_pub` (*

*`id` int(8) NOT NULL AUTO\_INCREMENT COMMENT 'Id cua bai bao duoc thu thap ve tu he thong DBSA',*

*`sbj\_id` int(8) DEFAULT NULL COMMENT 'Id cua tua de bai bao sau khi phan loai',*

*`astract` longtext COMMENT 'Tom tat cua bai bao',*

*`title` longtext COMMENT 'Tua de cua bai bao',*

*`year` int(4) unsigned DEFAULT NULL COMMENT 'Nam xuat ban cua bai bao',*

*`publisher` varchar(255) DEFAULT NULL COMMENT 'Nha xuat ban tai lieu',*

*`authors` varchar(250) DEFAULT NULL COMMENT 'Ten cac tac gia cua bai bao',*

*`links` longtext COMMENT 'Cac duong dan mo rong cua bai bao',*

*UNIQUE KEY `id` (`id`)*

*) ENGINE=InnoDB DEFAULT CHARSET=latin1*

* Thêm bảng **dbsa\_pub\_in\_dblp:**

*CREATE TABLE `dbsa\_pub\_in\_dblp` (  
  `id` int(8) DEFAULT NULL COMMENT 'id cua bai bao trong du lieu dblp',  
  `sbj\_id` int(8) DEFAULT NULL COMMENT 'id cua bang chu de bai bao',  
  `link` varchar(250) DEFAULT NULL COMMENT 'Cac lien ket mo rong cua bai bao'  
) ENGINE=InnoDB DEFAULT CHARSET=latin1*

* Thêm bảng **dbsa\_sbj:**

*CREATE TABLE `dbsa\_sbj` (  
  `id` int(8) unsigned NOT NULL AUTO\_INCREMENT COMMENT 'Id cua chu de bai bao',  
  `subj\_name` varchar(250) DEFAULT NULL COMMENT 'Ten cua chu de bai bao',  
  PRIMARY KEY (`id`)  
) ENGINE=InnoDB DEFAULT CHARSET=latin1 COMMENT='Luu thong tin chu de bai bao'*

Chú ý: Khi đã có dữ liệu thì file backup cũng sẽ có data.

* Chỉnh sửa tên comlum trong bảng **dblp\_author\_ref\_new**:

alter table dblp\_author\_ref\_new change id pub\_id int(8) unsigned;

***Hoàn tất quá trình tạo database cho chương trình.***

## Các bước backup dữ liệu của chương trình:

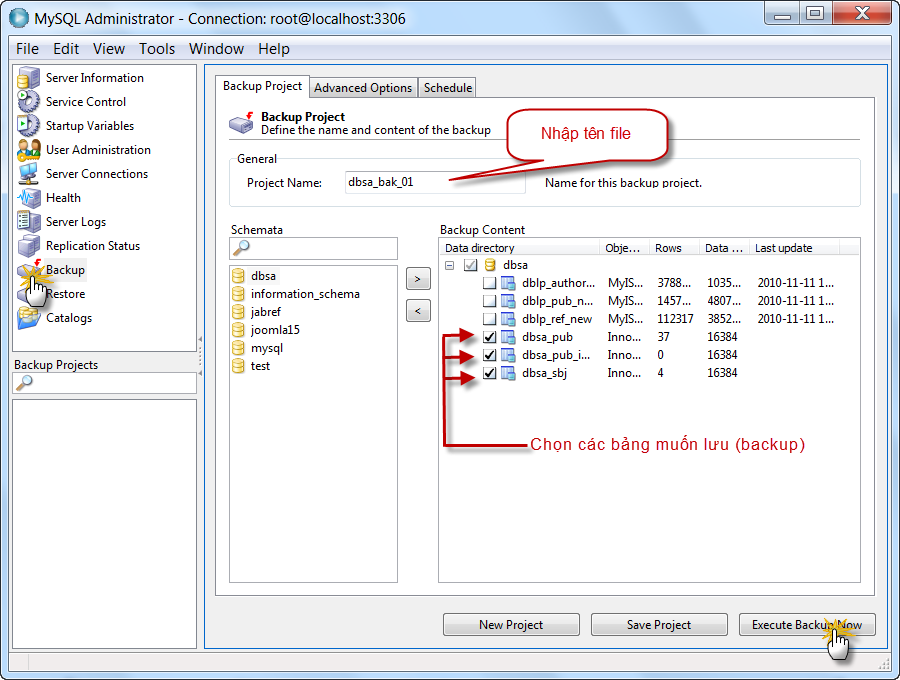
Khi muốn cập nhật database mới nhất của DBLP, việc dầu tiên là phải lưu các bảng hiện có của hệ thống đang có để tránh việc mất mát dữ liệu. Các bảng cần phải lưu (backup) là:

* **dbsa\_pub**
* **dbsa\_pub\_in\_dblp**
* **dbsa\_sbj**

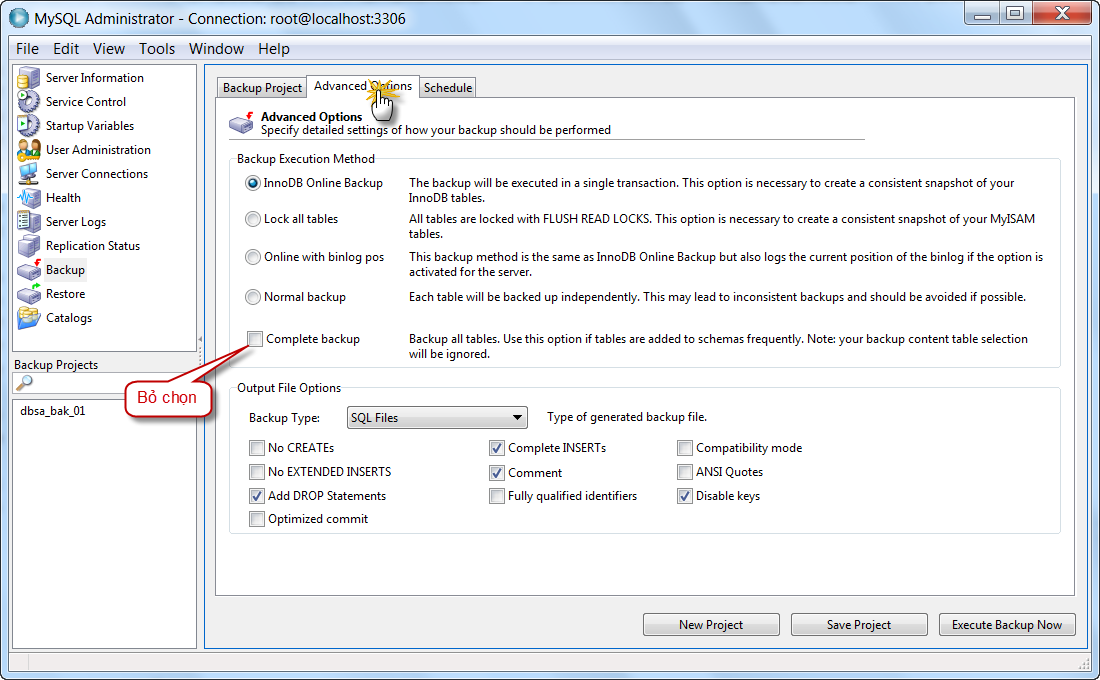
Sau đây là hướng dẫn các bước thực hiện lưu các bảng trên:

**Bước** 1:

Mở My Administrator 🡪 chọn Backup 🡪 chọn New Project 🡪 Nhập các thông số như hình 3, chọn cơ sở dữ liệu là dbsa, tên tập tin muốn lưu, và nhấn chọn 3 bảng: **dbsa\_pub, dbsa\_pub\_in\_dblp**, **dbsa\_sbj**. 🡪 Nhấn Execute backup now.



*Hình 3 – Lựa chọn backup dữ liệu (1)*



*Hình 4 – Lựa chọn backup dữ liệu (2)*

*Lưu ý*: Mặc định MySQL sẽ chọn lưu tất cả các bảng có trong database. Để lưu theo lựa chọn ở trên phải hủy chọn “complete backup” như hình 4.

Bước 3: phục hồi dữ liệu đã backup.

Đầu tiên chọn file backup từ máy đã được backup như ở trước, đúng các thông số như ở hình 2. Sau đó nhấn “Start restore” để bắt đầu quá trình phục hồi dữ liệu.

Như vậy chúng ta đã hoàn thành quá trình tạo cơ sở dữ liệu cho hệ thống xây dựng dữ liệu chỉ mục sử dụng Webcrawler.

# 4. HƯỚNG DẪN SỬ DỤNG CHƯƠNG TRÌNH.

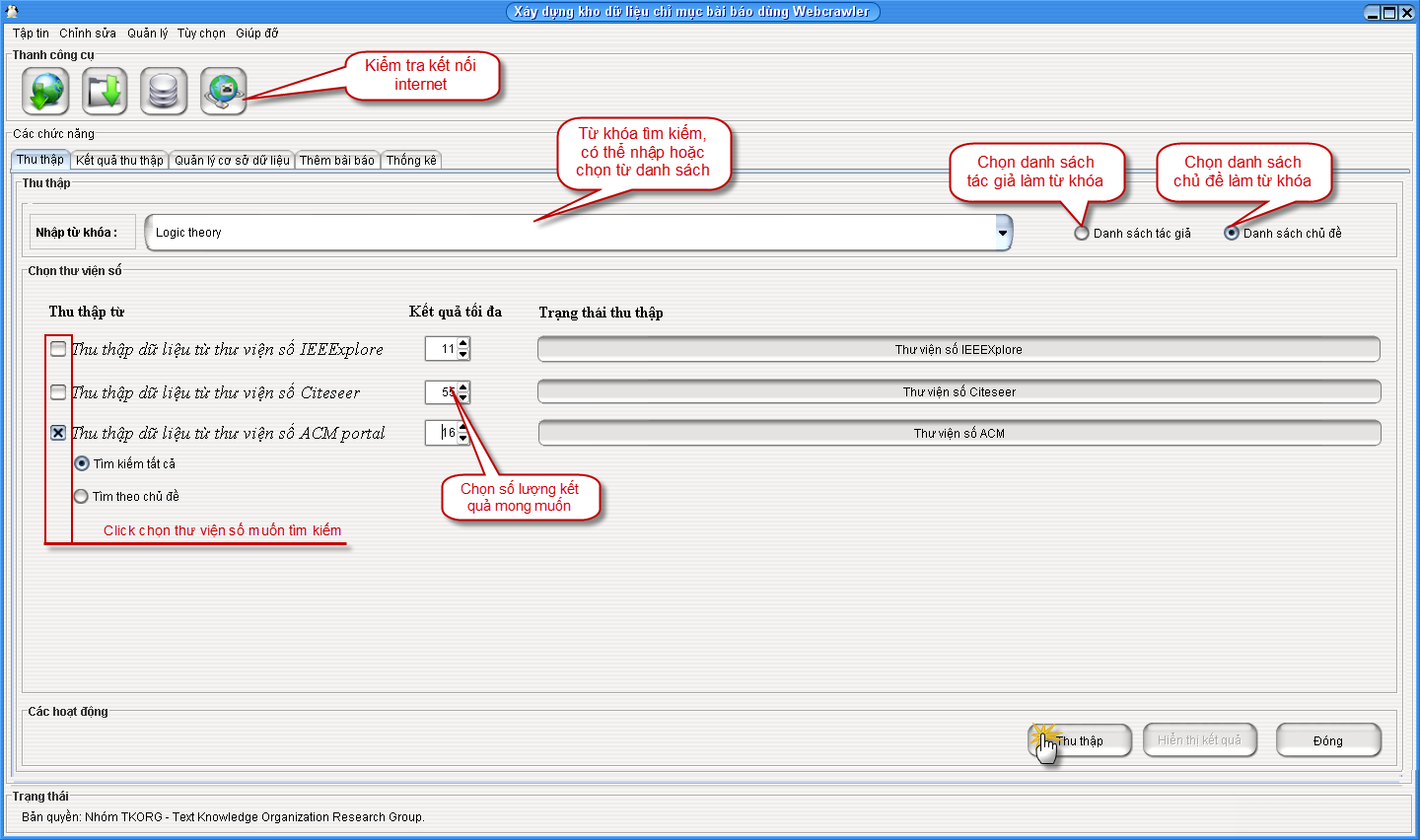
## 4.1 Giao diện chức năng thu thập thông tin Metadata từ các thư viện số.

Đây là giao diện đầu tiên khi người sử dụng khởi động chương trình.

Các thành phần trong giao diện như hình 1 gồm:

* Thanh trình đơn (1). (Menu bar)
* Thanh công cụ (2) (Tool bar)
* Các tab chức năng của chương trình (3)
* Thanh trang thái, hướng dẫn (4) (Status bar)

Các chức năng chính của hệ thống nằm trong phần các tab chức năng.

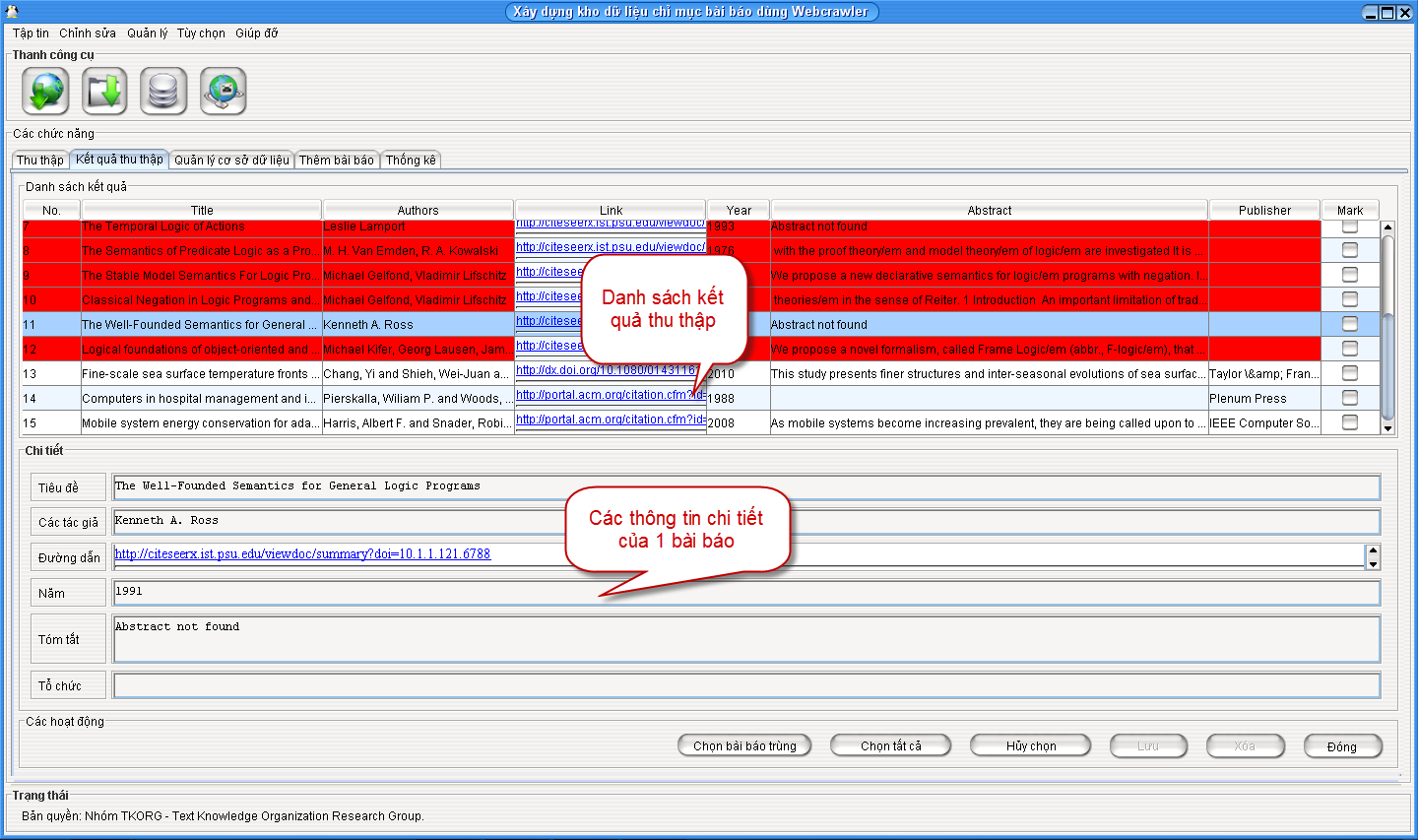


Hình 1 – Giao diện chính của hệ thống

## 4.2 Mô tả chức năng thu thập

Nhu hình 1, các thành phần và xử lý chính của chức năng này như sau:

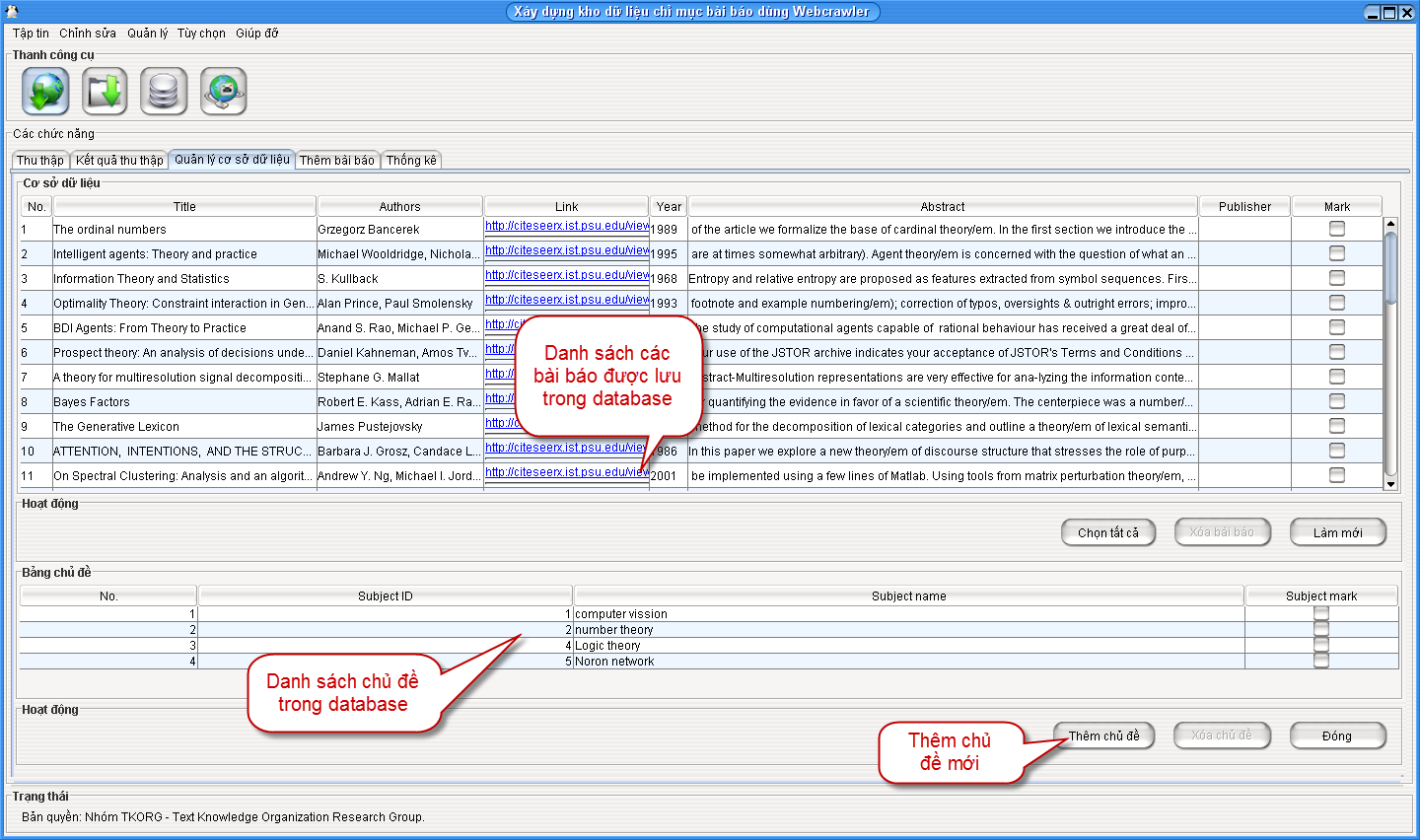
* Phần từ khóa: Có thể nhập một từ khóa mới hoặc chọn từ khóa theo gợi ý của hệ thống. Check vào nút danh sách tác giả hoặc danh sách chủ đề thay đổi danh sách từ khóa gợi ý.
* Phần lựa chọn các thông số đầu vào cho quá trình: chọn thư viện số muốn thu thập, số lượng kết quả trả về ứng với mỗi thư viện số đó.
* Phần thứ ba chứa các sự kiện nhấn nút “Thu thập” để bắt đầu quá trình thu thập, sau khi thu thập xong thì có thể nhấn nút “Hiển thị kết quả” để chuyển sang tab kết quả.



Hình 2 – Tab kết quả thu thập

* Sau khi kết thúc quá trình thu thập tab kết quả thu thập sẽ hiển thị danh sách kết quả như hình 2. Các bài báo bị tô màu là bài báo bị trùng với bài báo trong cơ sở dữ liệu. Khi chọn xem một bài báo, thông tin bài báo đó sẽ hiển thị ở phần thông tin chi tiết. người dùng có thể nhấn vào đường dẫn liên kết để truy cập trực tiếp bài báo đó trên thư viện số.
* Ở phần này người dùng có thể chọn các bài báo bằng cách check vào các nút ở cột bên phải bảng danh sách kết quả. Hoặc nhấn nút “chọn tất cả”, “chọn tất cả các bài trùng” để chọn nhanh. Sau đó người dùng có thể lưu hoặc xóa các bài đã chọn.

## 4.3 Chức năng quản lý cơ sở dữ liệu

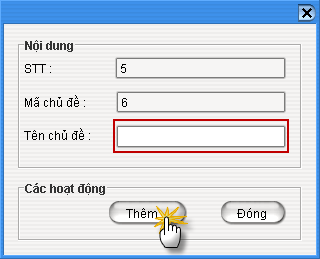


Hình 5 – Quản lý cơ sở dữ liệu

Đây là chức năng cho phép người sử dụng quản lý cơ sở dữ liệu của mình.

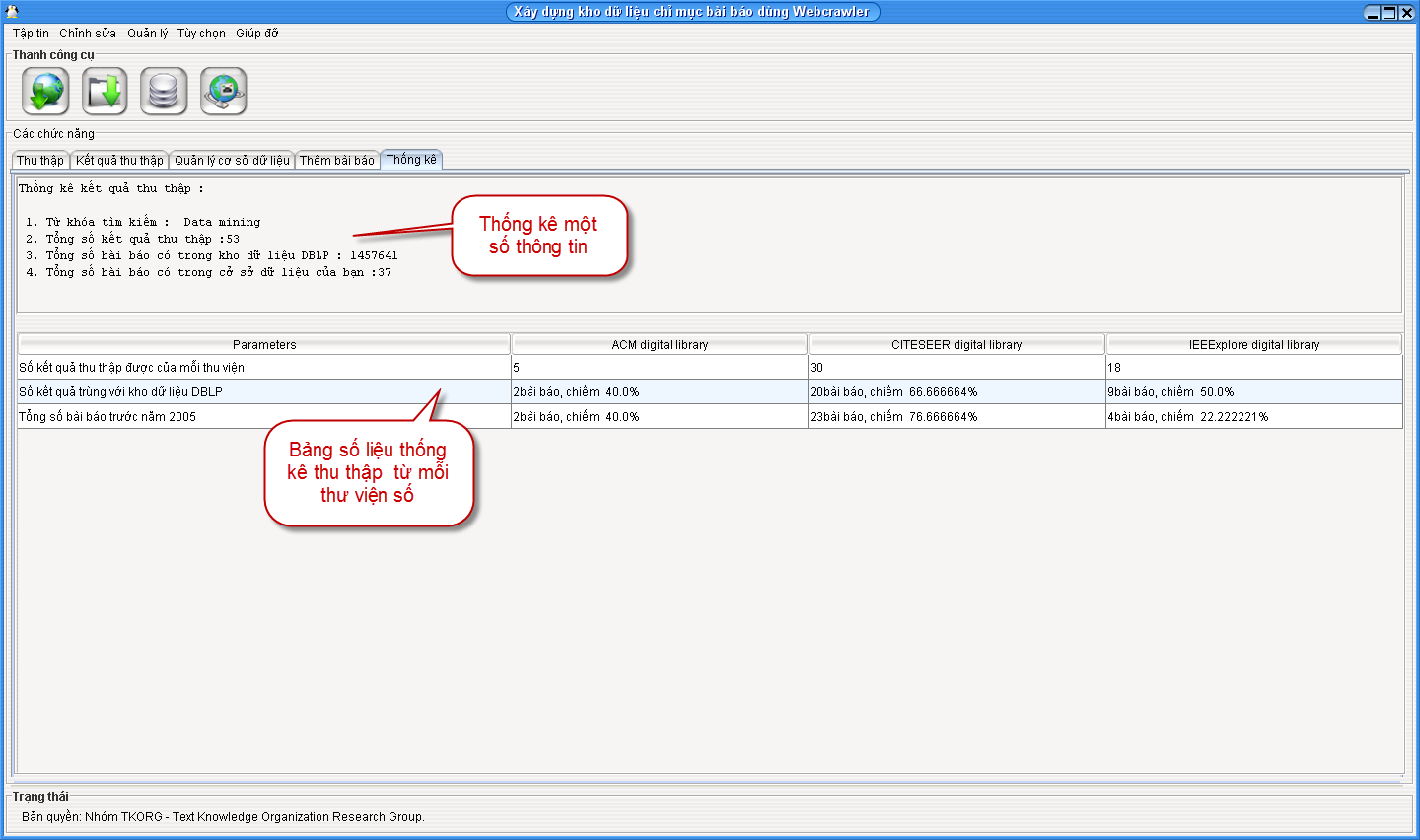
Chức năng này gồm hai thành phần chính như sau:

* Quản lý các bài báo khoa học đã được người dùng lưu trong cơ sở dữ liệu. Người dùng có thể xem danh sách bài báo và chọn xóa các bài báo theo ý muốn.
* Quản lý danh sách các chủ đề, tương tự như chức năng trên người dùng có thể xem và xóa các chủ đề theo ý muốn. Ngoài ra người dùng có thể thêm một chủ đề mới bằng cách nhấn vào nút “Thêm chủ đề”. Khi nhấn vào nút này hệ thống sẽ hiển thị lên chức năng thêm chủ đề như hình 6, sau đó người dùng nhập tên chủ đề và nhấn “Thêm mới”.



Hình 6 – Thêm chủ đề mới

## 4.4 Thông kê kết quả thu thập



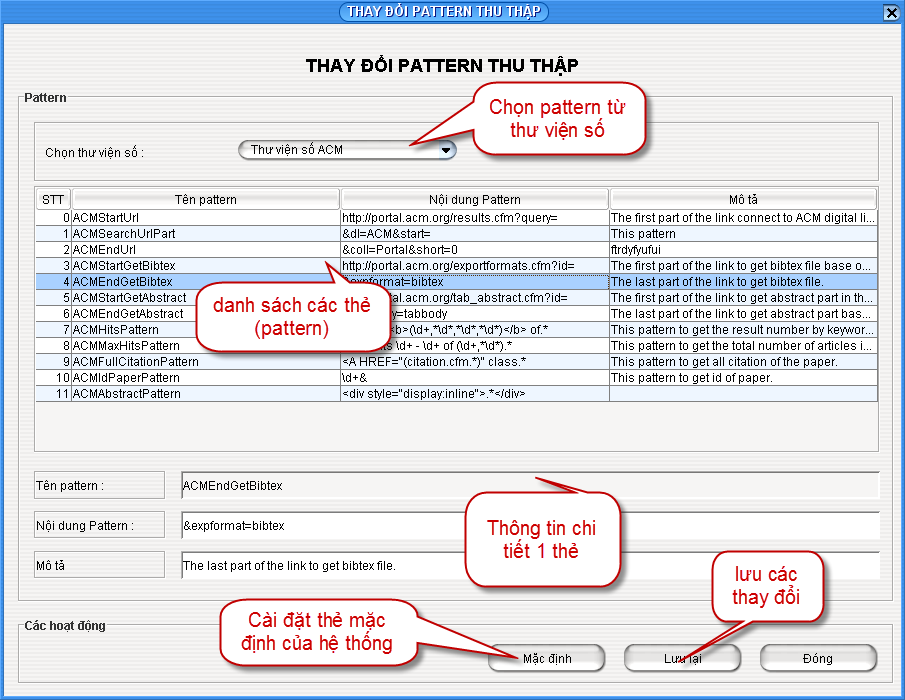
Hình 7- Bảng thống kê kết quả sau khi thu thập hoàn thành

Sau khi chức năng thu thập hoàn tất hệ thống sẽ tự động thống kê các kết quả từ mỗi thư viện số, kiểm tra, so sánh… và đưa ra kết quả như hình 7

## 4.5 Chức năng thay đổi các thẻ rút trích (Pattern)

- Chức năng đổi các thẻ (pattern). Đây là chức năng định nghĩa các thẻ để truy xuất các thư viện số và thu thập các thông tin từ các thư viện số. Người dùng có thể chọn một thẻ và thay đổi nội dung sau đó lưu lại, hoặc cài đặt các thẻ theo mặc định cùa hệ thống như hình 8 đã mô tả rõ.

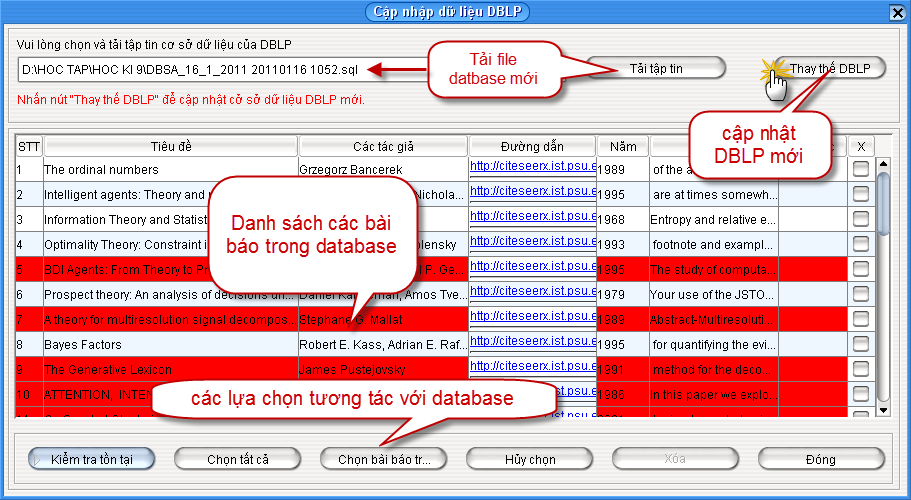
***Lưu ý:*** *khi thay đổi nội dung các thẻ sẽ dễ dẫn tới hệ thống không hoặc động tốt và làm việc không chính xác.*



Hình 8 – Chỉnh sửa thông tin các thẻ của các thư viện số(pattern)

## Chức năng cập nhật cơ sở dữ liệu mới của DBLP

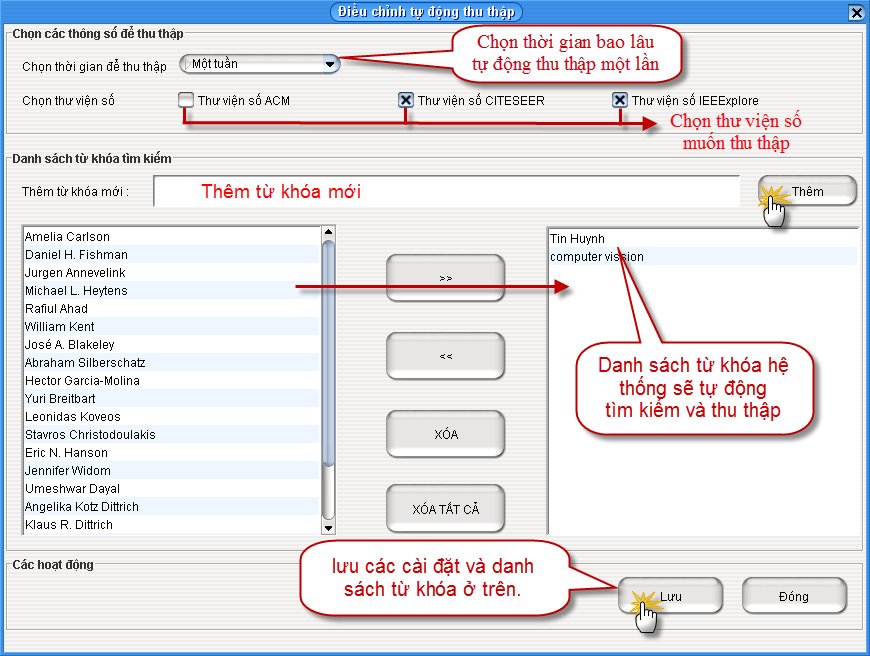
- Để thực hiện chức năng này, đầu tiên bạn phải tải về bộ cơ sở dữ liệu mới nhất của DBLP từ địa chỉ <http://dblp.l3s.de/dblp++.php>. Sau đó bạn chọn chức năng cập nhật cơ sở dữ liệu DBLP từ thanh trình đơn. Từ giao diện của hệ thống chọn load tập tin vừa tải về tiếp theo nhấn nút “Thay thế DBLP”. Hệ thống sẽ nhận sự kiện và thay thế cơ sở dữ liệu mới thay cho cơ sở dữ liệu cũ. Sau khi cập nhật xong hệ thống sẽ kiểm tra so sánh dữ liệu của DBLP mới với dữ liệu các bài báo khoa học đã được bạn lưu từ trước. Hệ thống sẽ hiển thị ra danh sách các bài báo như hình 9, những bài báo bị trùng sẽ được tô màu. Tại đây cho phép người dùng có thể chọn các bài báo và xóa theo ý muốn.



Hình 9 – Chức năng cập nhật cơ sở dữ liệu DBLP

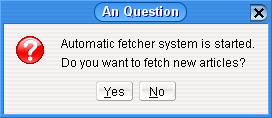
## 4.7 Chức năng tự động thu thập dữ liệu:

- Chức năng này cho phép người dùng có thể tự động thu thập dữ liệu theo định kì trên các thư viện và danh sách từ khóa do người dùng chọn như hình 10. Sau khi người dùng nhấn nút lưu thì hệ thống sẽ lưu lại tất cả các thông tin mà người dùng đã chọn. Danh sách từ khóa người dùng có thể thêm mới hoặc chọn các gợi ý cùa hệ thống ở bên cột trái.



*Hình 10 – Cài đặt thông tin cho chức năng tự động thu thập dữ liệu.*

- Sau khi cài đặt xong như ở trên, mỗi khi hệ thống được khởi động lên thì sẽ kiểm tra xem đã đến thời gian tự động thu thập hay chưa. Nếu đã đến thời gian tự động theo như cài đặt trước, hệ thống sẽ bật chức năng lên như hình-11. Xác nhận xem người dùng có muốn thu thập không, nếu có thì hệ thống sẽ tự động thu thập tất cả các bài mới, loại bỏ các bài trùng với cơ sở dữ liệu đã có và lưu xuống cơ sở dữ liệu.



Hình 11 – Xác nhận người dùng có muốn tự động thu thập hay không