**BÁO CÁO CRAWLER**

# Giới thiệu Crawler

Crawler là chương trình tự động tìm kiếm tài nguyên trên internet như trang web , hình ảnh , tài liệu word , pdf hay Postscripts, cho phép máy tìm kiếm đánh chỉ số sau đó. Quá trình thực hiện được gọi là Web crawling hay spidering, rất nhiều các công cụ tìm kiếm trên thế giới sử dụng spidering để cập nhật kho dữ liệu website của mình. Chẳng hạn như google dùng nó để lấy các thông tin trên các website ồi cập nhật vào cơ sở dữ liệu của nó giúp người dùng search ra các trang theo ý muốn. Web crawler là 1 loại của cái gọi là **bot** (là 1 tác tử thực hiện tự động, đại loại nghĩa giống như spider). Đại loại, nó bắt đầu từ danh sách các địa chỉ URL được gọi là seeds (hạt giống). Nó sẽ vào các địa chỉ này lọc thông tin rồi tìm ra các địa chỉ URL khác thêm chúng vào danh sách các địa chỉ đã duyệt qua gọi là **crawl frontier**. Sau nó nó lại lặp lại quá trình đó duyệt qua những URL mới. Cứ thế, cứ thể nó lần qua rất nhiều địa chỉ website và thu thập rất nhiều nội dung khác nhau giúp ta dễ dàng tìm kiếm thông tin mình cần.

Một số bọ tìm kiếm phổ biến của các máy tìm kiếm :

* Googlebot của [Google](http://www.google.com/)
* MSNBot của [MSN](http://www.msn.com)
* Slurp của [Yaho](http://www.yahoo.com)o
* Scooter của Alta Vista
* Baidu của Baidu

Vì vậy điều cần thiết đầu tiên để tạo nên một cỗ máy tìm kiếm tuyệt hảo đó là có một crawler thông minh và khả năng cập nhật dữ liệu nhanh và tốt trên một khối lượng dự trữ cực lớn.

Sauk hi tìm hiểu một số crawler mã nguồn mở có trên mạng tìm được như JSpider, Crawler của Nutch . Yêu cầu hiện tại để thực hiện luận văn là tìm và sử dụng đưowjc crawler có khả năng download các file html từ các website nhạc để lấy được các link media trong file html dùng để indexing cho Search Enginer . Vì crawler là một trong những yếu tố cực kỳ quan trọng trong chiến lược tìm kiếm của các máy tìm kiếm danh tiếng như Google, yahoo , Baidu… vì vậy khả năng dùng được crawler của các cỗ máy tìm kiếm trên là không thể . Hiện nay có rất nhiều phần mềm về crawler được bán có bản quyền bên cạnh đó các phần mềm mã nguồn mở cũng rất phong phú đa dạng được giới thiệu như **[Aspseek](http://en.wikipedia.org/wiki/Aspseek" \o "Aspseek)** , **[arachnode.net](http://en.wikipedia.org/wiki/Arachnode.net" \o "Arachnode.net)** ,**[DataparkSearch](http://en.wikipedia.org/wiki/DataparkSearch" \o "DataparkSearch)** ,**[GNU Wget](http://en.wikipedia.org/wiki/Wget" \o "Wget),** [**GRUB**](http://en.wikipedia.org/wiki/Grub_%28search_engine%29) **,** [**Heritrix**](http://en.wikipedia.org/wiki/Heritrix) , [**ht://Dig**](http://en.wikipedia.org/wiki/Ht-/dig) ,[**HTTrack**](http://en.wikipedia.org/wiki/HTTrack) ,[**ICDL Crawler**](http://en.wikipedia.org/wiki/ICDL_crawling) ,[**mnoGoSearch**](http://en.wikipedia.org/wiki/MnoGoSearch) ,[**Nutch**](http://en.wikipedia.org/wiki/Nutch) ,[**Pavuk**](http://en.wikipedia.org/wiki/Pavuk) ,[**YaCy**](http://en.wikipedia.org/wiki/YaCy) **…**

# Crawler được sử dụng

Yêu cầu được đặt ra cho crawler được sử dụng trong luận văn là có khả năng tải được hầu hết các trang của một site media trên internet về máy tính cục bộ dưới dạng file html , có khả năng resum khi bị ngắt kết nối download giữa chừng, có chức năng redownload khi dữ liệu trên trang web bị thay đổi, có chức năng hẹn giờ tự động cập nhật lại dữ liệu . Jspider sau khi tìm hiểu các ví dụ thì không thấy khả quan trong quá trình tìm kiếm và download các html từ các trang web media như bamboo.com , nhac8.com , nhaccuatui.com, www3.nhac.vui.vn, mp3.zing.vn, vnmusic.com, … Mã nguồn mở Nutch tìm hiểu thì thấy rất khó để tách riêng dữ liệu được tải về làm dữ liệu để làm đầu vào để indexing riêng , nhất nhiều phần mềm copy website được sử dụng nhưng cũng không đáp ứng được yêu cầu trên . Ngoài ra còn khảo sát các chương trình khác như Webcopier, [WebRipper\_1.32 free ,](http://my.opera.com/hieu_bf/blog/?tag=Ti%E1%BB%87n%20%C3%8Dch%20M%E1%BA%A1ng&startidx=20&nodaylimit=1) [WebsiteWatcher4.42](http://www.vn-zoom.com/f99/website-watcher-v4-41-a-43137.html), WebDataExtractor\_wde , Mã nguồn mở HTTrack đáp ứng được những yêu cầu đặt ra. HTTrack có khả năng tải website về máy local như một phiên bản thứ 2 với một số trang web được yêu cầu. *<http://en.wikipedia.org/wiki/HTTrack> HTTrack được giới thiệu trên wiki ,* <http://www.httrack.com/> website giới thiệu về mã nguồn .

* Các trang web nhạc đã được duyệt như :

<http://www.nhac8.com/>, vnmusic.com.vn , nhaccuatui.com, www3.nhac.vui.vn , mp3.zing.vn, vnmusic.com, [www.yeucahat.com/](http://www.yeucahat.com/),[clip.vn](http://www.clip.vn), [vtc.com.vn](http://www.vtc.com.vn), [sonic.vn](http://sonic.vn/front/music/), [nghenhac.tv](http://www.nghenhac.tv), [nhacyeu.info](http://www.nhacyeu.info), [nhacnen.net](http://Nhacnen.net), [tuoitho.net](http://tuoitho.net), [enhac.net](http://enhac.net), , [music.vietvoice.net](http://music.vietvoice.net" \t "_blank), [cuasoamnhac.net](http://cuasoamnhac.net), [hoathachthao.info](http://hoathachthao.info), [tialia.net](http://tialia.net), [notnhactre.net](http://notnhactre.net), [nhacviet.vietnamnet.vn](http://nhacviet.vietnamnet.vn), [nghenhacso.net](http://www.nghenhacso.net), [tinhcaviet.com](http://tinhcaviet.com), [Woim.net](http://www.woim.net), [Khuctutinh.net](http://khuctutinh.net), [Loitraitim.com](http://loitraitim.com), [24h.com.vn](http://www.24h.com.vn), [Vntube.com](http://Vntube.com), [Yeuamnhac.com](http://Yeuamnhac.com) , [Theoyeucau.com](http://Theoyeucau.com), [Nghenhac.info](http://nghenhac.info), <http://music.vietfun.com> <http://www.music4u.info>.

* Các chương trình crawler đã từng kiểm ta và sử dụng:

Webcopier, [WebRipper\_1.32 free ,](http://my.opera.com/hieu_bf/blog/?tag=Ti%E1%BB%87n%20%C3%8Dch%20M%E1%BA%A1ng&startidx=20&nodaylimit=1) [WebsiteWatcher4.42](http://www.vn-zoom.com/f99/website-watcher-v4-41-a-43137.html), WebDataExtractor\_wde , Nutch, Jspider, telepot.pro.

* Các website đã tải về và lấy được link trong file html khi dùng chương trình HTTrack :

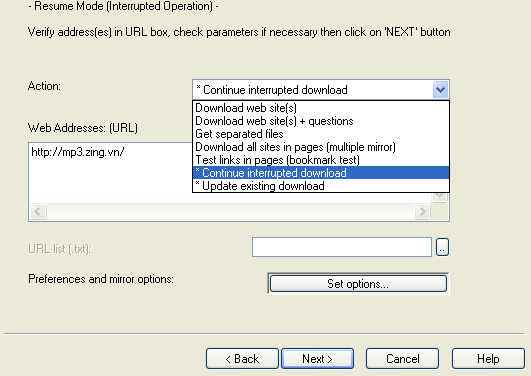
[www.mp3.zing.vn](http://www.mp3.zing.vn) , <http://music.yeucahat.com> , <http://vnmusic.com.vn> , , <http://hoang-huy.net> , …

Đặc điểm chung của hầu hết các website media đó là đều dùng javascript để mã hóa link download , ngăn chặn các địa chỉ download liên tục có nhiều kết nối khi dùng các chương trình crawler download website vì vậy các website tìm được thỏa mãn không nhiều. Tại thời điểm hiện tại viết bài khảo sát này trang <http://baamboo.com> cũng chỉ crawler về được một vài trang web nhất định ví dụ như : bennhac.com. , benxua.com , sử dụng trang tìm kiếm trung gian [http://mp3.timmau.com/](http://mp3.timmau.com/#Play,15679), chacha.vn, nhaccuatui.com, [www.nhac1000.com](http://www.nhac1000.com) , <http://www.canhac.com/> điều này chứng tỏ còn rất nhiều trang web media ở Việt Nam bamboo chưa truy cập và lấy dữ liệu được.

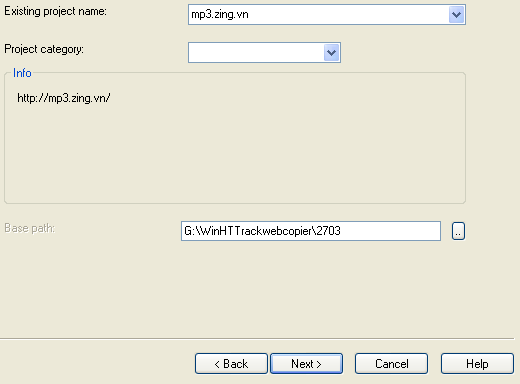
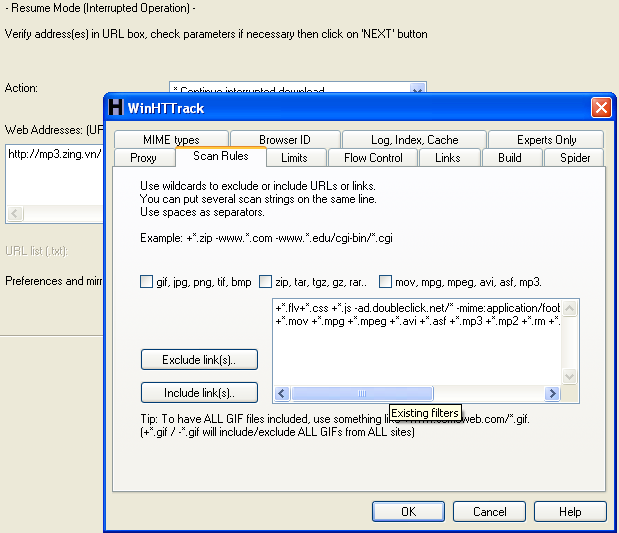
# Hướng dẫn sử dụng

Ở lần chạy đầu tiên, HTTrack sẽ yêu cầu thiết lập Proxy và ngôn ngữ chính thức sử dụng trong chương trình. Nên OK để đồng ý với những hiệu chỉnh mặc định của nó và bắt đầu quá trình tải website mong muốn của mình theo 3 bước như sau:   
- Bước 1: Tạo project mới   
Để thuận tiện hơn cho người sử dụng, HTTrack sẽ lưu tất cả các thông số về một quá trình tải website cần tải, nơi lưu site trên máy tính, scan rules... vào một file project có dạng \*.whtt. Vì vậy nên dễ dàng quản lý nhiều tác vụ tải site khác nhau, mở lại hay nâng cấp website đã tải..   
Trên cửa sổ của chương trình, bấm Next và nhập tên project vào khung Project name. Sau đó, chọn nơi cần lưu website trên máy bằng cách bấm vào nút hình dấu cộng (+). Lưu ý, nên đặt tên project theo địa chỉ website cần tải, chẳng hạn như [mp3.zing.com](http://mp3.zing.com) ([www.mp3.zing.com](http://www.mp3.zing.com)) để tiện phân biệt với các project khác sau này.

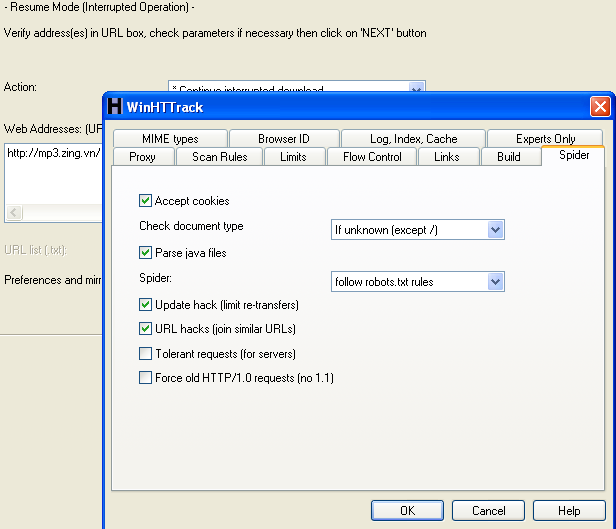
Ở mục Action với những project đầu tiên được thiết lập thì dùng một trong năm yêu cầu ban đầu :

* Download web site(s)
* Download web site(s) + questions
* Get separated file
* Download all sites in pages ()
* Test links in pages(bookmark test) 

Next để tiếp tục qua bước 2.   
Bước 2: Thiết lập các thông số cần thiết   
Đây là bước quan trọng nhất trong toàn bộ thao tác vì nó quyết định website tải về sẽ có những gì và cách thức tải như thế nào.   
Đầu tiên, nhập vào địa chỉ website ở khung Web Addresses (URL), ví dụ như [www.mp3.zing.com](http://www.mp3.zing.com) , hoặc có thể thêm các link khác nếu muốn.Nếu website cần tải có yêu cầu đăng nhập (Login), bấm nút Add URL và điền vào các thông số cần thiết như mật mã và tên tài khoản.

  
Sau khi đã hoàn tất phần địa chỉ website cần tải, bạn hãy bấm tiếp vào nút Set options để có thêm một số hiệu chỉnh khác. Như đã nói ở trên, HTTrack cho phép thiết lập bộ lọc giúp xác định các dạng file cần tải. Để thực hiện điều này thì trên hộp thoại vừa hiện ra, bạn chuyển qua thẻ Scan Rules, sau đó bấm chọn vào 3 dạng file cho sẵn của chương trình là gif,jpg,png... nếu muốn tải các dạng file hình ảnh; zip,tar,tgz... với dạng file nén và mov,mpg,mpeg... đối với các file phim và nhạc số. Lưu ý là có thể thiết lập thêm các dạng file khác bằng cách nhập trực tiếp vào khung theo công thức sau:   
+\*.dạng file để thêm vào dạng file cần tải mới và -\*.dạng file để ngăn không cho HTTrack tải dạng file này. Ví dụ như công thức +\*.pdf -\*.txt sẽ có nghĩa là lấy các file pdf và bỏ các file txt. 

Nếu quan tâm đến dạng liên kết, cấu trúc của website... có thể tùy chọn trong các hiệu chỉnh nâng cao khác cũng trên hộp thoại này của chương trình. Khi đã vừa ý, hãy bấm OK, sau đó bấm Next và Finish để hoàn tất bước 2 và cũng để bắt đầu quá trình tải .

Trong phần hiệu chỉnh option có phần spider để quét các trang web có liên kết với website hiện hành. Phần này mục tiêu đặt ra không yêu cầu đặt lên đầu , vì với mức độ của luận văn , chi phí và thời gian không đáp ứng được nhiều, vì vậy chỉ download trực tiếp các website yêu cầu.  
Bước 3: Các thao tác căn bản với website đã tải   
Nếu mọi chuyện suôn sẻ, sẽ có được website mong muốn trên máy tính. Cần biết là không phải website nào cũng cho phép bạn tải xuống vì những lý do bảo mật. Để duyệt website vừa tải, hãy mở thư mục đã lưu site và bấm đúp chuột vào file index.html.   
HTTrack còn có 2 thao tác rất hữu dụng ở hộp danh sách Action trong bước 2 mà bạn nên tìm hiểu là:   
-Continue interrupted download: Tiếp tục lại một quá trình tải site nếu gặp trục trặc phải ngưng giữa chừng.   
- Update esixting download: Đồng bộ hóa website trên máy tính với website tương ứng trên Internet.