MISP Web Scraper

CTIS - October 19/20 - 2022

cudeso.be
We Secure You
https://www.cudeso.be
koen.vanimpe@cudeso.be

Koen Van Impe

- **Freelancer**
  - Incident response, threat intelligence, security monitoring
- **Open source contributions**
  - MISP modules, taxonomies, automation and integration with DFIR tools, ...
  - "*MISP tip-of-the-week*"
- **BelgoMISP**
  - Belgian MISP User Group
- **OSINT threat feed**
  - botvrij.eu

koen.vanimpe@cudeso.be

https://www.cudeso.be

https://www.vanimpe.eu

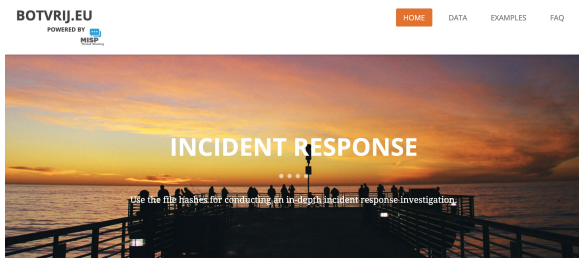
https://github.com/cudeso

@cudeso

Why?

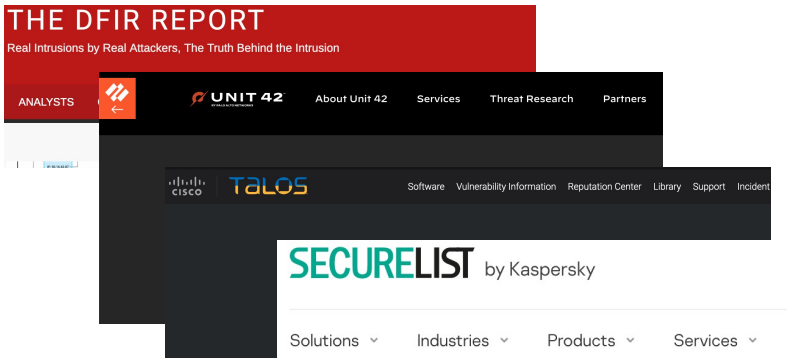
How it started: adding threat events to botvrij.eu

Public reports | Blog posts | Whitepapers | “OSINT”



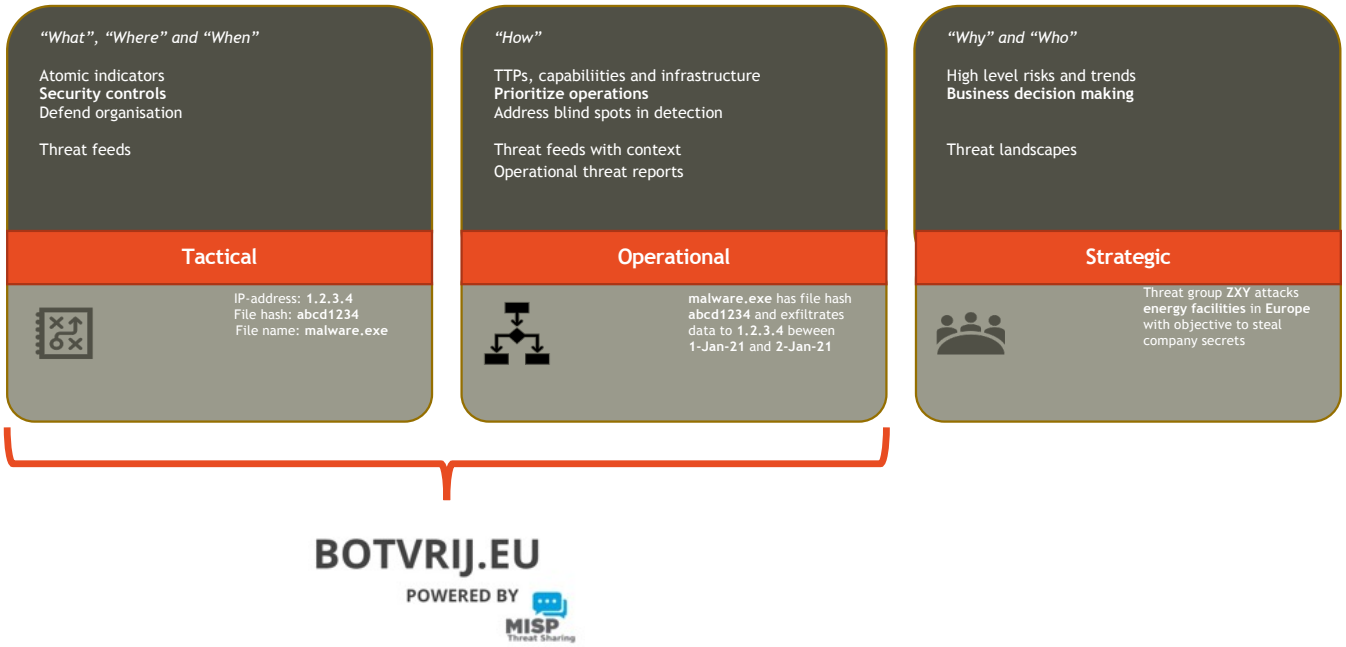
osint:source-type="blog-post" x | tlp:white x | type:OSINT x

- A lot of **copy and pasting**
  - Tedious
  - Time consuming
  - Risk of errors

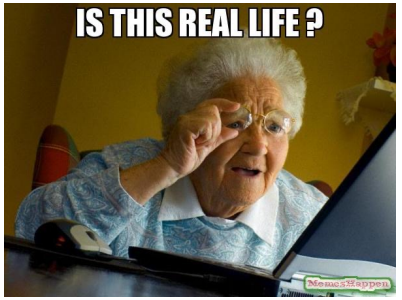


Avoid being just an indicator list

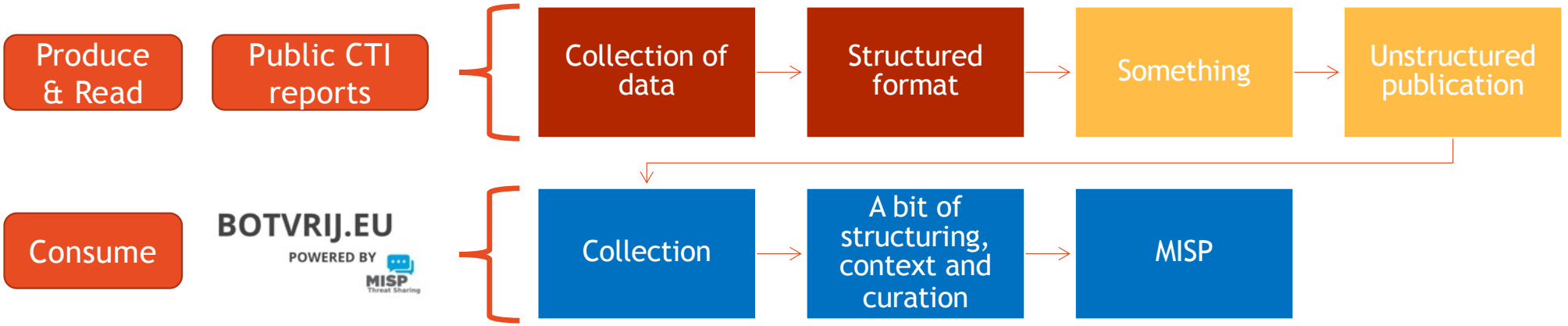
- Focus on **tactical** and **operational** intelligence
- Include context for events and attributes



Sources used for the OSINT feed

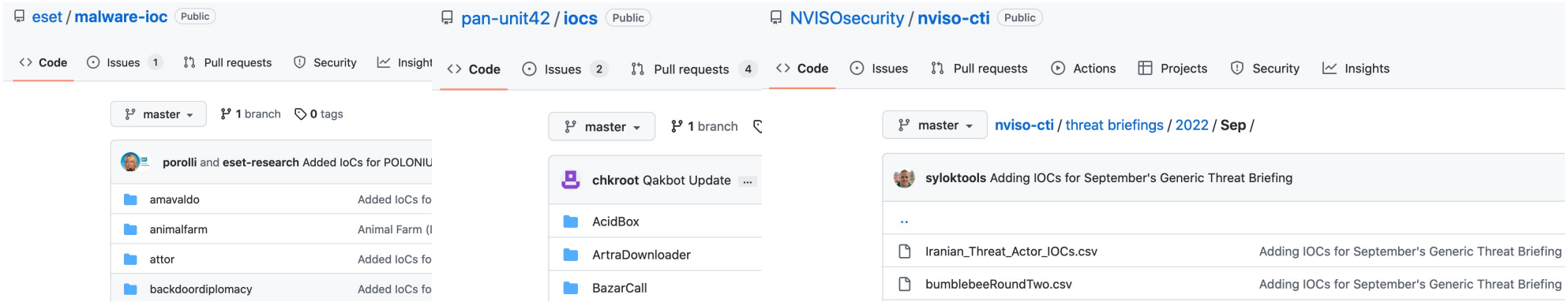


- A lot of public reports *mostly* contain **unstructured** information
  - Or structured but not consistent in format
    - CSV in format x, CSV in format y, TXT files
- Reports sometimes **focus more on readers**, than on consumers
- Workflow (presumably not just for me ...)



Why are public sources so difficult to consume?

- It's not about *solving the sharing problem*
  - But about **making public reports more easily consumable**
- There are commercial interests
  - But providing structured information publicly can be a valuable "**teaser**" for your professional CTI services
- Not all is bad



What would make this work easier?

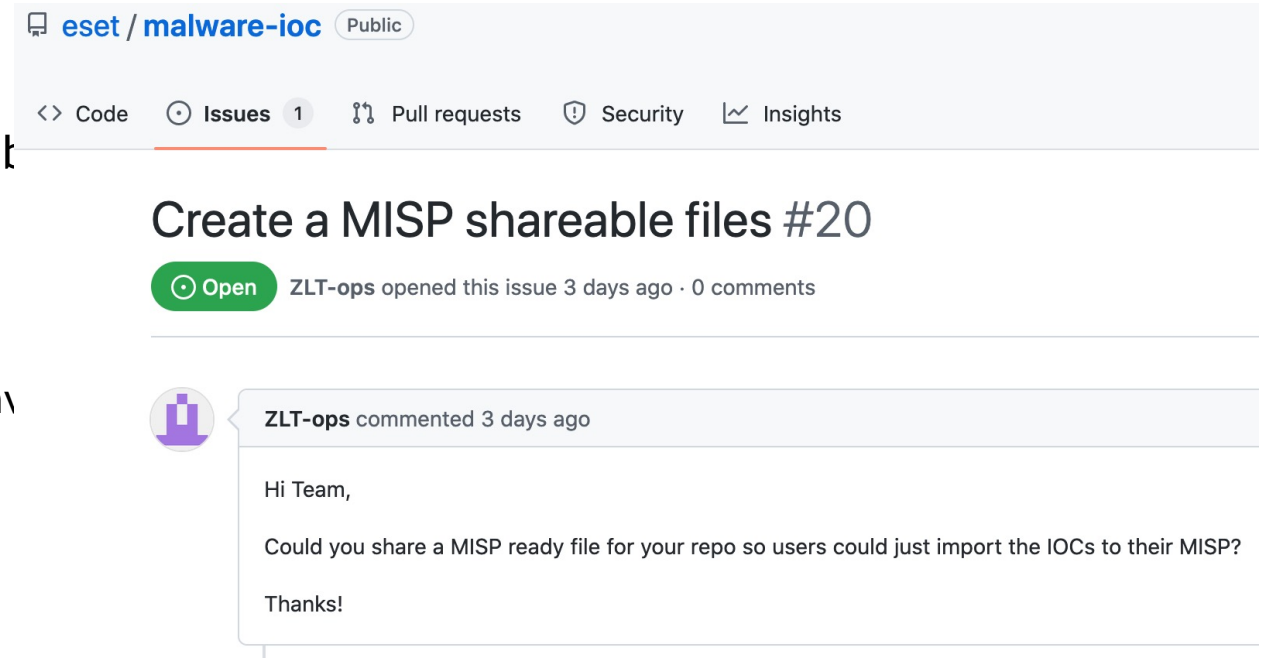
- Formats?
  - **STIX**
    - Low adoption
    - Certainly what concerns public reports
  - **CSV/TXT**
    - Not always consistent
      - Field sequence, naming conventions
  - **PDF/XLS/...**
    - No

What would make this work easier?

- Formats?
  - **STIX**
    - Low adoption
    - Certainly what concerns public reports
  - **CSV/TXT**
    - Not always consistent
      - Field sequence, naming conventions
  - **PDF/XLS/...**
    - No
  - Pending everyone starts using the **MISP core format** ...

What would make this work easier?

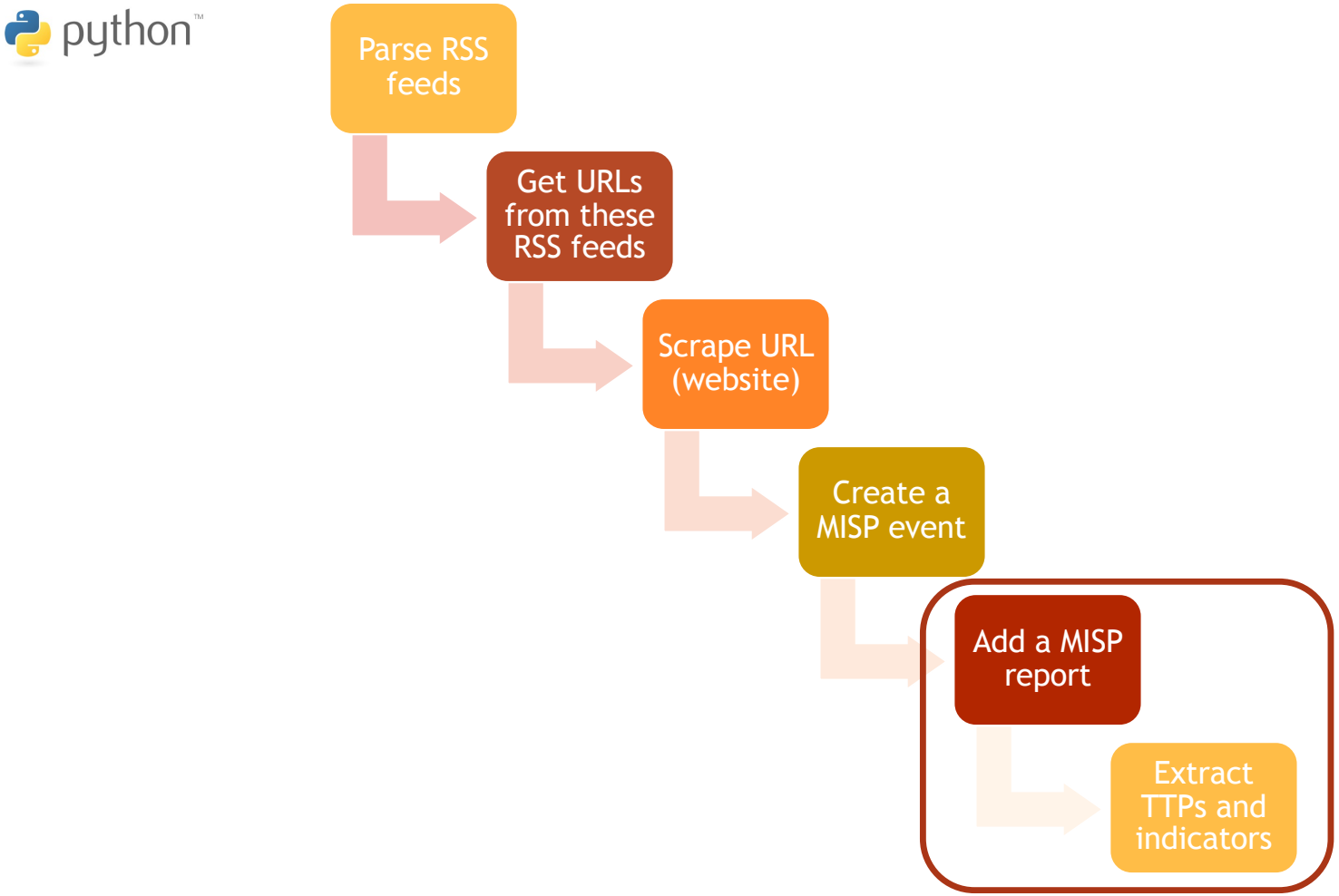
- Formats?
  - **STIX**
    - Low adoption
    - Certainly what concerns pub
  - **CSV/TXT**
    - Not always consistent
      - Field sequence, naming conv
  - **PDF/XLS/...**
    - No
  - Pending everyone starts using the **MISP core format** ...



MISP web scraper

MISP web scraper

MISP web scraper



Build on MISP reports

"CTI report"

References to attributes and objects

References to taxonomies and galaxy matrices

In raw text or markdown

Is part of an event and distributed with events

Build on MISP reports

"CTI report"	References to attributes and objects	References to taxonomies and galaxy matrices	In raw text or markdown	Is part of an event and distributed with events
1	Indicators	TTPs		

Build on MISP reports

"CTI report"

References to attributes and objects

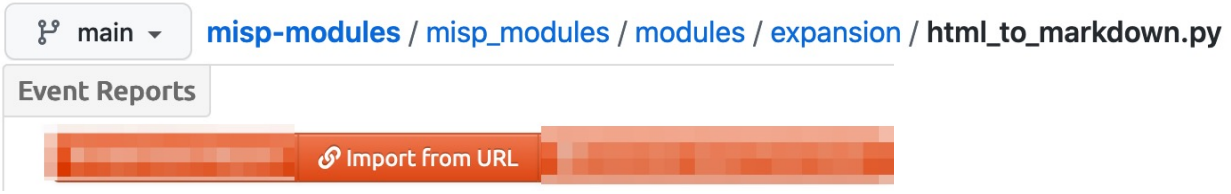
References to taxonomies and galaxy matrices

In raw text or markdown

Is part of an event and distributed with events

1. Indicators
2. TTPs

2. Convert “website” content to markdown



Build on MISP reports

"CTI report"

References to attributes and objects

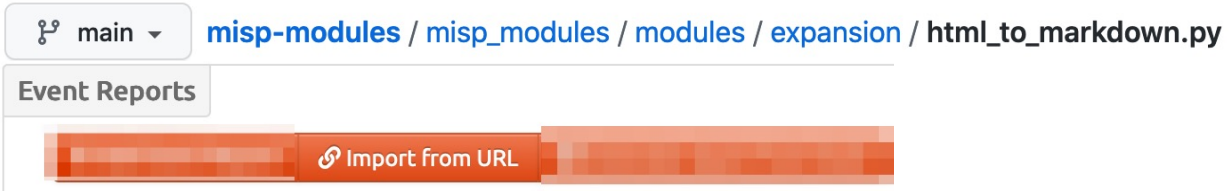
References to taxonomies and galaxy matrices

In raw text or markdown

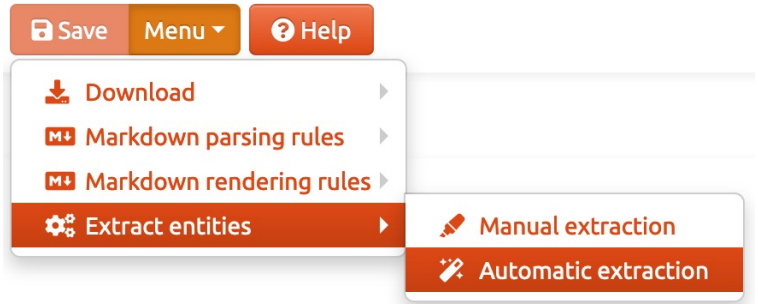
Is part of an event and distributed with events

1. Indicators
   TTPs

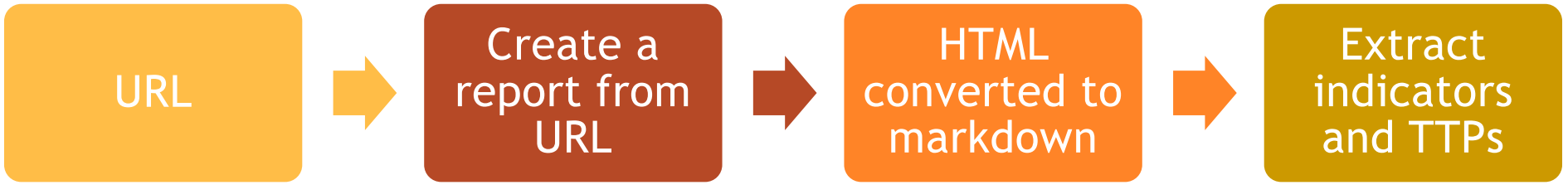
2. Convert “website” content to markdown



3.



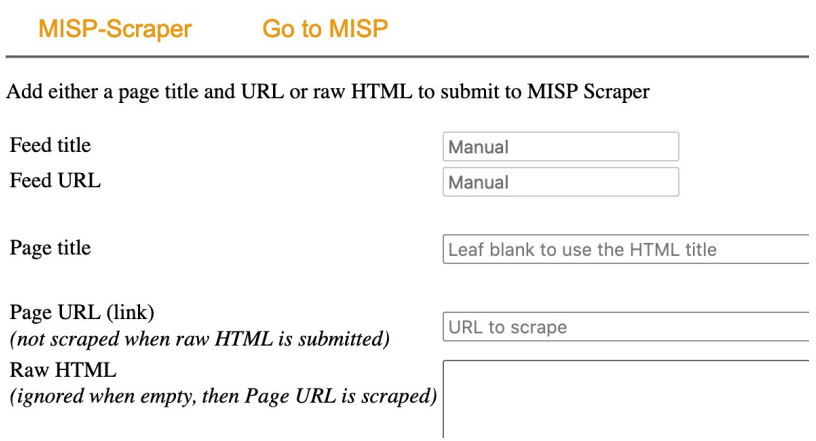
Easy wins with the MISP report feature

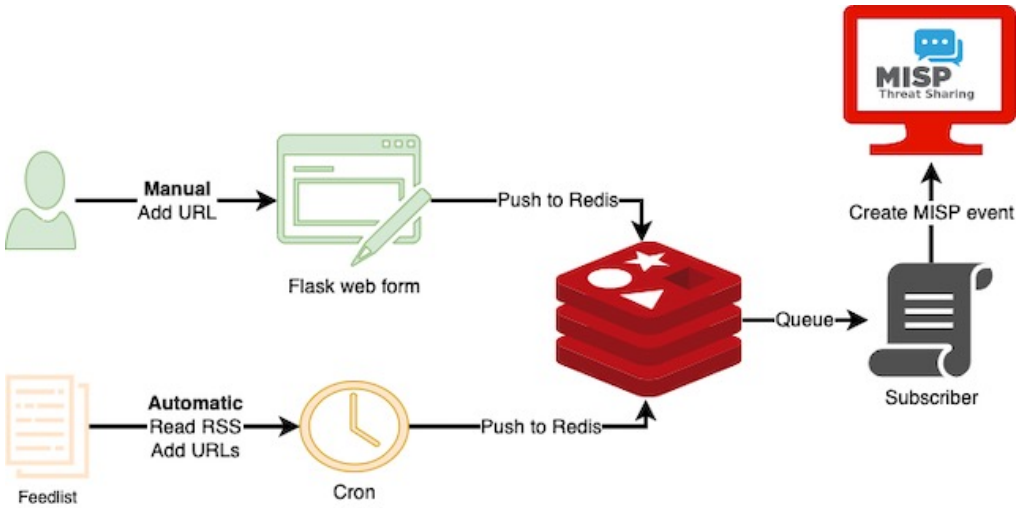


MISP web scraper components

Flask

- Web form
- **Manual:** add a URL to Redis

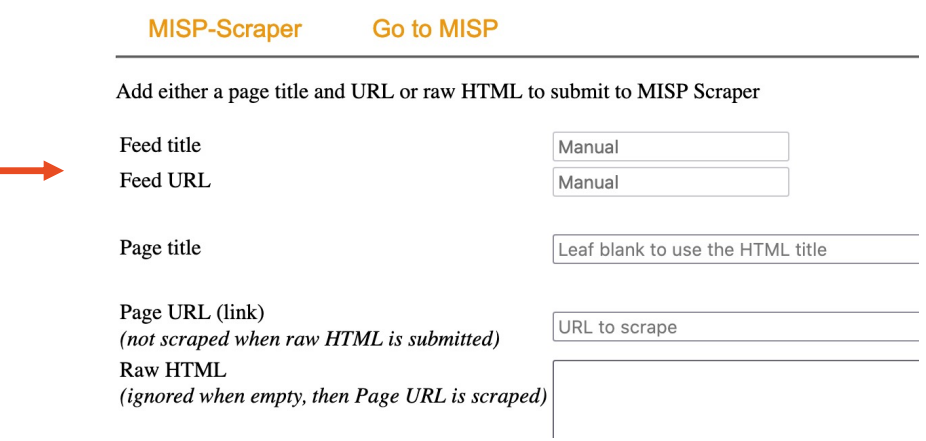




MISP web scraper components

Flask

- Web form
- **Manual:** add a URL to Redis



Cron job

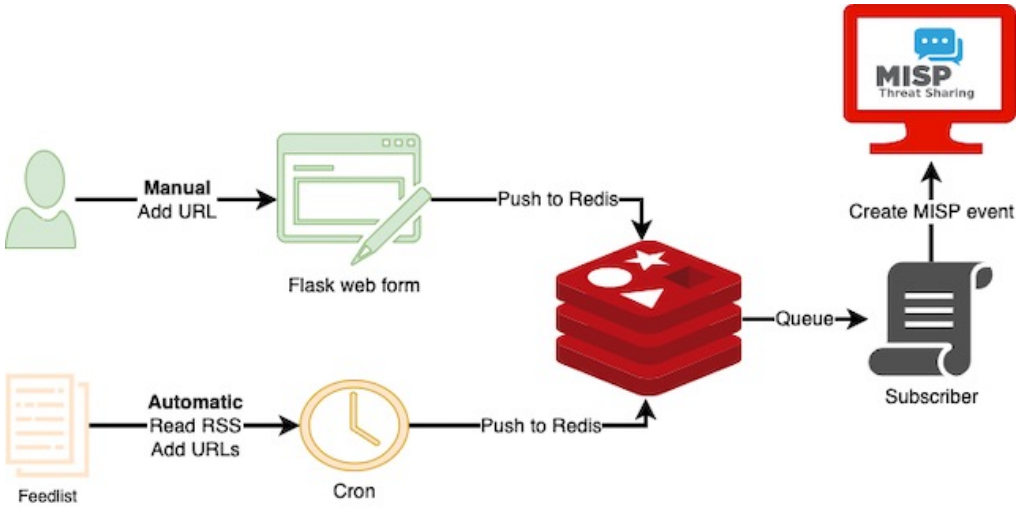
- Refresh RSS feed list (*feedlist.json*)
- **Automatic:** add URLs to Redis

Redis

- Push URLs to Redis

Subscribe to Redis queue

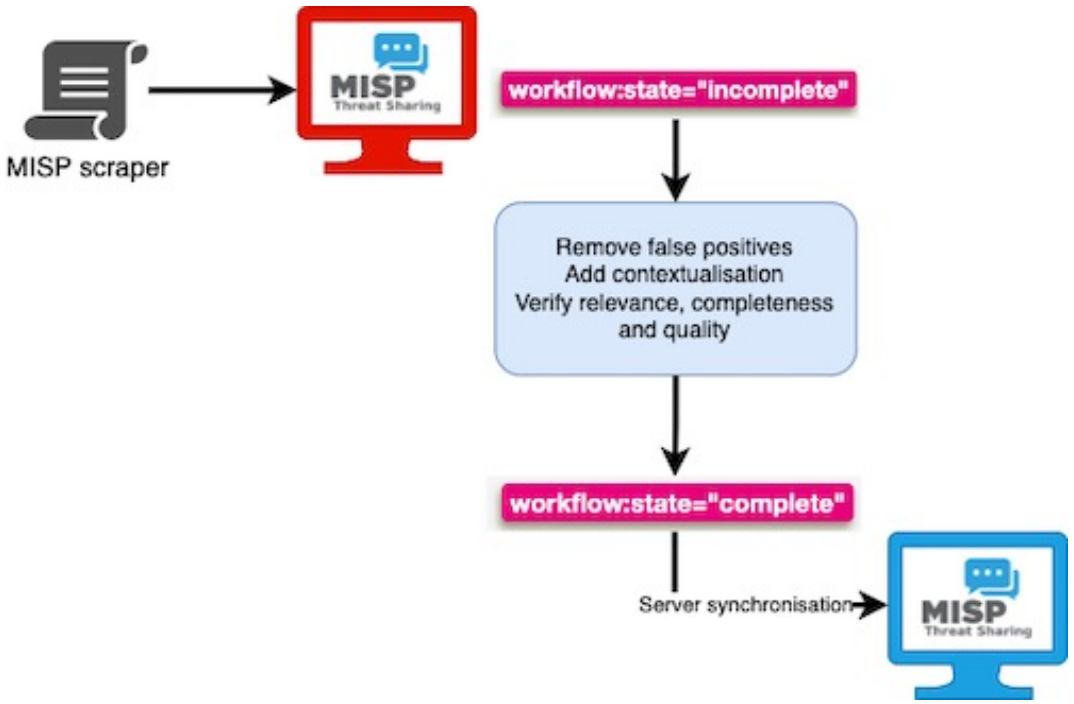
- Pick up URLs, crawl the website
- **Create MISP event and report**



Web scraper workflow

- **Scrape** in a “dirty” environment
  - Automated, “cron job”
  - Create MISP event

workflow:state="incomplete" x



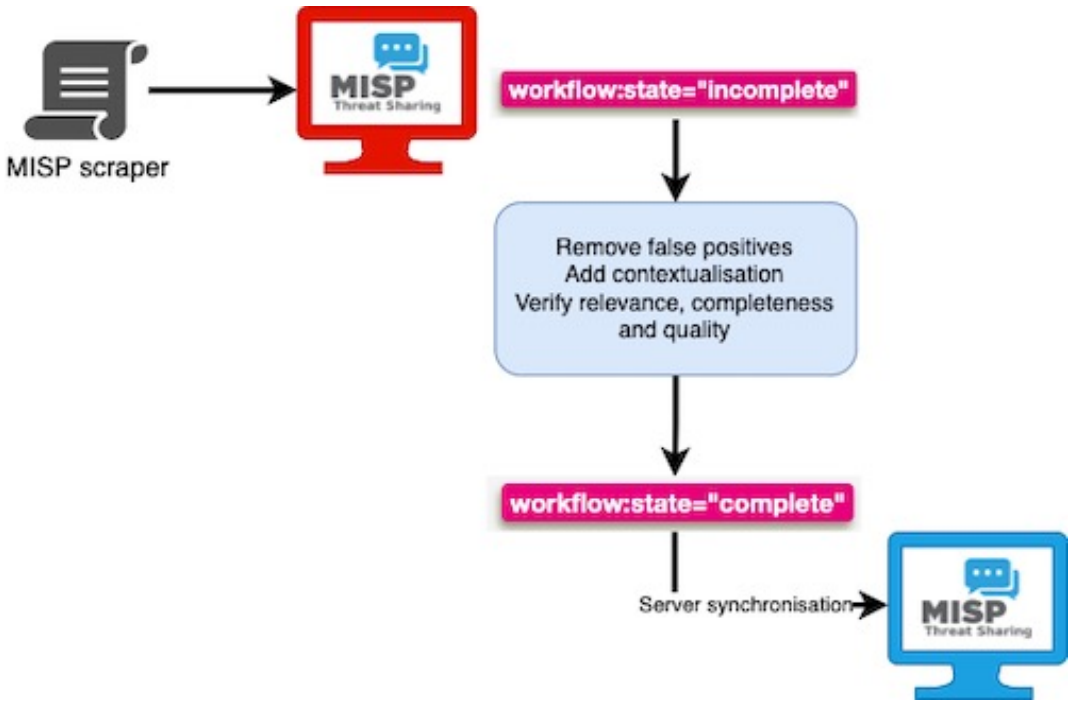
Web scraper workflow

- **Scrape** in a “dirty” environment
  - Automated, “cron job”
  - Create MISP event

workflow:state="incomplete" x

- **Curate** event
  - Scrape successful?
  - Relevant and complete?
  - Remove false positives

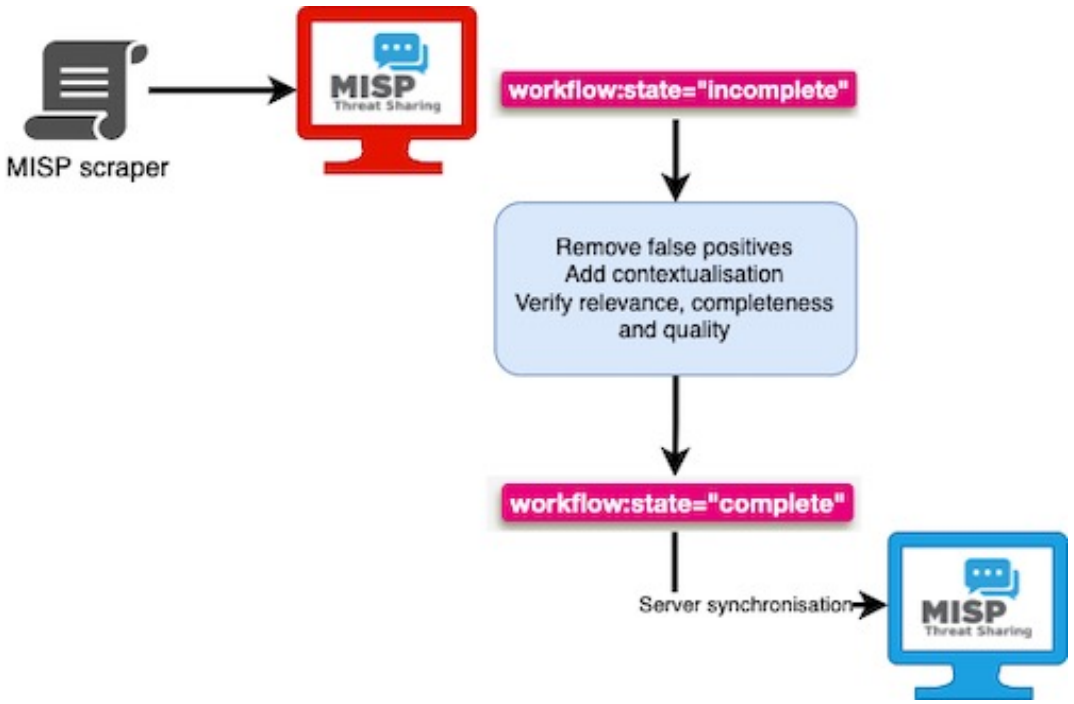
workflow:state="complete" x



Web scraper workflow

- **Scrape** in a “dirty” environment
  - Automated, “cron job”
  - Create MISP event

workflow:state="incomplete"



- **Curate** event
  - Scrape successful?
  - Relevant and complete?
  - Remove false positives

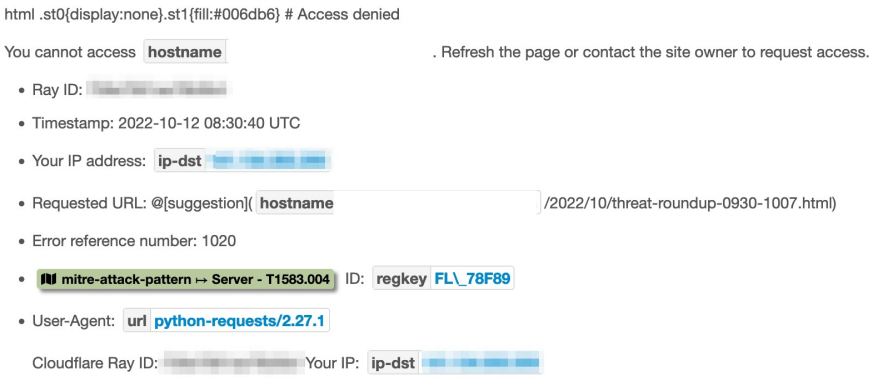
workflow:state="complete"

- Publish and **synchronise**
  - With your “clean” MISP

Curate scraped events (1/2)

- Analyst checks
  - Scrape **successful**?
    - HTTP error codes added as tags
    - Not “blocked” by a CDN

misp-scraper:HTTP=403
misp-scraper:HTTP=404



- **Relevant**?
  - “Threat report”?

Date	Org	Category	Type	Value
2022-10-12		Other	comment	Preventing Cryptocurrency Cyber Extortion
2022-10-12		External analysis	link	http://feeds.trendmicro.com/TrendMicroSimplySecurity
2022-10-12		External analysis	link	https://www.trendmicro.com/en_us/ciso/22/i/prevent-cyber-extortion.html

Page 1 of 1, showing 1 records out of 3 total, starting on record 1, ending on 3

- Wrong attribute **category** or **types**

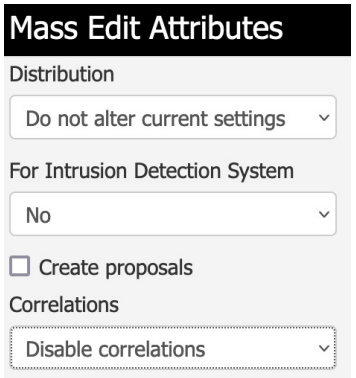
Payload delivery hostname 2022-09-23-iocs-for-icedid-and-cobalt-strike.txt.zip

Curate scraped events (2/2)

- **False positives**
  - MISP warninglists

- **Mass edit** attributes
  - Remove to_ids
    - Attributes primarily fit for context, not for detection or filtering
  - Remove correlation

- Mass **delete** attributes
  - Attributes irrelevant for the threat event



Payload delivery	filename	Trojan.BAT.ADFIND.YECGUT
Payload delivery	filename	Backdoor.Win32.SYSTEMBC.YXCFLZ
Payload delivery	filename	Backdoor.Win32.COBEACON.YXCH3
Payload delivery	filename	Ransom.Win32.PLAYCRYPT.YECGUT
Payload delivery	filename	Trojan.Win64.PRIVICMD.YXCHW
Payload delivery	filename	Ransom.Win32.PLAYDE.YACHWT

Make curation a bit easier (1)

- **Doubles**
  - Web scraper: **skip MISP events** already created
    - (based on event title + scraper tag)

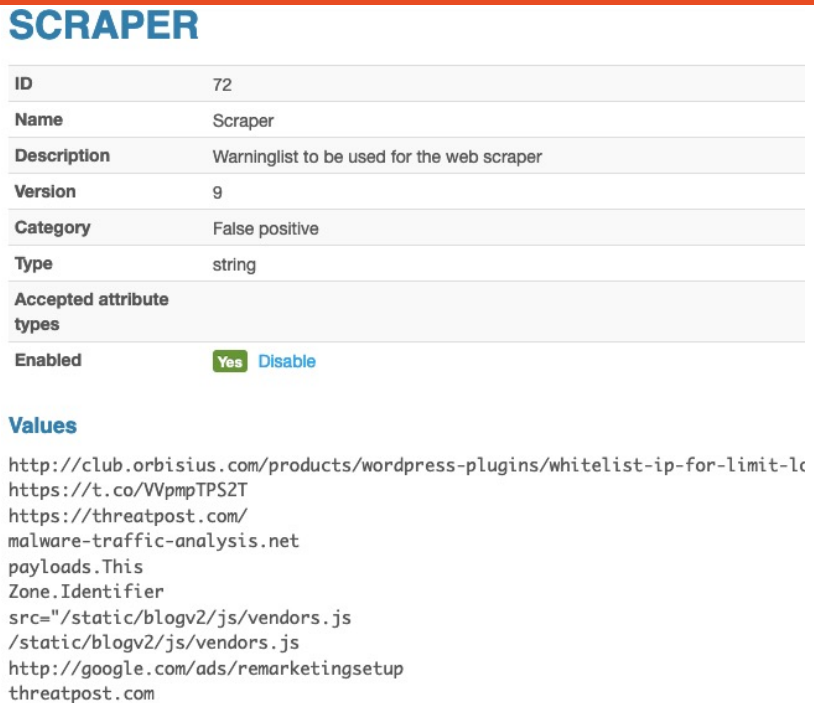
- **Retention**
  - Old events are automatically deleted
  - Avoid analyst overload

- **Classification** of scraped events

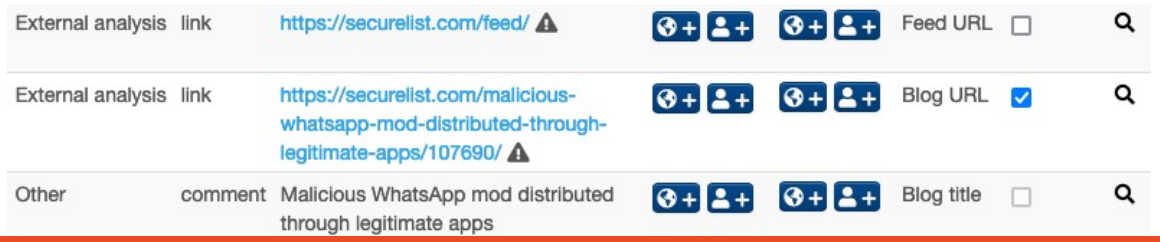
  - Tool and source
    - **Unique tag per source**
    - Allows f.e. to filter for all posts scraped from the source TrendMicro
  - Default TLP and workflow state

Make curation a bit easier (2)

- Custom **warninglist**
  - Via a MISP warninglist
  - Avoid adding known non-relevant attributes
    - URLs to partner websites
    - Social media
    - “Commercial” strings



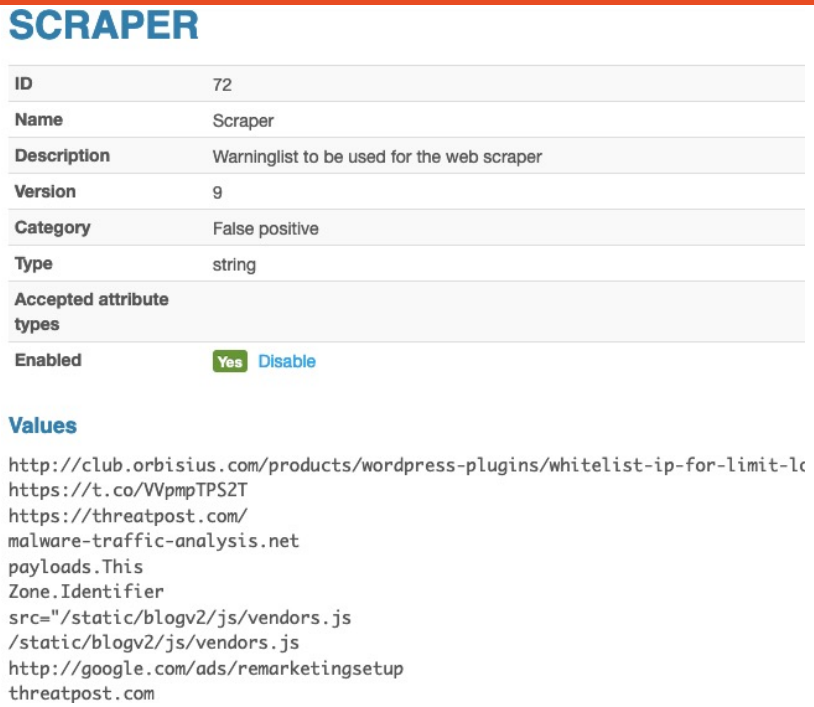
- Links and comment attributes to describe source
  - Feed, title
    - No correlation
  - Link to web page
    - Correlation



Make curation a bit easier (2)

- Custom **warninglist**
  - Via a MISP warninglist
  - Avoid adding known non-relevant attributes
    - URLs to partner websites
    - Social media
    - “Commercial” strings

Curation: events with less than 4 attributes can be ignored



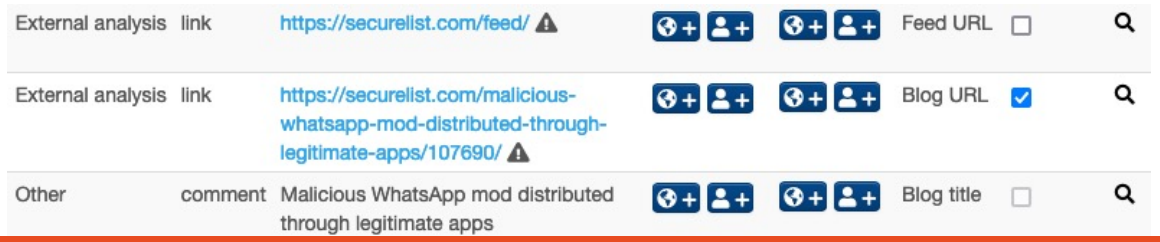
SCRAPER

ID	72
Name	Scraper
Description	Warninglist to be used for the web scraper
Version	9
Category	False positive
Type	string
Accepted attribute types	
Enabled	Yes Disable

Values

```
http://club.orbisius.com/products/wordpress-plugins/whitelist-ip-for-limit-lo
https://t.co/VVpmpTPS2T
https://threatpost.com/
malware-traffic-analysis.net
payloads.This
Zone.Identifier
src="/static/blogv2/js/vendors.js
/static/blogv2/js/vendors.js
http://google.com/ads/remarketingsetup
threatpost.com
```

- Links and comment attributes to describe source report
  - Feed, title
    - No correlation
  - Link to web page
    - Correlation



External analysis	link	https://securelist.com/feed/		Feed URL	☐
External analysis	link	https://securelist.com/malicious-whatsapp-mod-distributed-through-legitimate-apps/107690/		Blog URL	☑
Other	comment	Malicious WhatsApp mod distributed through legitimate apps		Blog title	☐

Scraped event

Preventing Black Basta Ransomware in 2022

Event ID	294
UUID	391357f3-d319-41f8-8525-ce2b083097b6
Creator org	CUDESO
Owner org	CUDESO
Creator user	
Protected Event (experimental)	
Tags	misp:tool="misp-scraper" x osint:source-type="blog-post" x misp:event-type="collection" x misp-galaxy:botnet="Qbot" x misp-galaxy:malpedia="Black Basta" x workflow:state="complete" x tlp:white x
Date	2022-09-12

Tags from scraper to ease curation

Scraped event

Preventing Black Basta Ransomware in 2022

Event ID	294
UUID	391357f3-d319-41f8-8525-ce2b083097b6
Creator org	CUDESO
Owner org	CUDESO
Creator user	
Protected Event (experimental)	
Tags	misp:tool="misp-scraper" x osint:source-type="blog-post" x misp:event-type="collection" x misp-galaxy:botnet="Qbot" x misp-galaxy:malpedia="Black Basta" x workflow:state="complete" x tlp:white x
Date	2022-09-12

MISP report

Event Reports

+ Add Event Report | Import from URL | Generate report from Event | All | Default | Deleted

ID	Name	Last update	Distribution	Actions
16	Report from - https://www.deepinstinct.com/blog/black-basta-ransomware-threat-emergence (1662992420)	2022-09-12 16:20:20	Inherit event	

Tags from scraper to ease curation

2022-09-12	Network activity	url	https://aazsbsgya565vlu2c6bzy6yfiebkcbtvvcytvolt33s77xypi7nypxyd.onion/	Black Basta TOR site			Inherit
2022-09-12	Payload delivery	sha256	9fce9ee85516533bae34fc1184a7cf31fa9f2c7889b13774f83d1df561708833	Black Basta Windows Variant			Inherit
2022-09-12	Payload delivery	sha256	203d2807df6ef531efbec7bfd109986de3e23df64c01ea4e337cbe5ba675248b	Black Basta Windows Variant			Inherit
2022-09-12	Payload delivery	sha256	0d6c3de5aebbbe85939d7588150edf7b7bdc712fceb6a83d79e65b6f79bfc2ef	Black Basta Linux/ESXi Variant			Inherit
2022-09-12	Payload delivery	sha1	2bee3f716b80273db9639376a296cf19cdba0f1a	Qbot payload certificate thumbprint (sha1)			Inherit
2022-09-12	Network activity	url	http://146.70.79.52/	Payload host			Inherit

Extracted attributes

Scraped event

Ragnar Locker Ransomware Targeting the Energy Sector

Event ID	292
UUID	8dbeaaac-a671-4a02-8dab-5eec2a1c935b
Creator org	CUDESO
Owner org	CUDESO
Creator user	koen.vanimpe@cudeso.be
Protected Event (experimental)	Event is in unprotected mode.
Tags	misp:tool="misp-scraper" x osint:source-type="blog-post" x misp:event-type="collection" x misp-galaxy:malpedia="RagnarLocker (Windows)" x workflow:state="complete" x tlp:white x
Date	2022-09-12

Scraped event

Ragnar Locker Ransomware Targeting the Energy Sector

Event ID	292
UUID	8dbeaaac-a671-4a02-8dab-5eec2a1c935b
Creator org	CUDESO
Owner org	CUDESO
Creator user	koen.vanimpe@cudeso.be
Protected Event (experimental)	Event is in unprotected mode.
Tags	misp:tool="misp-scraper" x osint:source-type="blog-post" x misp:event-type="collection" x misp-galaxy:malpedia="RagnarLocker (Windows)" x workflow:state="complete" x tlp:white x
Date	2022-09-12

Event Reports

+ Add Event Report | Import from URL | Generate report from Event | All | Default | Deleted

ID	Name	Last update	Distribution	Actions
14	Report from - https://www.cybereason.com/blog/threat-analysis-report-ragnar-locker-ransomware-targeting-the-energy-sector (1662984886)	2022-09-12 14:14:47	Inherit event	

2022-09-12	Payload delivery	filename	ntuser.dat.log				☑	☐
2022-09-12	Payload delivery	filename	bootfront.bin				☑	☐
2022-09-12	Payload delivery	sha256	9b62cdb57f4c34924333dfa3baefd993efeab68109580b682b074f0e73b63983			Ragnar Locker Binary	☑	☑
2022-09-12	Payload delivery	sha256	c2bd70495630ed8279de0713a010e5e55f3da29323b59ef71401b12942ba52f6			Ragnar Locker Binary	☑	☑
2022-09-12	Payload delivery	sha256	5469182495d92a5718e0e1dcdf371e92b79724e427050154f318de693d341c89			Ragnar Locker Binary	☑	☑
2022-09-12	Payload delivery	sha256	ec35c76ad2c8192f09c02eca1f263b406163470ca8438d054db7adcf5bfc0597			Ragnar Locker Binary	☑	☑
2022-09-12	Payload delivery	sha256	dd5d4cf9422b6e4514d49a3ec542cffb682be8a24079010cda689afbb44ac0f4			Ragnar Locker Binary	☑	☑
2022-09-12	Payload delivery	sha256	cf5ec678a2f836f859eb983eb633d529c25771b3b7505e74aa695b7ca00f9fa8			Ragnar Locker Binary	☑	☑
2022-09-12	Payload delivery	sha256	ce33096639fb5c51684e9e3a7c7c7161884ecad29e8d6ad602fd8be42076b8d4			Ragnar Locker Binary	☑	☑

What works well?

Always

- Hashes
- IP address
- Easy to identify with regexp

Often

- URLs
- Domain and hostname
- E-mail src
- TTPs
  - If not part of the commercial text on web pages

Sometimes

- Filenames
- Registry keys

- Good for indicator collection
- Some automatic TTPs
- Converting to objects remains a manual process
- Detection rules (yara) require manual edition

- **RSS reader for threat reports**

Installation and configuration

Requirements

- MISP **modules** installed and enabled
- Enable HTML to markdown module

Installation

- Setup Python **virtual environment**
- Create a MISP API key

Configuration

- **'scraper.py**
- Skelet config file in GitHub 'scraper.py.default'

```
git clone https://github.com/cudeso/misp-scraper
cd misp-scraper
virtualenv scraper
source scraper/bin/activate
pip install -r requirements.txt
cp scraper.py.default scraper.py
```

Configuration file

```
# General configuration
misp_scraper_log = "/var/log/misp/scraper.log"                  # Log file for the scraper
app_name = "MISP-Scraper"                                       # Used in the Flask server
manual_feed = "Manual"
manual_feedsource = "Manual"
flask_secret_key = "MyZsuePerSecret1984key"                     # Flask secret key. Change this to
flask_address = "127.0.0.1"                                     # IP address for the Flask server
flask_port = 5200
flask_certificate_file = "/etc/ssl/private/misp.local.crt"      # Point to a certificate file (fe.
flask_certificate_keyfile = "/etc/ssl/private/misp.local.key"   # Point to a certificate key file
logging_level = "error"
feedlist = "/home/ubuntu/misp-scraper/feedlist.json"            # Location of the RSS feedlist
rawhtml_distribution = 4
rawhtml_sharing_group_id = 4

# MISP configuration
misp_key = ""                                                   # MISP API key
misp_url = ""                                                   # MISP url
misp_verifycert = False
misp_distribution = 2                                           # Connected communities
misp_threat_level_id = 2                                        # Medium
misp_analysis_level = 2                                         # Completed
misp_scraper_event = "Scraper"                                  # Prefix of the MISP event
misp_retentiontime = "99d"                                      # Set this to >days than URLs are
misp_scraper_tags = ["misp:tool=\"misp-scraper\"", "osint:source-type=\"blog-post\"", "misp:event-
misp_warninglist = 0
misp_hard_delete_on_cleanup = False
```

Feedlist configuration file

JSON

List of feeds

- title
  - Used in the tagging of events
- url
  - URL to download the RSS feed

```
{
        "feeds": [
                {
                        "title": "Proofpoint",
                        "url": "https://www.proofpoint.com/us/threat-insight-blog.xml"
                },
                {
                        "title": "Red Canary",
                        "url": "https://www.redcanary.co/feed/"
                },
                {
                        "title": "vanimpe.eu",
                        "url": "https://www.vanimpe.eu/feed/rss/"
                },
                {
                        "title": "Threatpost",
                        "url": "http://threatpost.com/feed"
                }
        ]
}
```

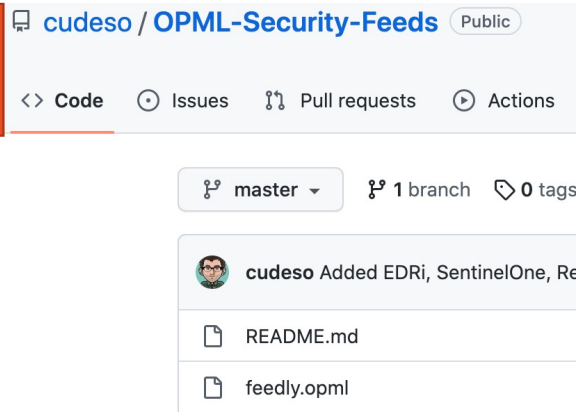
Feedlist configuration file

JSON

List of feeds

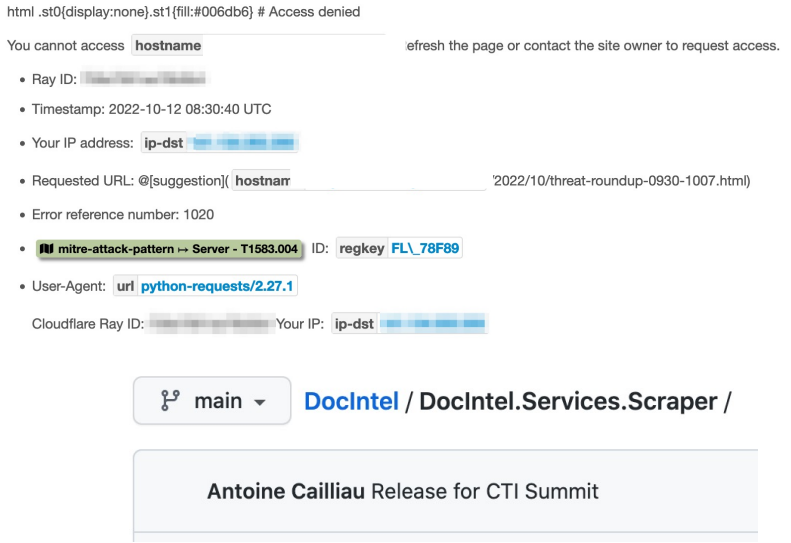
- title
  - Used in the tagging of events
- url
  - URL to download the RSS feed

- Need inspiration for RSS feeds?
  - https://github.com/cudeso/OPML-Security-Feeds



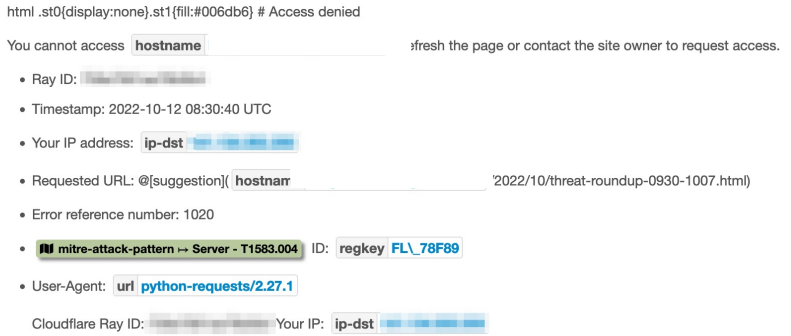
Room for improvement (1)

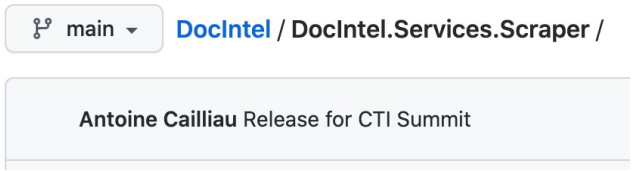
- Web crawler is **Python**
  - User-agent blocked
  - Improve with **ail-splash-manager**
    - Instead of handling the crawling in web-scraper, have AIL do the crawling
    - Microsoft Playwright ^^AIL



Room for improvement (1)

- Web crawler is **Python**
  - User-agent blocked
  - Improve with **ail-splash-manager**
    - Instead of handling the crawling in web-scraper, have AIL do the crawling
    - Microsoft Playwright ^^AIL
  - First steps done by allowing to submit raw HTML





- Include **screenshots**
  - Via ail-splash-manager?

- Deal with RSS sources containing ridiculous number of URLs
  - Feeds containing all articles going back to 20xx
  - Discover publish date and only consider recent reports

Room for improvement (2)

- Use Natural Language Processing (**NLP**) to analyse reports
  - FIRSTCON 2022

- **Bulk change category and type** of attributes
  - Similar as to bulk edit events
  - Would in general be a great feature to edit MISP events

- Use Feedly AI
  - Remove non-relevant clutter text from web pages

Room for improvement (3)

- Automatically **remove attributes** for specific scraped sources
  - F.e. always remove attribute xyz from source abc
    - Company labels etc.
  - Remove specific **CSS classes**

Room for improvement (3)

- Automatically **remove attributes** for specific scraped sources
  - F.e. always remove attribute xyz from source abc
    - Company labels etc.
  - Remove specific **CSS classes**
    - Need galaxies/clusters?

Room for improvement (3)

- Automatically **remove attributes** for specific scraped sources
  - F.e. always remove attribute xyz from source abc
    - Company labels etc.
  - Remove specific **CSS classes**
    - Need galaxies/clusters?
    - **Mass delete clusters**

```
<div class="social-copyright-row">
    <div class="inner-container">
        <div class="row">
            <ul class="col-md-6 social-media-links">
```

```
<div class="dropdown stretched-dropdown">
    <button class="menu-button button-default" type="button" data-toggle
        <span class="menu-button__icon icon-region"></span>
        <span class="menu-button__text">Region</span>
    </button>
        <div class="dropdown-menu align-">
            <ul class="menu-column col-xs-12 col-sm-4 col-md-3">
                <li class="dropdown-header">
                    The Americas
                </li>
                <li>
                    <a href="/en_us.html">
                        United States
                    </a>
                </li>
                <li>
                    <a href="/pt_br.html">
                        Brasil
                    </a>
                </li>
```

Misinformation Pattern

- Instagram
- Facebook
- Twitter

Target Information

- Australia
- Canada
- France
- Hong Kong
- Indonesia
- Ireland
- Malaysia
- New Zealand
- Philippines
- Singapore
- South Africa
- United Kingdom
- United States

Sector

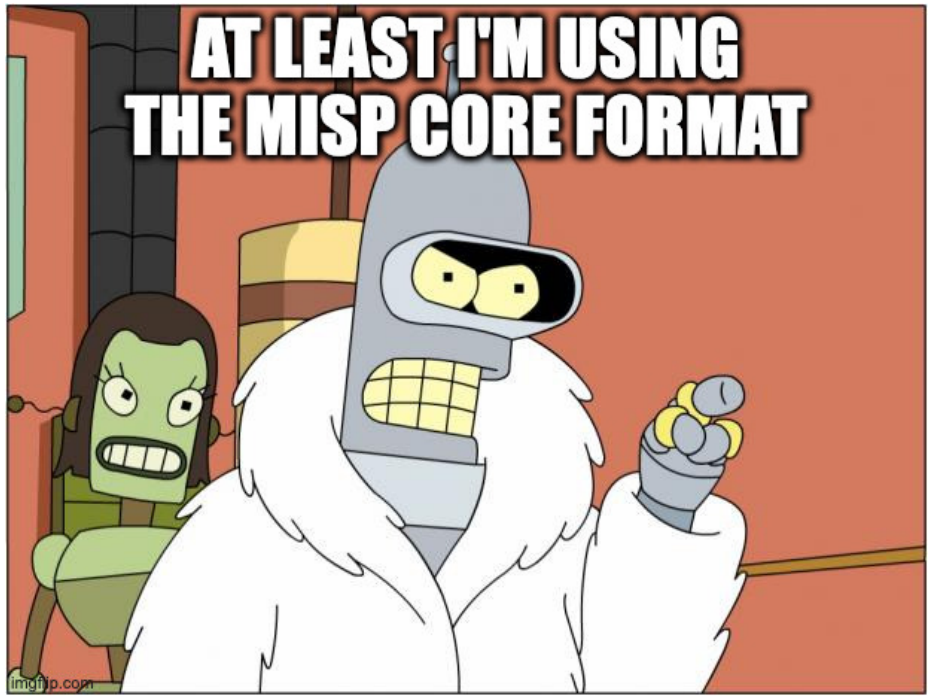
- Development
- Education
- Electric
- Gas
- Health
- Industrial
- Insurance
- Intelligence
- Legal
- Manufacturing
- Oil
- Technology
- Water

Room for improvement (4)

- Scraping websites to gather threat intel reports is far from ideal.

Room for improvement (4)

- Scraping websites to gather threat intel reports is far from ideal.
  - MISP web scraper is no longer needed and everyone uses the MISP core format.



Resources

Get it via GitHub

- https://github.com/cudeso/misp-scraper

Documentation on GitHub and on the MISP project website

- https://www.misp-project.org/2022/08/08/MISP-scraper.html/

Questions?