



Web Crawlers

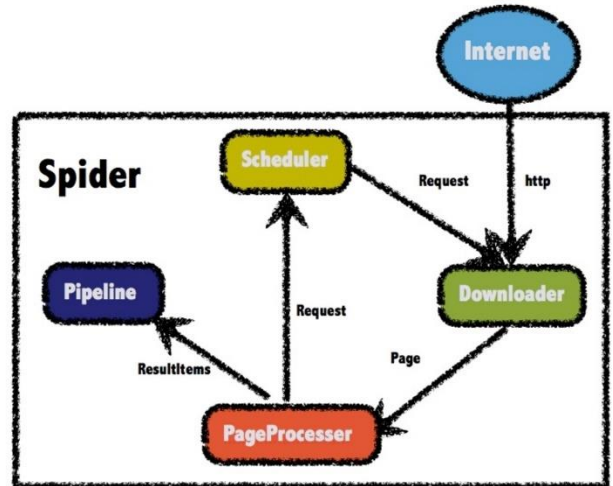
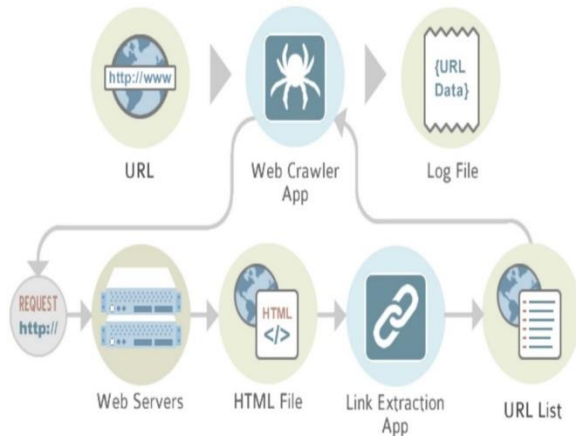
Information Security Inc.

Contents

- What are Web Crawlers?
- Ways to crawl a website
- References

What are Web Crawlers?

- Web crawlers are known by a variety of names – industry jargon labels them spiders or bots but technically they are referred to as web crawlers



```
root@LUCKY64:~# msiconsole
```

```
IIIIII dTb.dTb
II 4' v 'B
II 6. .P
II 'T; .;P'
II 'T; ;P'
IIIIII 'YvP'
```



```
I love shells --egypt
```

```
= [ metasploit v4.14.28-dev ]
+ -- ---[ 1662 exploits - 951 auxiliary - 293 post ]
+ -- ---[ 486 payloads - 40 encoders - 9 nops ]
+ -- ---[ Free Metasploit Pro trial: http://r-7.co/trymsp ]
```

```
msf > use auxiliary/crawler/msfcrawler
msf auxiliary(msfcrawler) > set RHOSTS 192.168.10.79
RHOSTS => 192.168.10.79
msf auxiliary(msfcrawler) > exploit
```

```
[*] Loading modules: /usr/share/metasploit-framework/data/msfcrawler
[*] Loaded crawler module Simple from /usr/share/metasploit-framework/data/msfcrawler/basic.rb...
[*] Loaded crawler module Comments from /usr/share/metasploit-framework/data/msfcrawler/comments.rb...
[*] Loaded crawler module Forms from /usr/share/metasploit-framework/data/msfcrawler/forms.rb...
[*] Loaded crawler module Frames from /usr/share/metasploit-framework/data/msfcrawler/frames.rb...
[*] Loaded crawler module Image from /usr/share/metasploit-framework/data/msfcrawler/image.rb...
[*] Loaded crawler module Link from /usr/share/metasploit-framework/data/msfcrawler/link.rb...
[*] Loaded crawler module Objects from /usr/share/metasploit-framework/data/msfcrawler/objects.rb...
[*] Loaded crawler module Scripts from /usr/share/metasploit-framework/data/msfcrawler/scripts.rb...
[*] OK
[*] URI LIMITS ENABLED: 10 (Maximum number of requests per uri)
[*] Target: 192.168.10.79 Port: 80 Path: / SSL:
[*] >> [200] /
[*] >> [200] /twiki/
[*] >> [200] /phpMyAdmin/
[*] >> [200] /mutillidae/
[*] >> [302] /dvwa/
[302] Redirection to: login.php
```

Ways to crawl a website

```
root@LUCKY64:~# httrack http://192.168.10.79 -O /root/Site
WARNING! You are running this program as root!
It might be a good idea to run as a different user
Mirror launched on Mon, 07 Aug 2017 03:03:55 by HTTrack Website Copier/3.49-2 [XR&CO!2014]
mirroring http://192.168.10.79 with the wizard help..
* 192.168.10.79/phpMyAdmin/phpmyadmin.css.php?lang=en-utf-8&convcharset=utf-8&token=9ee469a33b32a6dc48c85c600f485c6d&js_f
* 192.168.10.79/mutillidae/images/twitter.gif (1828 bytes) - OK
* 192.168.10.79/mutillidae/images/youtube_256_256.png (39246 bytes) - OK

* 192.168.10.79/mutillidae/index.php?page=documentation/how-to-access-Mutillidae-over-Virtual-Box-network.php (21603 byte
* 192.168.10.79/mutillidae/index.php?page=php-errors.php (22462 bytes) - OK
Bytes saved: 4,57MiB Links scanned: 6/75 (+65)
Time: 4min57s Files written: 69
Transfer rate: 12,27KiB/s (10,77KiB/s) Files updated: 0
Active connections: 4 Errors: 0

Current job: parsing HTML file (45%)
request - 192.168.10.79/dvwa/login.php 0B / 8,00KiB
```

```
root@LUCKY64:~/Site/192.168.10.79# pwd
~/root/Site/192.168.10.79
root@LUCKY64:~/Site/192.168.10.79# ls -la
total 52
drwxr-xr-x 12 root root 4096 Aug 7 03:09 .
drwxr-xr-x 4 root root 4096 Aug 7 03:03 ..
drwxr-xr-x 2 root root 4096 Aug 7 03:16 dav
drwxr-xr-x 3 root root 4096 Aug 7 03:12 dvwa
drwxr-xr-x 2 root root 4096 Aug 7 03:08 icons
-rw-r--r-- 1 root root 1162 Aug 7 03:02 index.html
drwxr-xr-x 6 root root 4096 Aug 7 03:16 mutillidae
drwxr-xr-x 3 root root 4096 Aug 7 03:09 oops
drwxr-xr-x 3 root root 4096 Aug 7 03:08 p
drwxr-xr-x 3 root root 4096 Aug 7 03:10 phpMyAdmin
drwxr-xr-x 3 root root 4096 Aug 7 03:09 rdiff
drwxr-xr-x 4 root root 4096 Aug 7 03:12 twiki
drwxr-xr-x 3 root root 4096 Aug 7 03:09 view
```

Ways to crawl a website

- Black Widow

The screenshot displays the Black Widow web crawler interface. The top navigation bar includes icons for Browser, Filters, Scanner, Link Errors, Emails, Ext Links, Structure, and NetSpy. The 'Scanner' icon is highlighted with a red box.

Locating the site or page to scan.

Use the browser to navigate to the site or page you would like to scan. Optionally, you may need to login to the site before you the Filters to control the scanning process.

Address: <http://192.168.10.79/>

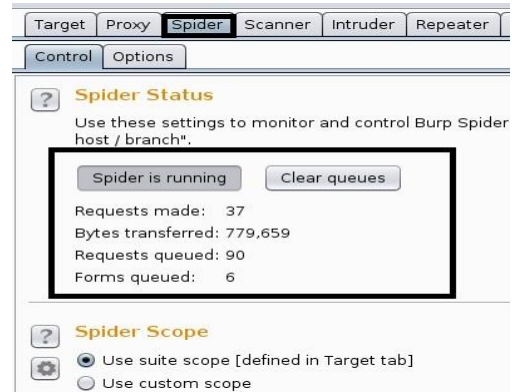
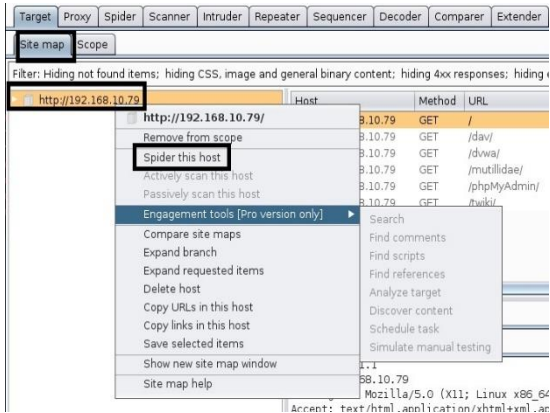
Scanner settings and status

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 select (from the Structure) what you need to download. Keep in mind that scanning with more than 6 connections may result in connect servers only allow this much. When you start the scan, the list below will show you the status for each scan connections.

Status	URL
100% of 453,477 bytes	http://192.168.10.79/twiki/TWikiDocumentation.html#Hiding_Control_Settings
100% of 453,477 bytes	http://192.168.10.79/twiki/TWikiDocumentation.html#The_Super_AdminGroup
100% of 453,477 bytes	http://192.168.10.79/twiki/TWikiDocumentation.html#Using_HTML
100% of 453,477 bytes	http://192.168.10.79/twiki/TWikiDocumentation.html#TWiki_Text_Formatting
100% of 453,477 bytes	http://192.168.10.79/twiki/TWikiDocumentation.html#TWiki_Editing_Shorthand
Response Received	http://192.168.10.79/twiki/TWikiDocumentation.html#HTML_and_TWiki_Usability

Ways to crawl a website

- Burp Suite Spider



Ways to crawl a website

- Scrapy framework
(<https://doc.scrapy.org/en/master/intro/tutorial.html>)

```
pip install scrapy
```

```
root@LUCKY64:~/opt3# scrapy -h
Scrapy 1.4.0 - no active project

Usage:
  scrapy <command> [options] [args]

Available commands:
  bench           Run quick benchmark test
  fetch           Fetch a URL using the Scrapy downloader
  genspider       Generate new spider using pre-defined templates
  runspider       Run a self-contained spider (without creating a project)
  settings        Get settings values
  shell           Interactive scraping console
  startproject    Create new project
  version         Print Scrapy version
  view            Open URL in browser, as seen by Scrapy

[ more ]         More commands available when run from project directory

Use "scrapy <command> -h" to see more info about a command
```


Ways to crawl a website

- Scrapy framework
(<https://doc.scrapy.org/en/master/intro/tutorial.html>)

```
root@LUCKY64:~# scrapy startproject simple_spider
New Scrapy project 'simple_spider', using template directory '/usr/local/lib/python2.7/dist-packages/scrapy/templates/project', created in:
/opt3/simple_spider

You can start your first spider with:
cd simple_spider
scrapy genspider example example.com

root@LUCKY64:~/opt3/simple_spider# pwd
/opt3/simple_spider
root@LUCKY64:~/opt3/simple_spider# ls -la
total 16
drwxr-xr-x  3 root root 4096 Aug  7 05:32 .
drwxr-xr-x 17 root root 4096 Aug  7 05:32 ..
-rw-r--r--  1 root root  270 Aug  7 05:32 scrapy.cfg
drwxr-sr-x  3 root root 4096 Aug  7 05:32 simple_spider

root@LUCKY64:~/opt3/simple_spider/simple_spider/spiders# scrapy genspider hackbot hackyourselffirst.troyhunt.com
Created spider 'hackbot' using template 'basic' in module:
simple_spider.spiders.hackbot
root@LUCKY64:~/opt3/simple_spider/simple_spider/spiders# cat hackbot.py
# -*- coding: utf-8 -*-
import scrapy

class HackbotSpider(scrapy.Spider):
    name = 'hackbot'
    allowed_domains = ['hackyourselffirst.troyhunt.com']
    start_urls = ['http://hackyourselffirst.troyhunt.com/']

    def parse(self, response):
        pass
```

Ways to crawl a website

- Scrapy framework
(<https://doc.scrapy.org/en/master/intro/tutorial.html>)

```
root@LUCKY64:/opt3/simple_spider/simple_spider/spiders# scrapy runspider hackbot.py
2017-08-07 06:14:48 [scrapy.utils.log] INFO: Scrapy 1.4.0 started (bot: simple_spider)
2017-08-07 06:14:48 [scrapy.utils.log] INFO: Overridden settings: {'NEWSPIDER_MODULE': 'simple_spider.spiders', 'SPIDER_LOADER_WARN_ONLY': True, 'SPIDER_MODULES': ['simple_spider.spiders'], 'ROBOTSTXT_OBEY': True, 'BOT_NAME': 'simple_spider'}
2017-08-07 06:14:48 [scrapy.middleware] INFO: Enabled extensions:
2017-08-07 06:14:49 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://hackyourselffirst.troyhunt.com/robots.txt> (referrer: None)
2017-08-07 06:14:49 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://hackyourselffirst.troyhunt.com/> (referrer: None)
2017-08-07 06:14:49 [scrapy.core.engine] INFO: Closing spider (finished)
```

Ways to crawl a website

- Scrapy framework
(<https://doc.scrapy.org/en/master/intro/tutorial.html>)
 - ▲ Example Spider (extract all links and follow them)

```
root@LUCKY64:~/scrapy/example_spiders/example_spider/spiders# cat bot.py
# -*- coding: utf-8 -*-

import scrapy
from scrapy.spiders import CrawlSpider, Rule
from scrapy.linkextractors import LinkExtractor
#from scrapy.linkextractors import LxmlLinkExtractor

class HackbotSpider(CrawlSpider):
    name = 'bot'
    start_urls = [
        'http://192.168.10.79/',
        'http://192.168.10.79/dvwa/'
    ]

    rules=[
        Rule(LinkExtractor(canonicalize=True, unique=True), follow=True, callback="parse")]

    def start_requests(self):
        for url in self.start_urls:
            yield scrapy.Request(url, callback=self.parse, dont_filter=True)

    def parse(self, response):
        print ('Processing..' + response.url)
        links = LinkExtractor(canonicalize=True, unique=True).extract_links(response)
        print links
```

```
scrapy runspider bot.py
```

```
017-06-07 07:47:28 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://192.168.10.79/dvwa/login.php (referer: None)
Processing...http://192.168.10.79/
Link(url='http://192.168.10.79/twiki/', text='Twiki', fragment='', nofollow=False), Link(url='http://192.168.10.79/phpMy
admin/', text='phpMyAdmin', fragment='', nofollow=False), Link(url='http://192.168.10.79/mutillidae/', text='Mutillidae',
fragment='', nofollow=False), Link(url='http://192.168.10.79/dvwa/', text='DVWA', fragment='', nofollow=False), Link(url
'http://192.168.10.79/dav/', text='WebDAV', fragment='', nofollow=False)]
```

References

- Wikipedia
https://en.wikipedia.org/wiki/Web_crawler
- ScienceDaily
https://www.sciencedaily.com/terms/web_crawler.htm
- Metasploit
<https://www.metasploit.com>
- HTTrack
<https://www.httrack.com>
- Black Widow
<http://softbytelabs.com/us/downloads.html>
- Burp Suite
<https://portswigger.net/burp>
- Scrapy
<https://scrapy.org/>