

5 Ways to Crawl a Website

 hackingarticles.in/5-ways-crawl-website

Raj

July 16, 2017

```
msf > use auxiliary/crawler/msfcrawler
msf auxiliary(msfcrawler) > set rhosts www.tptl.in
rhosts => www.tptl.in
msf auxiliary(msfcrawler) > set threads 10
threads => 10
msf auxiliary(msfcrawler) > exploit

[*] Loading modules: /usr/share/metasploit-framework/data/msfcrawler
[*] Loaded crawler module Simple from /usr/share/metasploit-framework
[*] Loaded crawler module Comments from /usr/share/metasploit-framework
[*] Loaded crawler module Forms from /usr/share/metasploit-framework
[*] Loaded crawler module Frames from /usr/share/metasploit-framework
[*] Loaded crawler module Image from /usr/share/metasploit-framework
[*] Loaded crawler module Link from /usr/share/metasploit-framework
[*] Loaded crawler module Objects from /usr/share/metasploit-framework
[*] Loaded crawler module Scripts from /usr/share/metasploit-framework
[*] OK
[*] URI LIMITS ENABLED: 10 (Maximum number of requests per uri)
[*] Target: www.tptl.in Port: 80 Path: / SSL:
[*] >> [200] /
[*] ERROR
[*] ERROR
[*] >> [200] /aboutus.php
[*] ERROR
[*] >> [200] /services.php
[*] ERROR
[*] >> [200] /projects.php
[*] ERROR
[*] >> [200] /clients.php
[*] ERROR
[*] >> [200] /contactus.php
[*] ERROR
[*] >> [404] /aboutus.html
[*] [404] Invalid link /aboutus.html
[*] >> [200] /dataplan.php
[*] ERROR
[*] >> [200] /videoplan.php
```

A Web crawler, sometimes called a spider, is an Internet bot that systematically browses the World Wide Web, typically for the purpose of Web indexing.

A Web crawler starts with a list of URLs to visit, called the seeds. As the crawler visits these URLs, it identifies all the hyperlinks in the page and adds them to the list of URLs to visit. If the crawler is performing archiving of websites it copies and saves the information as it goes. The archive is known as the repository and is designed to store and manage the collection of web pages. A repository is similar to any other system that stores data, like a modern-day database.

Let's Begin!!

Metasploit

This auxiliary module is a modular web crawler, to be used in conjunction with wmap (someday) or standalone.

use auxiliary/crawler/msfcrawler

msf auxiliary(msfcrawler)> set rhosts www.example.com

msf auxiliary(msfcrawler)> exploit

From, the screenshot you can see it has loaded crawler in order to exact hidden file from any website, for example, about.php, jquery contact form, html and etc which is not possible to exact manually from the website using the browser. For information gathering of any website, we can use it.

```
msf > use auxiliary/crawler/msfcrawler
msf auxiliary(msfcrawler) > set rhosts www.tptl.in
rhosts => www.tptl.in
msf auxiliary(msfcrawler) > set threads 10
threads => 10
msf auxiliary(msfcrawler) > exploit

[*] Loading modules: /usr/share/metasploit-framework/data/msfcrawler
[*] Loaded crawler module Simple from /usr/share/metasploit-framework
[*] Loaded crawler module Comments from /usr/share/metasploit-framework
[*] Loaded crawler module Forms from /usr/share/metasploit-framework
[*] Loaded crawler module Frames from /usr/share/metasploit-framework
[*] Loaded crawler module Image from /usr/share/metasploit-framework
[*] Loaded crawler module Link from /usr/share/metasploit-framework
[*] Loaded crawler module Objects from /usr/share/metasploit-framework
[*] Loaded crawler module Scripts from /usr/share/metasploit-framework
[*] OK
[*] URI LIMITS ENABLED: 10 (Maximum number of requests per uri)
[*] Target: www.tptl.in Port: 80 Path: / SSL:
[*] >> [200] /
[*] ERROR
[*] ERROR
[*] >> [200] /aboutus.php
[*] ERROR
[*] >> [200] /services.php
[*] ERROR
[*] >> [200] /projects.php
[*] ERROR
[*] >> [200] /clients.php
[*] ERROR
[*] >> [200] /contactus.php
[*] ERROR
[*] >> [404] /aboutus.html
[*] [404] Invalid link /aboutus.html
[*] >> [200] /dataplan.php
[*] ERROR
[*] >> [200] /videoplan.php
```

Httrack

HTTrack is a free and open source Web crawler and offline browser, developed by Xavier Roche

It allows you to download a World Wide Web site from the Internet to a local directory, building recursively all directories, getting HTML, images, and other files from the server to your computer. HTTrack arranges the original site's relative link-structure.

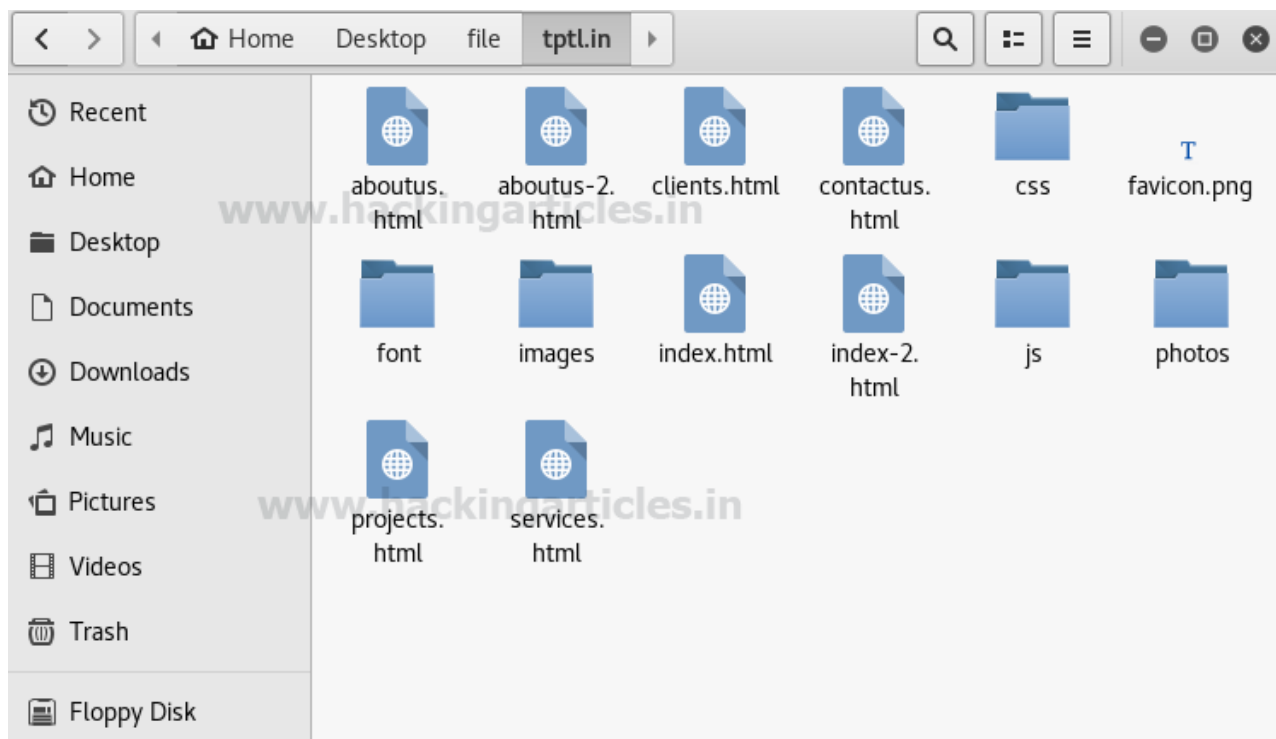
Type following command inside the terminal

```
httrack http://tptl.in -O /root/Desktop/file
```

It will save the output inside given directory /root/Desktop/file

```
root@kali:~# httrack http://tptl.in -O /root/Desktop/file
WARNING! You are running this program as root!
It might be a good idea to run as a different user
Mirror launched on Fri, 14 Jul 2017 05:48:49 by HTTrack Website Copier/3.49-2 [XR&CO'2014]
mirroring http://tptl.in with the wizard help..
^Ctptl.in/images/blockquote.html (510 bytes) - OK
** Finishing pending transfers.. press again ^C to quit.
Done.: tptl.in/images/carousel/right-arrow.png (0 bytes) - -1
Thanks for using HTTrack!
```

From given screenshot you can observe this, it has dumb the website information inside it which consist html file as well as JavaScript and jquery.



Black Widow

This Web spider utility detects and displays detailed information for a user-selected Web page, and it offers other Web page tools.

BlackWidow's clean, logically tabbed interface is simple enough for intermediate users to follow but offers just enough under the hood to satisfy advanced users. Simply enter your URL of choice and press Go. BlackWidow uses multi-threading to quickly download all files and test the links. The operation takes only a few minutes for small Web sites.

You can download it from [here](#).

Enter your URL **http://tptl.in** in Address field and press **Go**.



Click on **start** button given on the left side to begin URL scanning and select a folder to save the output file.

From the screenshot, you can observe that I had browse C:\Users\RAJ\Desktop\tptl in order to store output file inside it.

BlackWidow - http://tptl.in/

Browser | Filters | Scanner | Link Errors | Emails | Ext Links | Structure | NetSpy | About

Scanner settings and status

Before you begin the scan, review the settings at the bottom of this page. You can select to download as you scan, or, wait until the scan is done and select (from the Structure) what you need to download. Keep in mind that scanning with more than 6 connections may result in connection errors as most servers only allow this much. When you start the scan, the list below will show you the status for each scan connections.

Status	URL
Receiving Response	http://tptl.in/

New Scan

Start Scan

Pause Scan

Stop Scan

Start the scan at the URL defined in the 'Browser' section.

Filters and/or Expert script will be used to control the scan.

www.hackingarticles.in

Download while scanning

Select a download folder:

C:\Users\RAJ\Desktop\tptl

Preserve directory structure

Ignore duplicates (do not save)

Maximum connections: 6

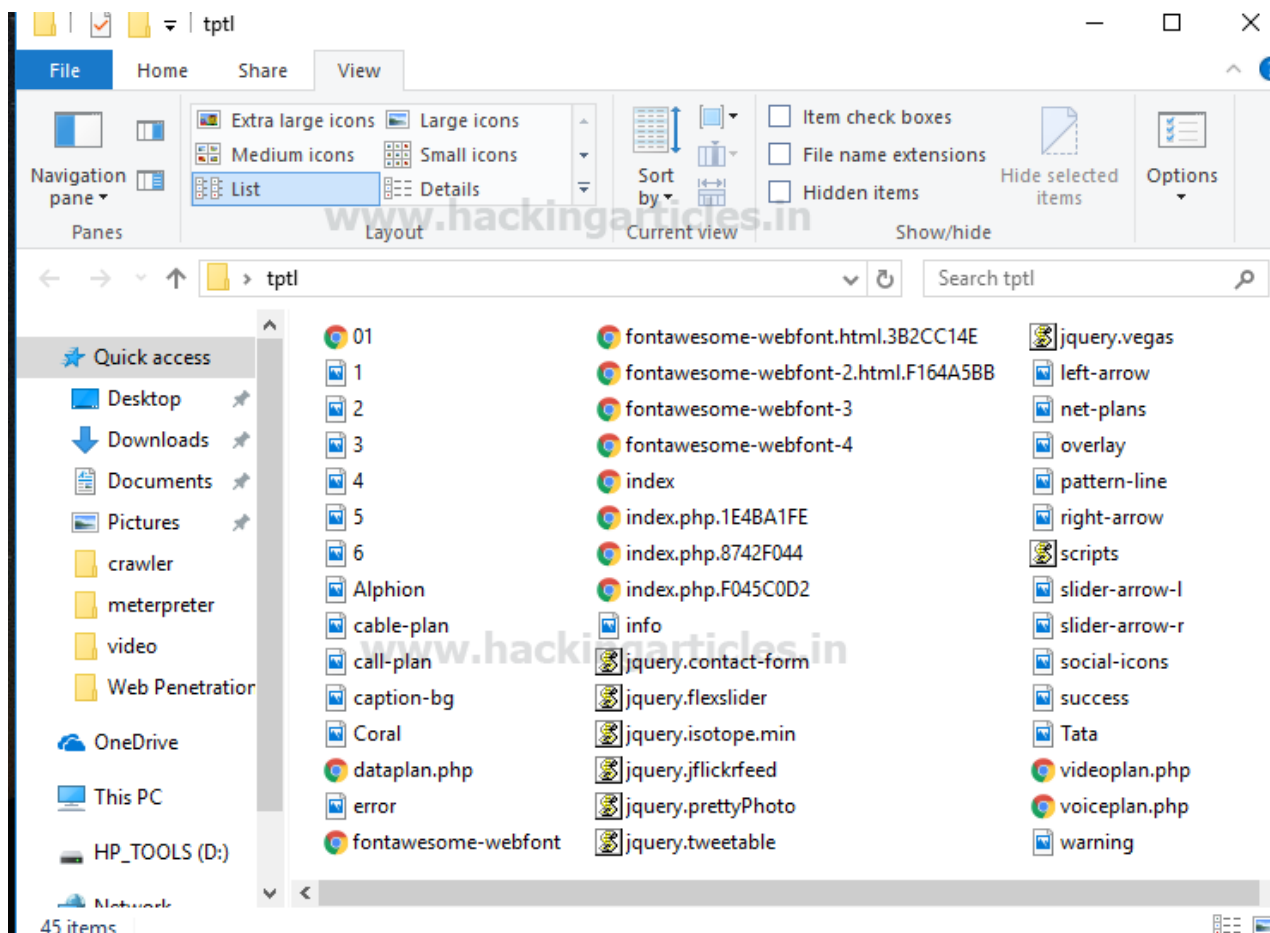
Slowdown the scan

Minimum (sec): 10

Maximum (sec): 60

Done: 00:00:00 | Scanned: 74 | Added: 66

When you will open target folder tptl you will get entire data of website either image or content, html file, php file, and JavaScript all are saved in it.



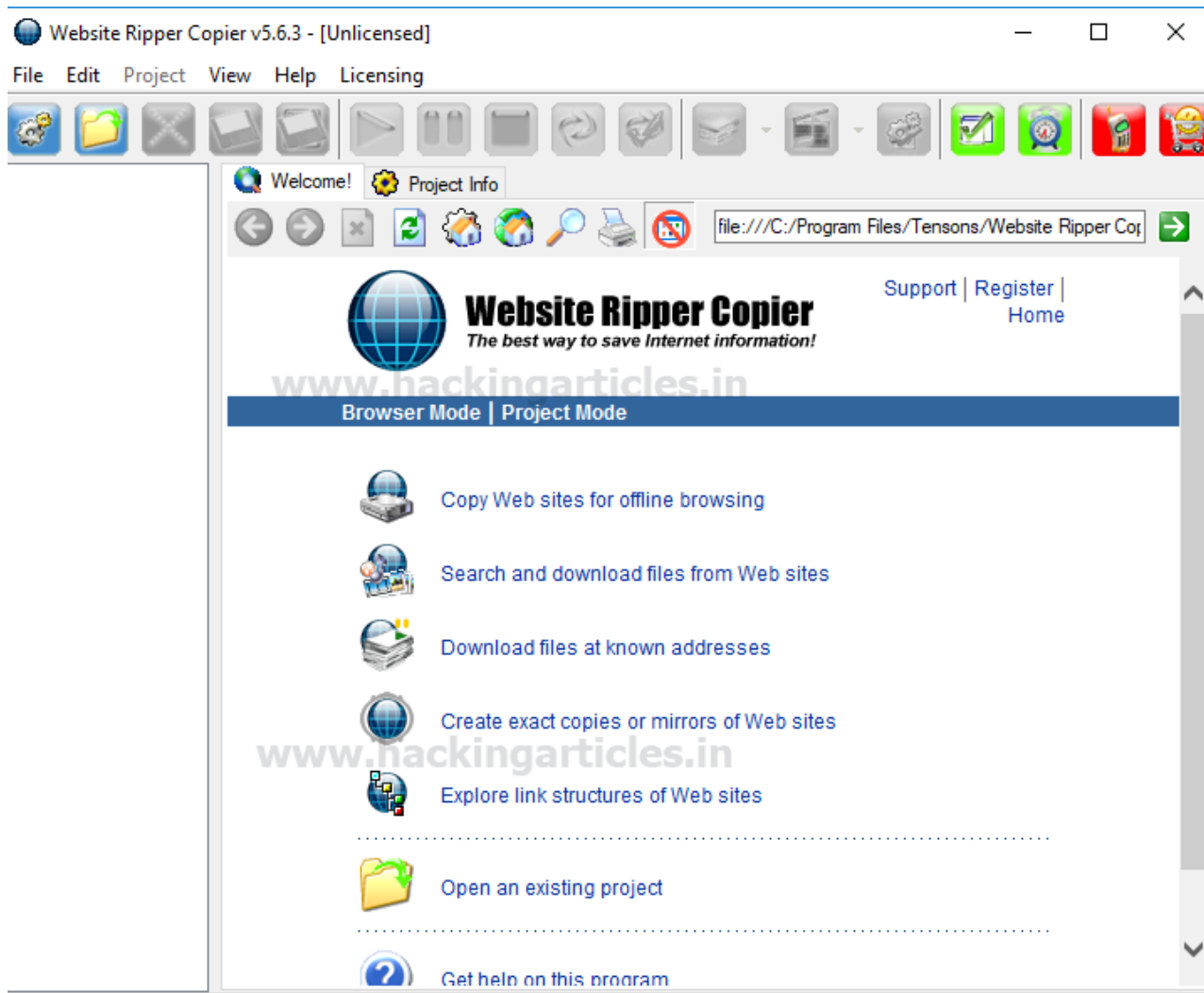
Website Ripper Copier

Website Ripper Copier (WRC) is an all-purpose, high-speed website downloader software to save website data. WRC can download website files to a local drive for offline browsing, extract website files of a certain size and type, like the image, video, picture, movie, and music, retrieve a large number of files as a download manager with resumption support, and mirror sites. WRC is also a site link validator, explorer, and tabbed antipop-up Web / offline browser.

Website Ripper Copier is the only website downloader tool that can resume broken downloads from HTTP, HTTPS and FTP connections, access password-protected sites, support Web cookies, analyze scripts, update retrieved sites or files, and launch more than fifty retrieval threads.

You can download it from [here](#).

Choose “websites for offline browsing” option.




Enter the website URL as `http://tptl.in` and click on **next**.

Enter the Starting Addresses

What Internet addresses would you prefer to explore and download data from?




Starting addresses

 Enter at least one Internet address to be the starting point for this project, one for each line. You can select from My Favorite URLs. (Valid address examples are "http://www.example.com" and "http://www.example.com/dir/page.htm".)

http://www.tptl.in

Web authorization and authentication (if needed)

 Web resources may be protected and require authorization or authentication to access them. If any of the above Internet addresses are password protected, enter the login info

Logins of Sites...

How to crawl protected sites?

< Back **Next >** Cancel

Mention directory path to save the output result and click **run now**.

Select the Project Configuration

How would you like to configure this project?



Save Parser Downloaders Mirroring Cookies Privacy Advanced

Save directory ...

Temporary directory

☒ As save directory
☐ Custom: ...

Save method

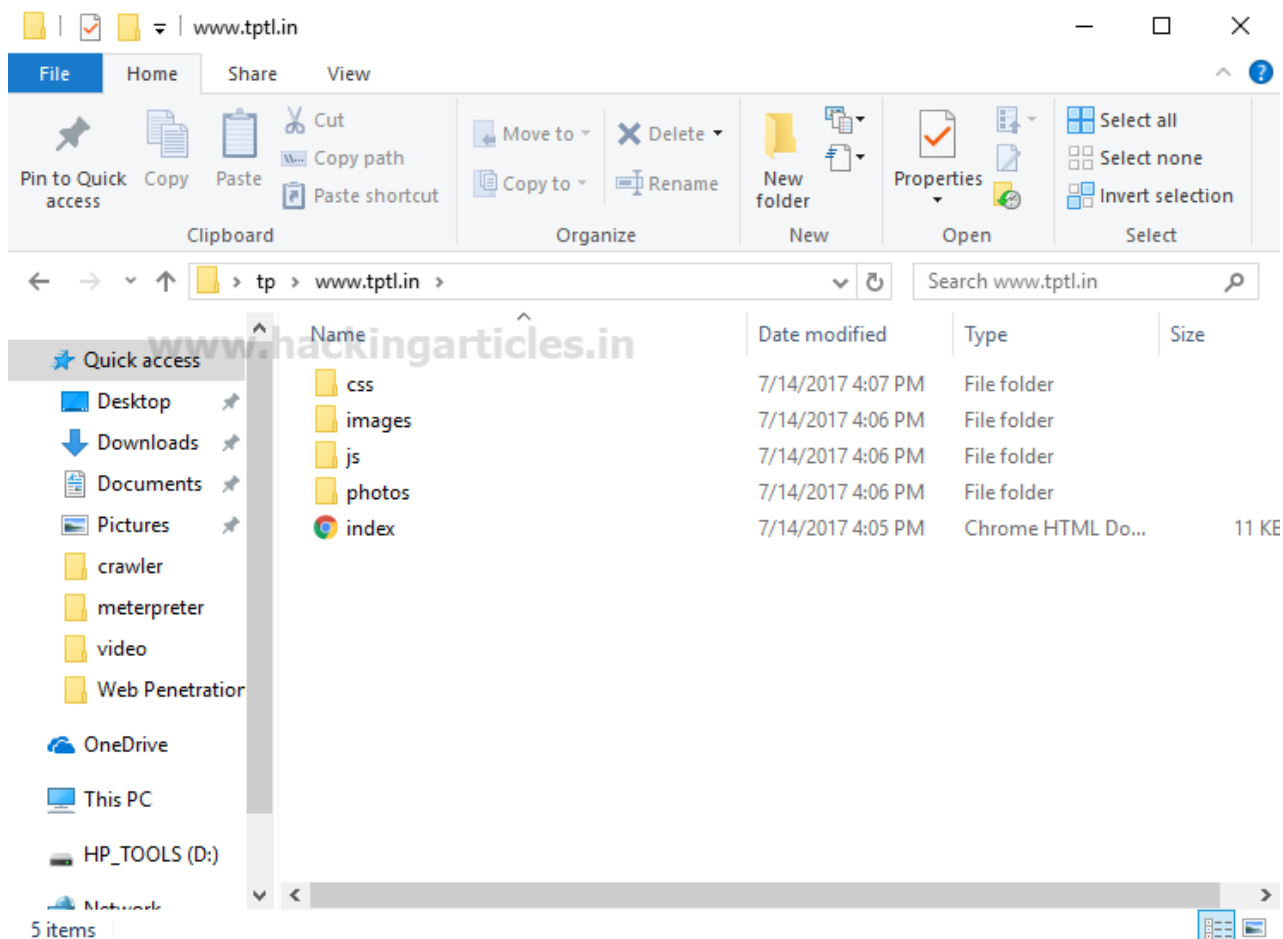
☐ Save all under one directory
☐ Organize by server name
☐ Organize by file extension
☒ Preserve directory and subdirectory structure

Filename pollution control

☒ Keep all files (by renaming conflicting filenames)
☐ Keep old files only (by disregarding new files)
☐ Keep new files only (by overwriting old files)

Run Later < Back **Run Now** Cancel

When you will open selected **folder tp** you will get fetched CSS,php,html and js file inside it.

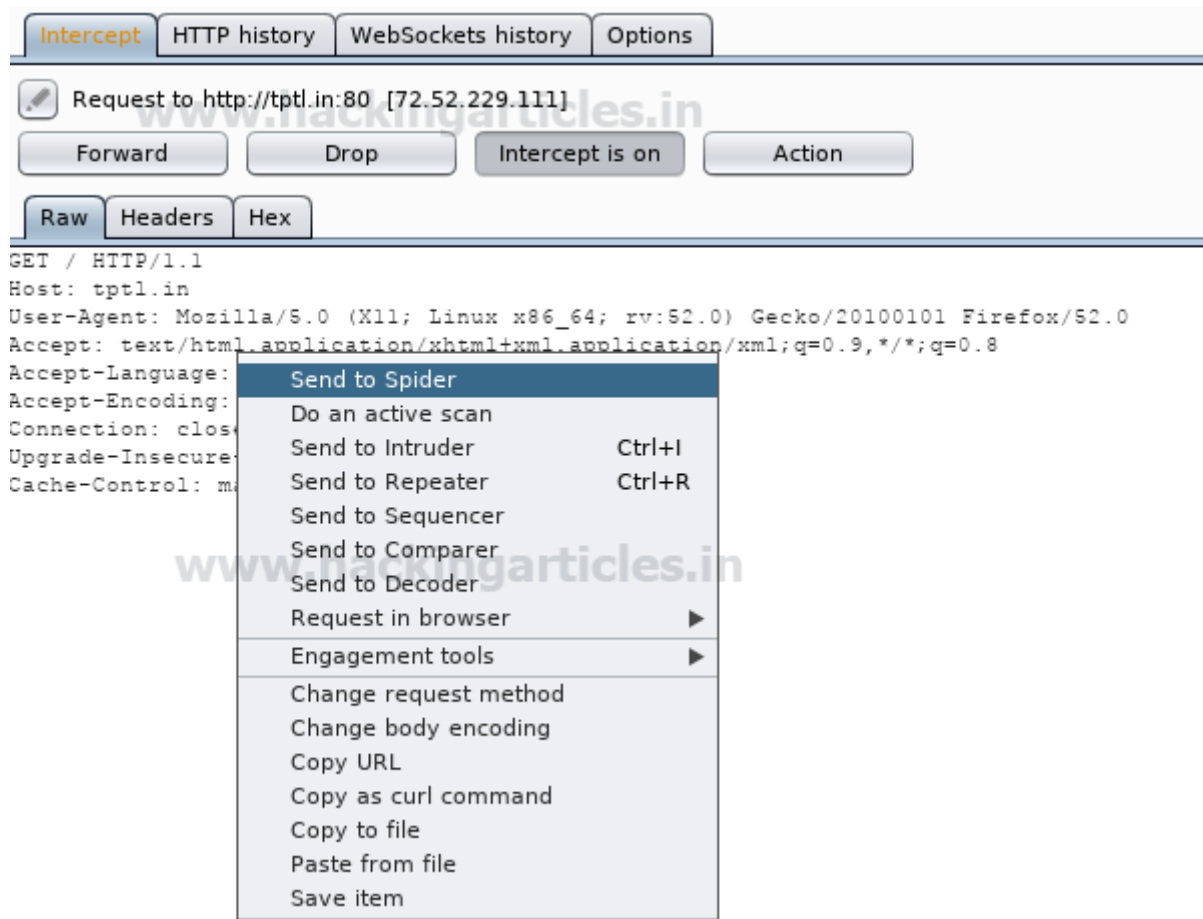


Burp Suite Spider

Burp Spider is a tool for automatically crawling web applications. While it is generally preferable to map applications manually. You can use Burp Spider to partially automate this process for very large applications, or when you are short of time.

For more detail read our previous articles from [here](#).

From given screenshot you can observe that I had fetched the http request of <http://tptl.in>. Now **send to spider** with help of action tab.



The targeted website has been added inside the **site map** under **the target** tab as a new scope for web crawling. From the screenshot, you can see it started web crawling of the target website where it has collected the website information in the form of php, html, and js.

Target

Proxy

Spider

Scanner

Intruder

Site map

Scope

Filter: Hiding not found items; hiding CSS, image and general binary content; hiding 4xx responses; hiding empty folders

- http://isotope.metafizzy.co
- http://jaysalvat.com
- http://lemonwi.se
- https://maps.google.co.in
- http://metafizzy.co
- http://platform.twitter.com
- http://player.vimeo.com
- http://sam.zoy.org
- http://theforest.net
- http://tptl.in**
 - /
 - aboutus.php
 - clients.php
 - contactus.php
 - CSS
 - dataplan.php
 - index.php
 - js
 - projects.php
 - services.php
 - videoplan.php
 - voiceplan.php
- http://twitter.com
- http://users.tpg.com.au
- http://vegas.jaysalvat.com
- http://wil.to
- http://www.alistapart.com
- http://www.apple.com
- http://www.flickr.com
- http://www.gnu.org
- http://www.kickstarter.com
- http://www.m-themes.eu

Contents

Method	URL	Params	Status	Length
GET	/		200	10906
GET	/aboutus.php		200	9095
GET	/clients.php		200	7605
GET	/contactus.php		200	9049
GET	/dataplan.php		200	7440
GET	/index.php		200	10906
GET	/js/jquery-1.8.1.min.js		200	93012
GET	/js/jquery-ui.min.js		200	28441
GET	/js/jquery.contact-form.js		200	1007
GET	/js/jquery.dcarousel.js		200	10591
GET	/js/jquery.fitvid.js		200	2883
GET	/js/jquery.flexslider.js		200	40106
GET	/js/jquery.isotope.min.js		200	16012

RequestResponse

RawHeadersHex

```

GET /js/jquery-ui.min.js HTTP/1.1
Host: tptl.in
Accept: */*
Accept-Language: en
User-Agent: Mozilla/5.0 (compatible; MSIE 9.0; Windows NT
6.1; Win64; x64; Trident/5.0)
Connection: close
Referer: http://tptl.in/

```

To gain more knowledge about hacking tools. Click on this [link](#).

Author: Aarti Singh is a Researcher and Technical Writer at Hacking Articles an Information Security Consultant Social Media Lover and Gadgets. Contact [here](#)