



AUGUST 4-5, 2021

ARSENAL

Introducing SubCrawl

A Framework for the Analysis and Clustering of Hacking Tools Found Using Open Directories

Josh Stroschein



Malware Analyst, HP Inc.

josh.stroschein@hp.com
@jstrosch

Patrick Schläpfer



Malware Analyst, HP Inc.

patrick.schlaepfer@hp.com
@stoerchl

Alex Holland

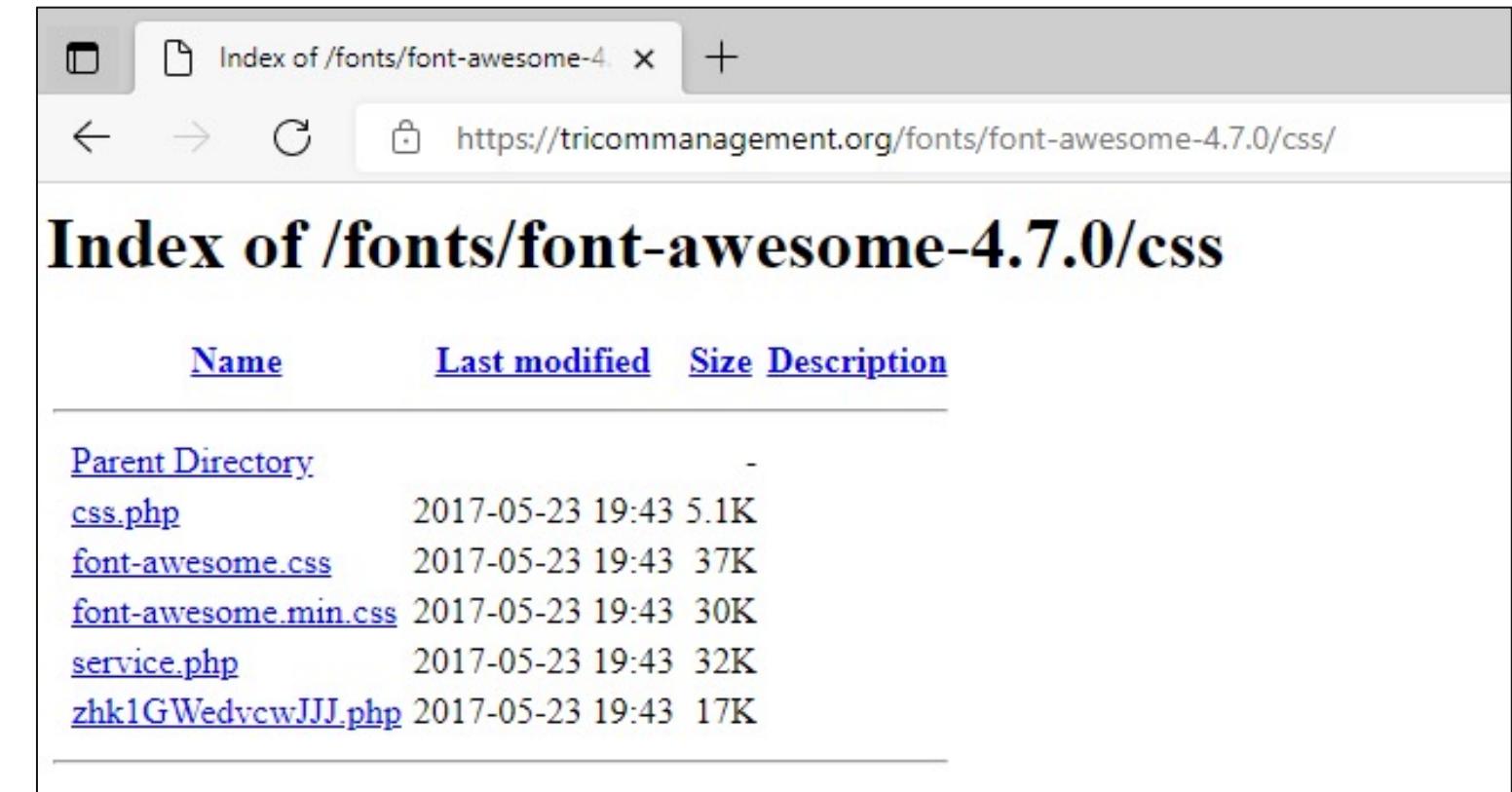


Malware Analyst, HP Inc.

alex.holland@hp.com
@cryptogramfan

What Are Open Directories?

- An open directory shows the file listing on a webserver
- Sometimes show additional information like timestamps, filesizes and descriptions
- Allow users to download files which are normally interpreted by the webserver
- Usually start with “Index of...”
- Parent and/or child directories can be traversed



The screenshot shows a web browser window with the following details:

- Address bar: https://tricommangement.org/fonts/font-awesome-4.7.0/css/
- Title bar: Index of /fonts/font-awesome-4.7.0/css
- Content area:
 - Index of /fonts/font-awesome-4.7.0/css**
 - Table with columns: Name, Last modified, Size, Description.
 - Data rows:

Name	Last modified	Size	Description
Parent Directory		-	
css.php	2017-05-23 19:43	5.1K	
font-awesome.css	2017-05-23 19:43	37K	
font-awesome.min.css	2017-05-23 19:43	30K	
service.php	2017-05-23 19:43	32K	
zhk1GWedvcwJJJ.php	2017-05-23 19:43	17K	

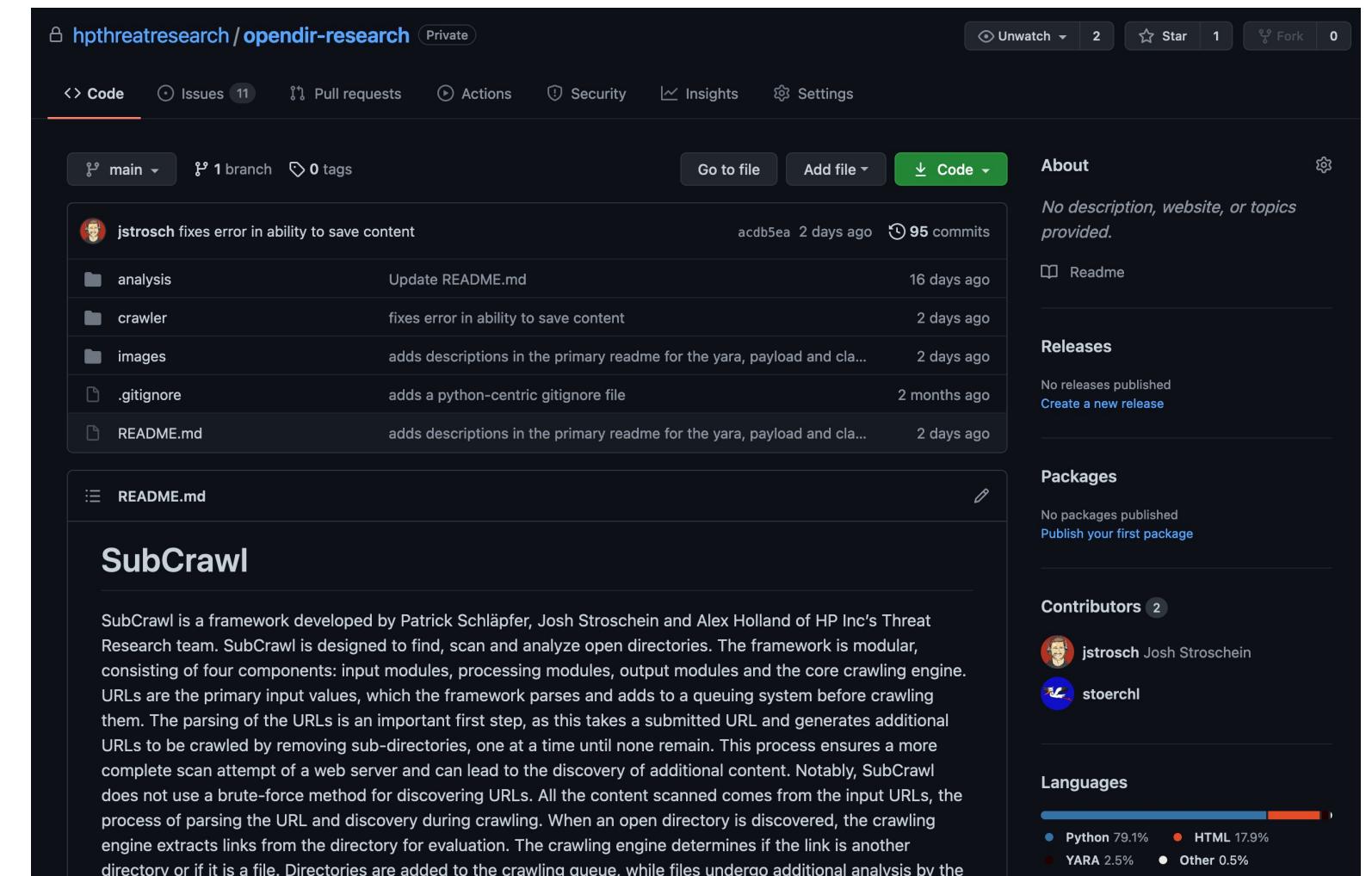
Why Open Directories?

From time to time, open directories are found on websites that are used for malware hosting and can thus provide malware analysts with a deeper insight into the used tools. Scanning and analyzing a large amount of open directories hosting malware can potentially answer two interesting questions:

- Do threat actors compromise the websites or buy access to it?
- How are the websites compromised?

Introducing SubCrawl

- To systematically detect and analyze such open directories, we developed SubCrawl.
- SubCrawl is a modular framework that scans URLs for open directory properties and analyzes their content using various processing modules.
- A few examples of such processing modules are:
PayloadProcessing
YARAProcessing
ClamAVProcessing
TLSHProcessing



The screenshot shows the GitHub repository page for `hpthreatresearch/opendir-research`. The repository is private, has 2 stars, 0 forks, and 11 issues. The code tab is selected, showing a commit history from `jstrosch` fixing errors in saving content. The README.md file is expanded, providing a detailed description of SubCrawl as a modular framework for finding, scanning, and analyzing open directories. It highlights four components: input modules, processing modules, output modules, and a crawling engine. The repository also includes sections for About, Releases, Packages, Contributors, and Languages.

About
No description, website, or topics provided.

Releases
No releases published
[Create a new release](#)

Packages
No packages published
[Publish your first package](#)

Contributors 2

-  **jstrosch** Josh Stroschein
-  **stoerchi**

Languages

Python 79.1%	HTML 17.9%
YARA 2.5%	Other 0.5%

<https://github.com/hptrustresearch/subcrawl>

SubCrawl Modes of Operation

- SubCrawl offers two different operation modes.
 - Run-Once Mode*
 - Service Mode*
- The run-once mode is suitable if you want to quickly scan a website and evaluate the results directly on the console.
- The service mode is intended for scanning a larger number of websites over a longer period of time. The results are saved for later analysis using a configured storage module.

 	283	misp@stoerchl.ch	2021-06-24	jx2.info
 	3860	patrick.schlapfer@hp.com	2021-05-03	Domain track event
  	21	misp@stoerchl.ch	2021-06-24	upload.vina-host.com
 	34	misp@stoerchl.ch	2021-06-24	www.56kdw.com
  	16	misp@stoerchl.ch	2021-06-24	simhisancak.xyz
 	21	misp@stoerchl.ch	2021-06-24	downloadas.xyz
  	17	misp@stoerchl.ch	2021-06-24	down.esudaa.com
 	18	misp@stoerchl.ch	2021-06-24	immigrationentry.xyz
 				
 				
 				

SubCrawl Storage Modules

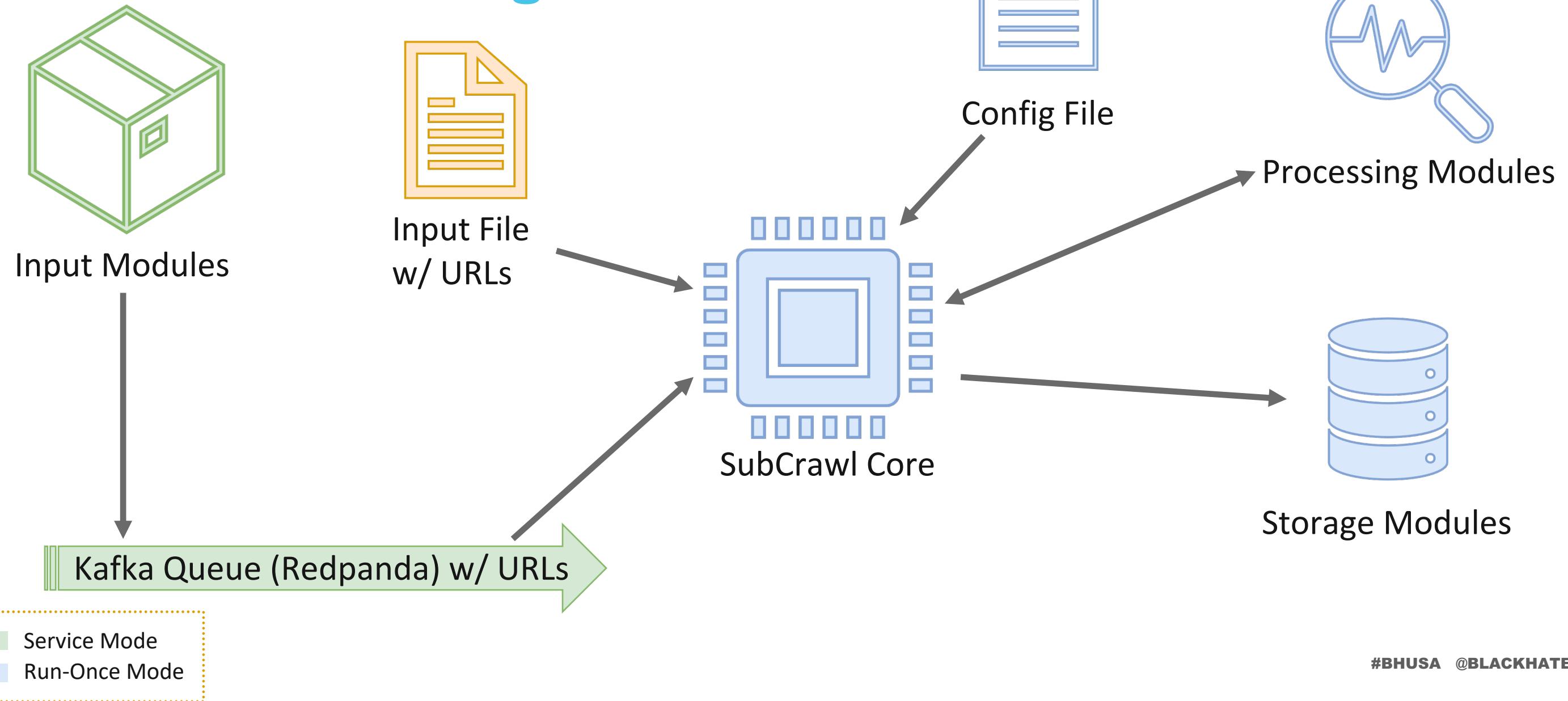
- The results from the processing modules can then be printed on the command line or saved in the desired form for later analysis based on the activated storage modules.
- Currently the following additional storage modules are implemented:
 - SQLiteStorage*
 - MISPStorage*

```
stoerchl@ homebase  crawler: python3 recursive-crawl.py -f urls.txt -s ConsoleStorage -p YARAProcessing,TLSHProcessing,JARMProcessing
[REDACTED]
~~ Harvesting the Open Web ~~

2021-06-23 11:30:56,595 - SubCrawl - INFO - [*] Loaded storage module: ConsoleStorage
2021-06-23 11:30:56,596 - SubCrawl - INFO - [*] Loaded processing module: YARAProcessing
2021-06-23 11:30:56,597 - SubCrawl - INFO - [*] Loaded processing module: TLSHProcessing
2021-06-23 11:30:56,598 - SubCrawl - INFO - [*] Loaded processing module: JARMProcessing
2021-06-23 11:30:56,599 - SubCrawl - INFO - [*] Parsing input sources...
2021-06-23 11:30:56,606 - SubCrawl - INFO - [*] Found 2 hosts to scrape
2021-06-23 11:30:56,607 - SubCrawl - DEBUG - Generated new URL: https://narayanhitihomestay.com/
2021-06-23 11:30:56,607 - SubCrawl - DEBUG - Generated new URL: https://narayanhitihomestay.com/dameon-kuphal/
2021-06-23 11:30:56,607 - SubCrawl - DEBUG - Generated new URL: https://fedhaminerals.com/
2021-06-23 11:30:56,608 - SubCrawl - DEBUG - Generated new URL: https://fedhaminerals.com/7qPxcmczwcdb/
2021-06-23 11:30:56,608 - SubCrawl - INFO - [*] Done parsing URLs, ready to begin scraping 2 hosts and 4 URLs... starting in 0 seconds!
2021-06-23 11:30:56,627 - SubCrawl - DEBUG - Starting down path... https://narayanhitihomestay.com/
2021-06-23 11:30:56,628 - SubCrawl - DEBUG - Starting down path... https://fedhaminerals.com/
2021-06-23 11:30:57,390 - SubCrawl - DEBUG - [*] Discovered: https://narayanhitihomestay.com/\'?ND\
2021-06-23 11:30:57,729 - SubCrawl - DEBUG - [*] Discovered: https://fedhaminerals.com/dr--maude-carroll-dds/
2021-06-23 11:30:57,761 - SubCrawl - DEBUG - [*] Discovered: https://narayanhitihomestay.com/\'?MA\
2021-06-23 11:30:58,106 - SubCrawl - DEBUG - [*] Discovered: https://narayanhitihomestay.com/\'?SA\
2021-06-23 11:30:58,462 - SubCrawl - DEBUG - [*] Discovered: https://narayanhitihomestay.com/\'?DA\
2021-06-23 11:30:58,562 - SubCrawl - DEBUG - [*] Discovered: https://fedhaminerals.com/license.php
2021-06-23 11:30:58,822 - SubCrawl - DEBUG - [*] Discovered: https://narayanhitihomestay.com/ttt/
2021-06-23 11:30:59,197 - SubCrawl - DEBUG - [*] Discovered: https://narayanhitihomestay.com/ttt/\'?ND\
[!!] YARA matches https://fedhaminerals.com/license.php (protected_webshell)
2021-06-23 11:30:59,576 - SubCrawl - DEBUG - [*] Discovered: https://narayanhitihomestay.com/ttt/\'?MA\'
```

SubCrawl Run-once Mode Demo

The SubCrawl Engine



Service Mode Setup

- To make not only the run-once mode easy to use, but also the service mode, we wrapped it in a Docker container.
- As already shown in the framework architecture, the service mode relies on a queuing system. For this purpose, we use Redpanda, which is also started directly as a separate Docker container.

```
crawler > docker-compose.yml
 1  version: "3.9"
 2  services:
 3    |   web:
 4    |     build: .
 5     |     ports:
 6     |       - 8000:8000
 7
 8
 9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
FROM python:3.8-slim-buster
WORKDIR /subcrawl
RUN apt-get -y update
RUN apt-get -y upgrade
RUN apt-get -y install build-essential gcc \
    yara magic supervisor clamav-daemon \
    clamav-freshclam clamav-unofficial-sigs
COPY requirements.txt requirements.txt
COPY supervisor/supervisord.conf /etc/supervisor/
RUN mkdir /var/log/subcrawl
RUN pip3 install -r requirements.txt
RUN freshclam
EXPOSE 8000
COPY . .
CMD ["bash", "run.sh"]
- PLAINTEXT://0.0.0.0:29092,OUTSIDE://0.0.0.0:9092
- --advertise-kafka-addr
- PLAINTEXT://redpanda:29092,OUTSIDE://localhost:9092
image: docker.vectorized.io/vectorized/redpanda:latest
container_name: redpanda-1
ports:
- 9092:9092
- 29092:29092
```

Service Mode Setup

- The two developed modules from URLhaus and Phishtank are used as URL input. The URLs to be scanned are added to the queue, which is continuously processed by the crawler.
- The processing and storage modules that are to be used must be entered in the config file

```
processing_modules:  
  - ClamAVProcessing  
  - WebshellProcessing  
  - JARMProcessing  
  - TLSHProcessing  
  - YARAProcessing  
storage_modules:  
  - SqliteStorage
```

Service Mode Setup

- To build and start the container execute the following command:

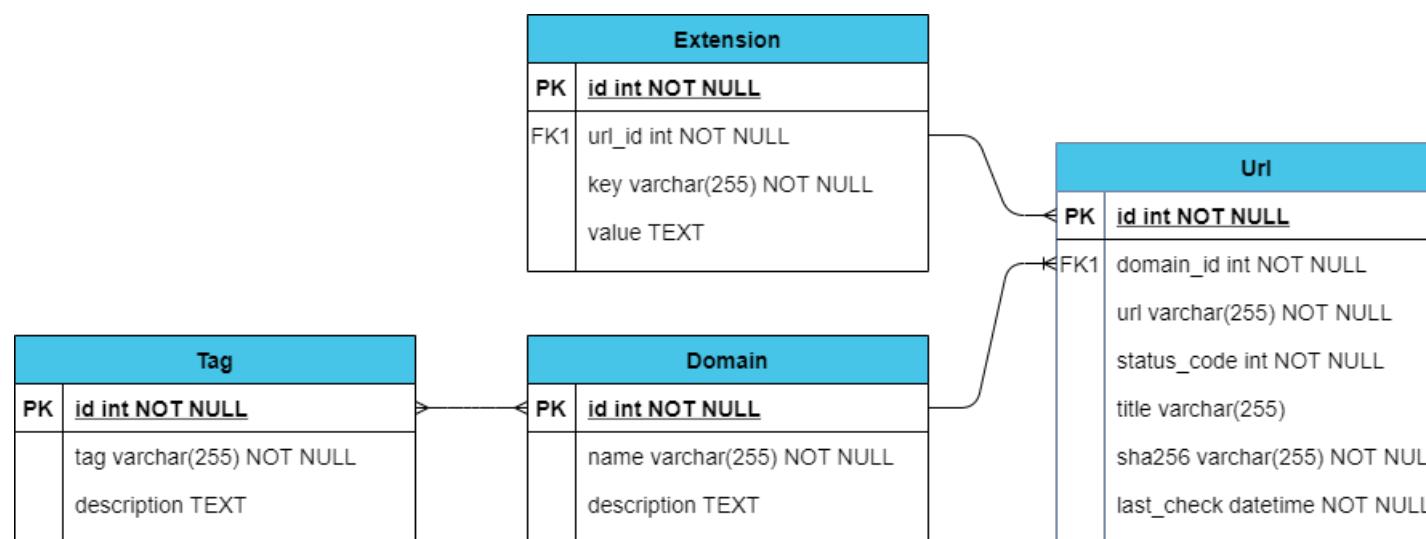
docker-compose up –build

- As soon as the containers are built and running, the web interface is reachable on port 8000

```
web_1    | Starting ClamAV daemon: clamd .
web_1    | Starting supervisor: supervisord.
web_1    | [2021-06-23 12:19:06 +0000] [368] [INFO] Starting gunicorn 20.0.4
web_1    | [2021-06-23 12:19:06 +0000] [368] [INFO] Listening at: http://0.0.0.0:8000 (368)
web_1    | [2021-06-23 12:19:06 +0000] [368] [INFO] Using worker: sync
web_1    | [2021-06-23 12:19:06 +0000] [373] [INFO] Booting worker with pid: 373
web_1    | [2021-06-23 12:19:06 +0000] [374] [INFO] Booting worker with pid: 374
web_1    | [2021-06-23 12:19:06 +0000] [375] [INFO] Booting worker with pid: 375
web_1    | [2021-06-23 12:19:06 +0000] [377] [INFO] Booting worker with pid: 377
```

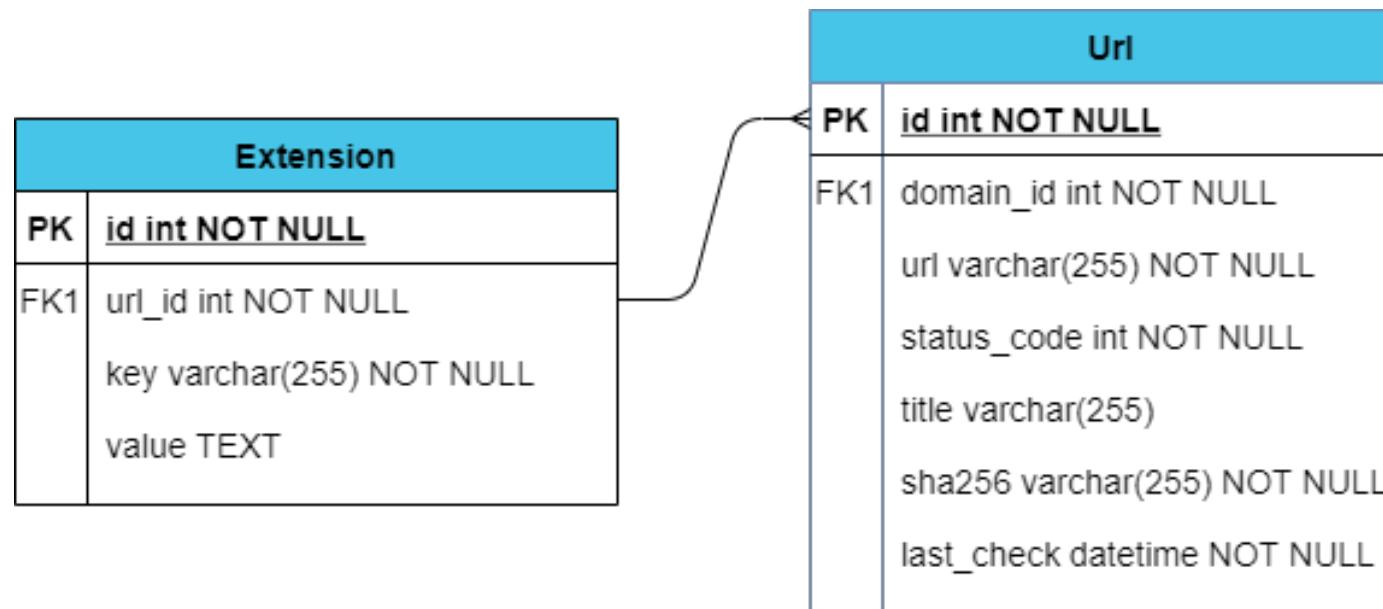
IMAGE	COMMAND	CREATED	STATUS	PORTS	NAMES
crawler_web	"bash run.sh"	7 minutes ago	Up 7 minutes	0.0.0.0:8000->8000/tcp,	crawler_web_1
docker.vectorized.io/vectorized/redpanda:latest	"/usr/bin/rpk redpan..."	2 days ago	Up 7 minutes	0.0.0.0:9092->9092/tcp, 0.0.0.0:29092->29092/tcp, 9644/tcp	redpanda-1

SQLite Storage Module - Database



- The SQLite Storage module is intended for longer storage of scan results.
- SQLite is configured as default storage module in service mode if the supplied Docker container is used.
- The database schema is kept very simple with most tables being self-explanatory.

SQLite Storage Module - Database



- The “Extension” table is used to save HTTP headers of the scanned website as well as results of the processing modules
- Simple key – value table connected with an URL

SQLite Storage Module - Import

```
def store_result(self, result_data):
    # Load URLHaus tags
    url_info = dict()
    r = requests.get(self.cfg["misp"]["urlhaus_api"], allow_redirects=True)
    csv_data = io.StringIO(r.content.decode("utf-8"))
    counter = 0
    while counter < 8:
        next(csv_data)
        counter += 1

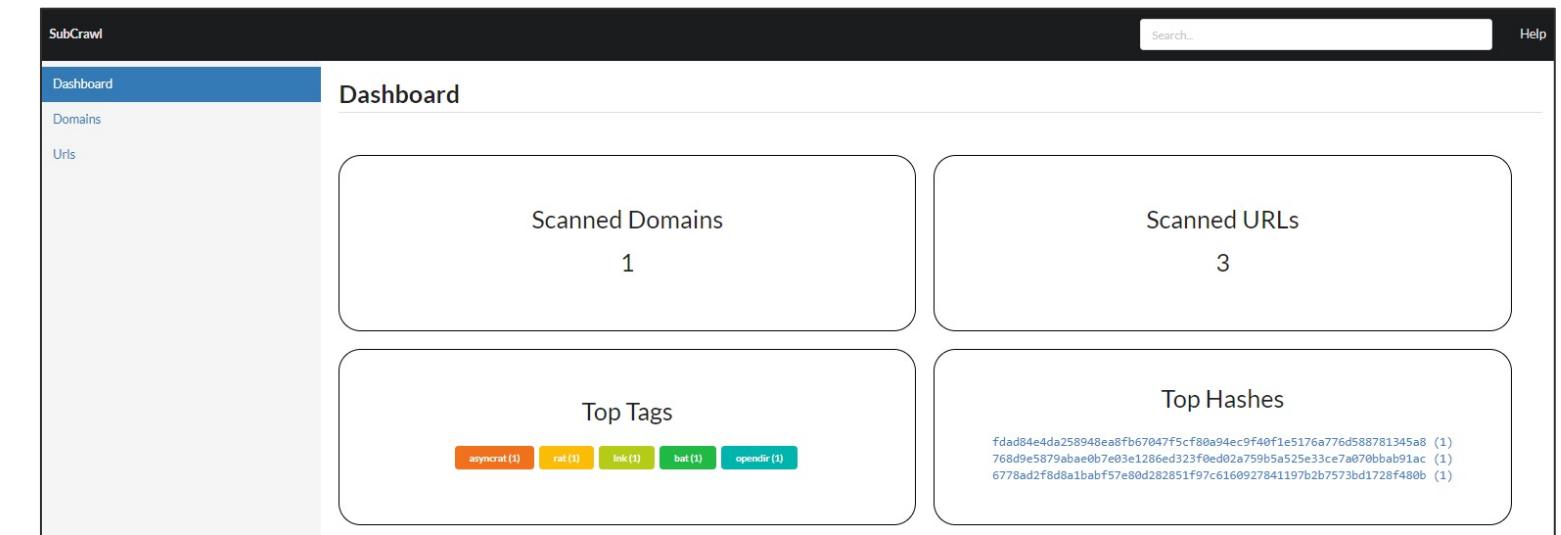
    csv_reader = csv.DictReader(csv_data)
    for row in csv_reader:
        domain = urlparse(row["url"]).netloc
        if domain not in url_info:
            url_info[domain] = set()
        url_info[domain].update(row["tags"].lower().split(","))

    result_data["urlhaus_tags"] = url_info
```

- The implementation of the SQLite storage module imports the database model, which is based on peewee from the utils folder.
- To enrich the data with additional contextual information, the storage module imports the tags from URLhaus and links them to the scanned domains.

SQLite Storage Module - Visual

- In order to analyze and manage the results of the scanned web pages in a user-friendly way, we have developed a simple web application.
- This allows the user to view, partially edit, delete or also search for data.
- On the home page, we implemented simple metrics that show how many domains and URLs have already been scanned and which tags and hashes occur the most.



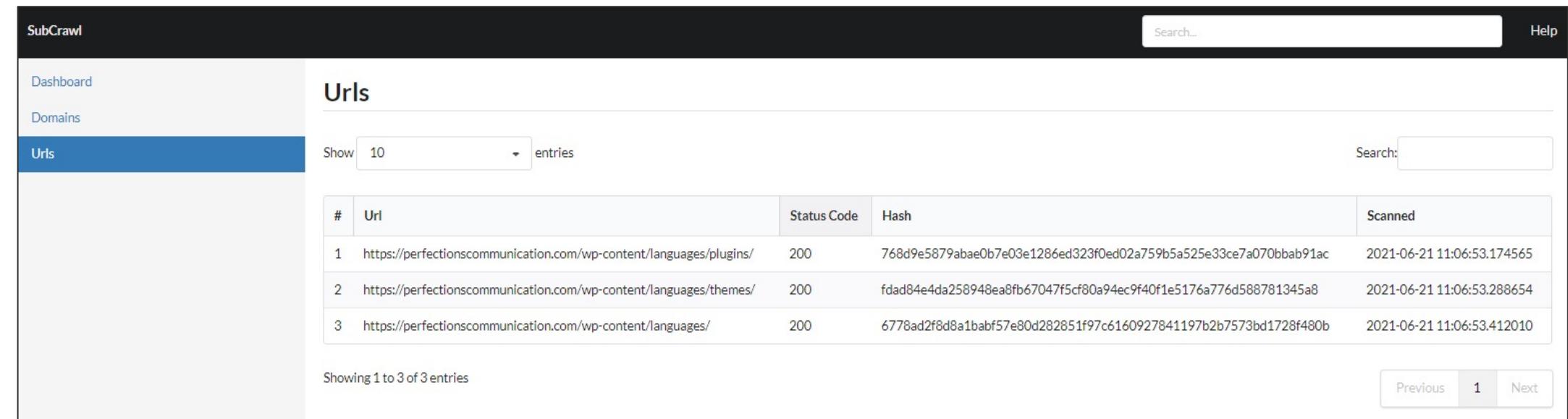
The screenshot shows the 'Dashboard' page of the SubCrawl web application. The left sidebar has 'Dashboard' selected, along with 'Domains' and 'Urls'. The main area is titled 'Dashboard' and contains four cards: 'Scanned Domains' (1), 'Scanned URLs' (3), 'Top Tags' (with tags: 'asyncrat (1)', 'rat (1)', 'link (1)', 'bat (1)', 'opendir (1)'), and 'Top Hashes' (with hashes: 'fdad84e4da258948ea8fb67047f5cf80a94ec9f40f1e5176a776d588781345a8 (1)', '768d9e5879abae0b7e03e1286ed323fe0ed2a759b5a525e33ce7a070bbab91ac (1)', '6778ad2f8d8a1babf57e80d282851f97c6160927841197b2b7573bd1728f480b (1)').

SQLite Storage Module - Visual

- The search can be used to query all the fields of the URLs, using the following syntax:

fieldname:value

- The result is then displayed in a list of found URLs.



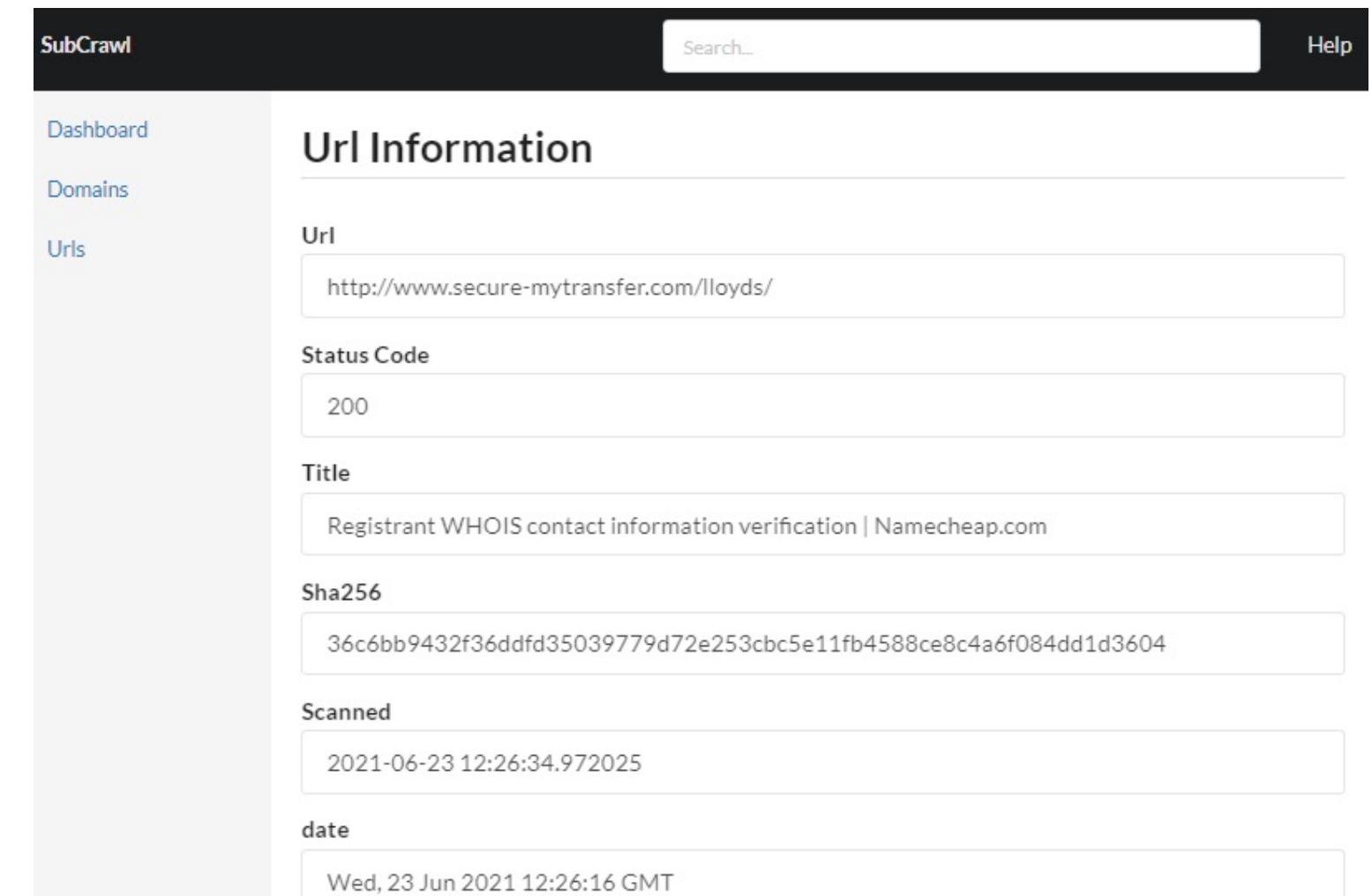
The screenshot shows the SubCrawl application interface. The left sidebar has navigation links: Dashboard, Domains, and Urls (which is currently selected and highlighted in blue). The main content area is titled "Urls". It includes a search bar at the top right labeled "Search..." and another search bar below it labeled "Search: [empty input field]". Below these are two sets of filters: "Show 10 entries" and "Search: [empty input field]". The central part of the screen is a table with the following data:

#	Url	Status Code	Hash	Scanned
1	https://perfectioncommunication.com/wp-content/languages/plugins/	200	768d9e5879abae0b7e03e1286ed323f0ed02a759b5a525e33ce7a070bbab91ac	2021-06-21 11:06:53.174565
2	https://perfectioncommunication.com/wp-content/languages/themes/	200	fdad84e4da258948ea8fb67047f5cf80a94ec9f40f1e5176a776d588781345a8	2021-06-21 11:06:53.288654
3	https://perfectioncommunication.com/wp-content/languages/	200	6778ad2f8d8a1babf57e80d282851f97c6160927841197b2b7573bd1728f480b	2021-06-21 11:06:53.412010

At the bottom of the table, it says "Showing 1 to 3 of 3 entries". On the far right, there are buttons for "Previous", "1", and "Next".

SQLite Storage Module - Visual

- Additionally, there exists a detail view for each url and domain.
- Most of the fields can not be modified.
- It is possible to add a description to a domain



The screenshot shows a web-based application interface for 'SubCrawl'. The top navigation bar includes the 'SubCrawl' logo, a search bar with placeholder text 'Search...', and a 'Help' link. On the left, a sidebar menu lists 'Dashboard', 'Domains', and 'Urls'. The main content area is titled 'Url Information' and displays the following details for the URL `http://www.secure-mytransfer.com/lloyds/`:

- Url:** `http://www.secure-mytransfer.com/lloyds/`
- Status Code:** 200
- Title:** Registrant WHOIS contact information verification | Namecheap.com
- Sha256:** `36c6bb9432f36ddfd35039779d72e253cbc5e11fb4588ce8c4a6f084dd1d3604`
- Scanned:** 2021-06-23 12:26:34.972025
- date:** Wed, 23 Jun 2021 12:26:16 GMT

SubCrawl Service Mode Demo

MISP Storage Module

- Our SQLite user interface is rather basic and does not offer extended analysis capabilities
- MISP has a lot more functionalities and is widely used in the security community
- We did not choose MISP as default storage module because of the installation and configuration complexity
- Many MISP Docker images exist but they still need to be configured by the user

```
└─misp:  
    misp_url: https://localhost  
    misp_api_key: API_KEY_Goes_HERE  
    urlhaus_api: https://urlhaus.abuse.ch/downloads/csv_recent/  
    domain_event: 0
```

MISP Storage Module – Domain Events

Home Event Actions Dashboard Galaxies Input Filters Global Actions Sync Actions Administration Logs

Welcome! Last login was on Tue, 22 Jun 21 07:28:39 +0000

List Events

Add Event Import from... REST client

List Attributes Search Attributes

View Proposals Events with proposals View delegation requests

Export Automation

Events

« previous 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 next »

	Published	Creator org	Owner org	ID	Clusters	Tags	#Attr.	Creator user	Date	Info	Distribution	Actions
<input type="checkbox"/>	✓	Aesir	Aesir	? 3803		      	19	misp@stoerchi.ch	2021-06-23	ledsupplies.net.au	Community	  
<input type="checkbox"/>	✓	Aesir	Aesir	? 60			3737	patrick.schlapfer@hp.com	2021-05-03	Domain track event	Community	  
<input type="checkbox"/>	✓	Aesir	Aesir	? 3802		      	20	misp@stoerchi.ch	2021-06-23	coflilipense.edu.co	Community	  
<input type="checkbox"/>	✓	Aesir	Aesir	? 3801		      	17	misp@stoerchi.ch	2021-06-23	rthdimension.co.za	Community	  
<input type="checkbox"/>	✓	Aesir	Aesir	? 3800		   	233	misp@stoerchi.ch	2021-06-23	107.172.205.128	Community	  
<input type="checkbox"/>	✓	Aesir	Aesir	? 3799		      	30	misp@stoerchi.ch	2021-06-23	ibpa.cl	Community	  
<input type="checkbox"/>	✓	Aesir	Aesir	? 3798		      	20	misp@stoerchi.ch	2021-06-23	vasicomunicanti.it	Community	  
<input type="checkbox"/>	✓	Aesir	Aesir	? 3797		      	18	misp@stoerchi.ch	2021-06-23	domucmayinbacninh.net	Community	  
<input type="checkbox"/>	✓	Aesir	Aesir	? 3796		      	21	misp@stoerchi.ch	2021-06-23	poo-logix.com	Community	  

Could not locate the PGP public key.

Powered by MISP 2.4.143 - 2021-06-23 14:28:29

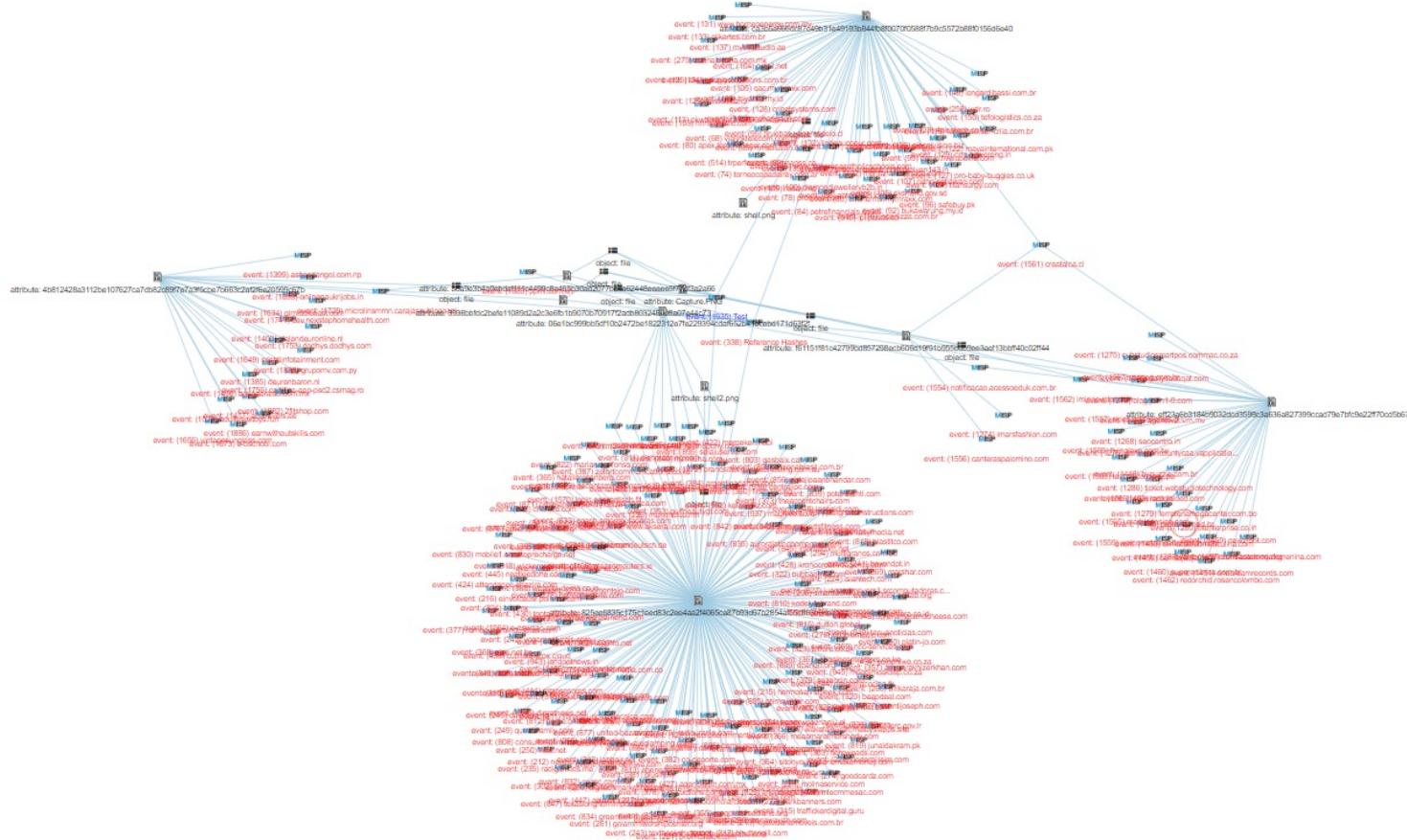
- One event is created per domain
- Events are enriched with tags taken from URLhaus like the SQLite storage module
- One defined event is used to track already scanned domains

MISP Storage Module – opendir Object

Object name: opendir-url []						
References: 0 +						
2021-06-21	Network activity	url: url	https://highend.pk/wp-content/plugins/goodlayers-core-twitter/twitteroauth/src/viewer.php			<input checked="" type="checkbox"/>
2021-06-21	Payload delivery	sha256: sha256	825ae6835c175c1eed83c2ee4aa2f4065ca87b93d97b2854af55c863b0dec ddc			<input checked="" type="checkbox"/> 211 212 213 215 Show 262 more...
2021-06-21	Other	title: text	None			<input checked="" type="checkbox"/>
2021-06-21	Other	status-code: text	200			<input type="checkbox"/>
2021-06-21	Other	header: text	Mon, 21 Jun 2021 22:36:20 GMT			<input type="checkbox"/> Date
2021-06-21	Other	header: text	Apache			<input type="checkbox"/> Server
2021-06-21	Other	header: text	timeout=5, max=100			<input type="checkbox"/> Keep-Alive
2021-06-21	Other	header: text	Keep-Alive			<input type="checkbox"/> Connection
2021-06-21	Other	header: text	chunked			<input type="checkbox"/> Transfer-Encoding
2021-06-21	Other	header: text	text/html; charset=UTF-8			<input type="checkbox"/> Content-Type
2021-06-21	Other	tlsh: text	T195B012003423CE20470D1034D7C25E190958F318610248842004456 666E849583B0FC4			<input checked="" type="checkbox"/> 449 802 803 804 Show 157 more...
2021-06-21	Other	yara: text	protected_webshell			<input checked="" type="checkbox"/> 3160 3255 3258 3259 Show 107 more...

- Each scanned URL is combined into one MISP object with its attributes
- MISP displays all related events for each attribute
- A related event also contains such an attribute
- E.g., interesting for found webshell logins as displayed in image

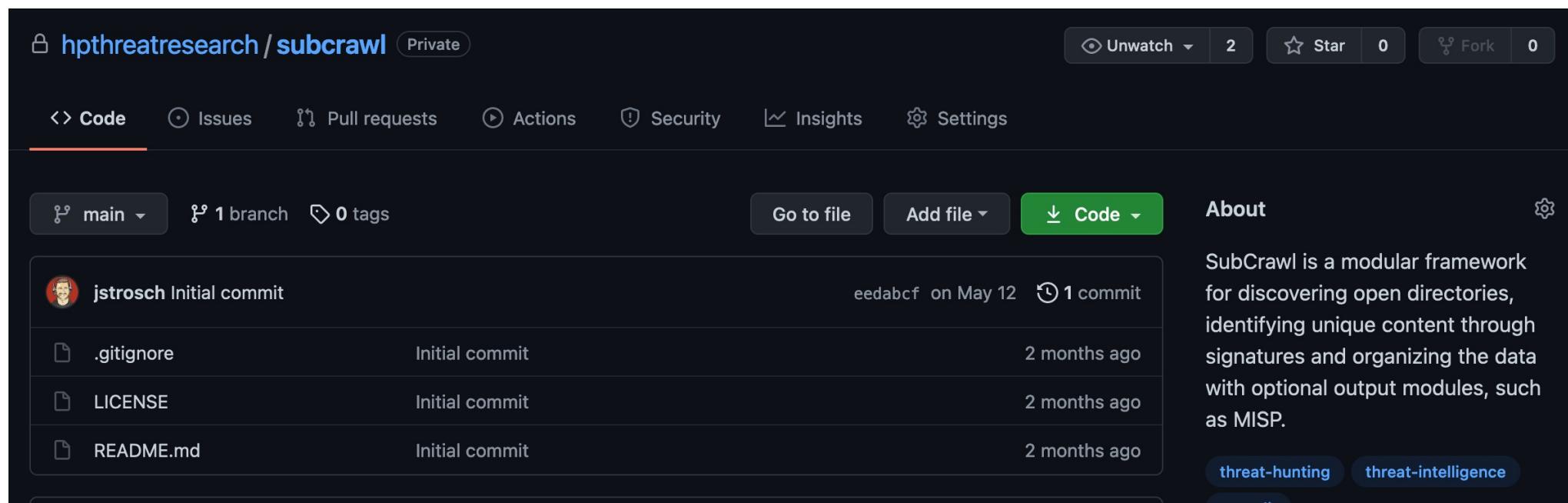
MISP Storage Module – Correlation Graph



- MISP offers the out of the box functionality to create correlation graphs
- With a large number of scanned domains, the clustering of used web servers, scripting technologies or found hashes can be interesting
- However, one limitation of MISP is that clustering based on similarity hashes is not possible
- Scripts for this can be found directly in our GitHub repository

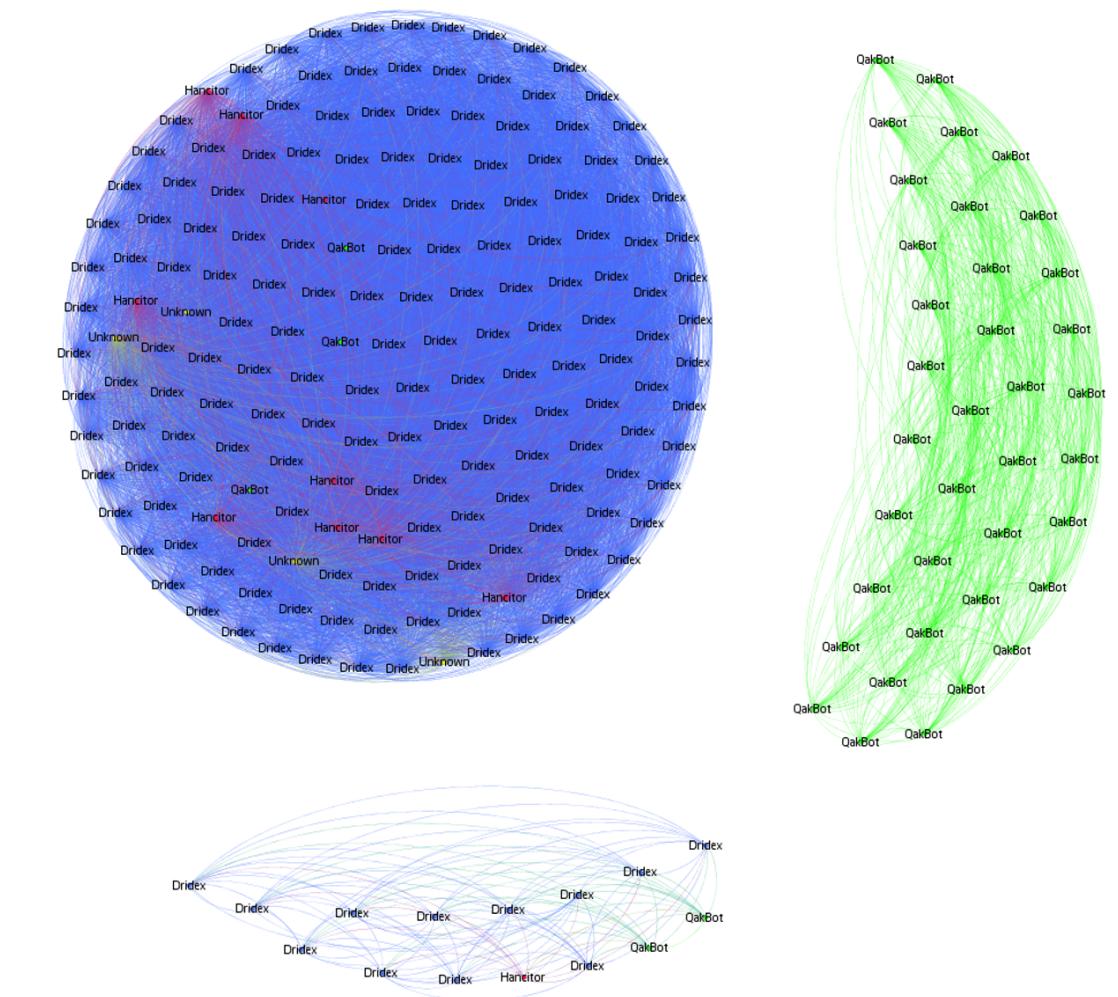
SubCrawl is Now Available!

- Github - <https://github.com/hpthreatresearch/subcrawl>
- @jstrosch @stoerchl @cryptogramfan



Upcoming Events

- Virus Bulletin Localhost – Oct 2021
- <https://vbllocalhost.com/presentations/introducing-subcrawl-a-framework-for-the-analysis-and-clustering-of-hacking-tools-found-using-open-directories/>
- Focus on data that we discovered, patterns in threat actor activity and how we were able to track tools such as web shells



Prevalent Dridex webshells on compromised hosts



Thank You!!