OSINT Project

Open Source INTelligence

Information surrounds us. Whether at home, in social settings, or even in the workplace. Likewise, to understand, track, and evaluate the vast amounts of information produced in business, many companies utilize open-source intelligence.
The main idea behind open-source intelligence (OSINT) involves collecting, analyzing, and generating actionable insights from publicly available sources. Although OSINT has been around for as long as records have been collected and analyzed by government-based transactions, companies have only gradually unlocked its true potential in recent decades.

But what is OSINT?

Open Source Intelligence (OSINT) refers to all information that can be found publicly, mostly via the internet, without breaching any copyright or privacy laws. Under this definition, a wide array of sources can be considered a part of OSINT. For instance, information posted publicly on social media websites, posts on discussion forums and group chats, unprotected websites directories, and any piece of information that can be found by searching online.

OSINT for Ethical Hacking

IT security professionals utilize OSINT search techniques and tools to discover weaknesses in friendly IT systems, so such vulnerabilities can be closed before threat actors discover them. Commonly found vulnerabilities include:

- Accidental leaking of sensitive information on social media sites. For example, an unaware employee may post a personal photo in the

server room showing the type of security devices used to secure a corporate network.

- Open ports and insecure services running can be discovered when scanning the subject network for vulnerabilities using specialized tools.
- Outdated operating system versions, software, and any content management systems already in use.
- Leaked information found on data leak repositories or across the darknet.

Proposed targets

Two targets have been chosen for the OSINT project and they are both websites related to Italian government institutions.
The first target is the website of the “Latina” Province, the Italian province in which I live, which has approximately 120k inhabitants.
The second one is the website of the province of “Ragusa”.

OSINT Search Pipeline

Let's consider a scenario in which we need to find information related to some topics on the web. For this purpose, we first need to search and perform analysis until we get the exact results. We expect these operations to be very time-consuming, especially searching for information about the date of creation of the IT infrastructure, which company worked on it, and what technologies have been used. We expect, hopefully, all these pieces of information to be present in institutional government files, deeply stored in some web archive.
Since the process mentioned above can be very time-consuming, we need intelligence tools that allow us to perform all the previous operations within seconds. Finally, by running multiple tools, we can collect, and later use, all the information, even the correlated ones, related to the target.

Step 1: Find pieces of information about the web infrastructure in institutional papers

The Italian government has a transparency policy called “transparent administration”, in which all the institutional entities must publish, online, all the contracts and notices about every single cent spent by them. This could be useful to find out on which date the web infrastructure has been created and by whom (the cost is not relevant for our purpose); on the other hand, the “transparency” policy does not oblige the winning company of the public announcement to publish in which way they will deliver the service, using what technology. To discover this, and so, to find out which technologies have been used and what are their current vulnerabilities of them, the OSINT tools come in handy.

Step 2: Find useful pieces of information with Shodan

Shodan is a database of billions of publicly available IP addresses, and it’s used by security experts to analyze network security. It is a search engine for hackers to see exposed assets considering that, when compared to other search engines, Shodan provides you the results that make more sense related to security professionals. It mainly includes information related to assets that are being connected to the network. The devices may vary from laptops, traffic signals, computers, and various other IoT devices. This open-source tool mainly helps the security analyst in identifying the target and testing it for different vulnerabilities, passwords, services, ports, and so on.

Step 3: Exploit Google capabilities via Google Dorks

Google Dorks came into existence in 2002, and it gives effective results with excellent performance. This query-based open-source intelligence tool is mainly developed and created to help users in targeting the index or search results appropriately and effectively. Google Dorks provides a flexible way of searching for information by using some operators, and perhaps it is also

called Google Hacking. These operators make the search easier to extract information.

Step 4: Collect data of the target via SpiderFoot

SpiderFoot is an open-source reconnaissance tool available for Linux and Windows. It has been developed using Python language with high configuration and runs virtually on any platform. It integrates with an easy and interactive GUI with a powerful command-line interface.
It has automatically enabled us to use queries over 100+ OSINT sources to grab the intelligence on emails, names, IP addresses, domain names, etc. It collects an extensive range of information about a target, such as netblocks, e-mails, web servers, and many more. The data collected from SpiderFoot will provide a wide range of information about our specific target. It provides clear insights about possible hacking threats which lead to vulnerabilities, data leaks, and other vital information.

Step 5: Draw Conclusions about what has been found

At this point, all the informations that have been found can be merged to find all the vulnerabilities of that specific target and possible mitigations.

OSINT Search First Target comune.latina.it

Step 1:

On this webpage, it is possible to check every public announcement done by the city of Latina from 2014 onwards. After several attempts in query form, searching for the keywords (in Italian) such as: “Information technologies”, “website”, “IT Services”, etc. I have found that the following documents are the ones relevant to the knowledge of the web infrastructure:

- The first one states that the current web infrastructure has been created and set up 4 years ago, in the year 2018

Data inizio pubblicazione	13/03/2018
Data fine pubblicazione	31/12/2023

  and it holds until the 31/12/2023, with the following front-end components (the pdf documents have been translated from Italian to English with google translate):

> - that at present the following systems are active on the front-ends of the institutional site:
> a) word-press environment of the institutional site
> b) common agenda; c) wordpress environment; d)
> zimbra e-mail service; e) EFA Antivirus Antispam
> f) Pegaso Web; g) pfSense; h) virtual systems
> rescue environments; i) e-mail systems rescue
> environment; **EXPECTED** that by virtue of the
> provisions contained in the rules on transparency
> and anti-corruption in public administrations
> (Legislative Decree 31 August 2013, no.101 "Urgent provisions

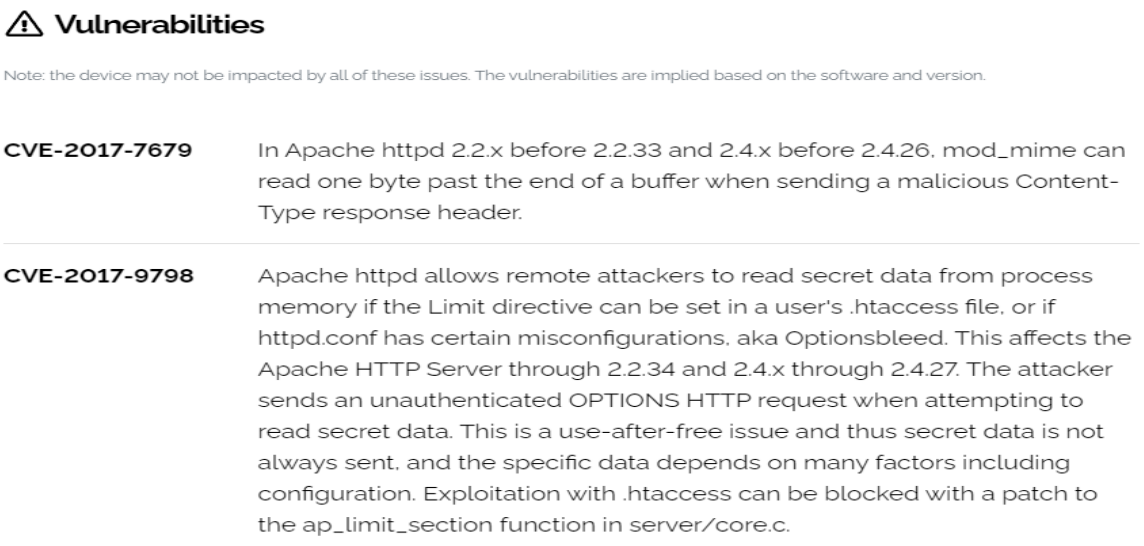
- The second one is a more recent document that states that the SSL certificates and web hosting have been updated and paid on the date 30/12/2021 for the current year:

> **Given,** again, that with Resolution no. 2194/2021 the hosting services of the website of the Organization, the e-mail server and other various servers, as well as numerous ancillary services, better specified in this deed, were renewed to the company Panservice sas;

Step 2:

About the control using the Shodan web service, first of all, it is required to find the public IP address of *https://www.comune.latina.it/*, which is found via, and is “212.66.100.22”. After inserting it on Shodan, many pieces of information have been displayed, and the most interesting ones are the following:

- The open ports on the webserver are the following ones, i.e., 21, 80, and 443. In which the 80 and 443 are correct but 21 could be avoided.
- The SSL certificate is correctly validated, as we have seen before, and therefore not expired. The validity is up until 20/5/2022.
- FTP, linked to the port 21 open, is often thought of as a “not secure” file transfer protocol. This is mainly due to FTP sending data in clear text and offering an anonymous option with no password required.
- The most interesting fact that Shodan displays is the current version of the Apache Server, which is 2.4.18.
  Shodan has integrated the possibility to show to the user the current vulnerabilities of the Apache server, via the CVE, i.e., the Common Vulnerability and Exposure which is a system that provides a reference method for publicly known information-security vulnerabilities and exposures. An example of them is the following:



Showed just like this, it is not immediate to understand the level of danger of each vulnerability. To overcome this, it is possible to get, for each vulnerability, an index that explains better how dangerous they are, on a scale from 0 to 10, that is the CVSS, Common Vulnerability Scoring System.

The problem is that Shodan provides this information just to the paid users, but, knowing the version of the apache server, it is possible to get the CVSSs of the vulnerabilities regarding that server on the CVEDetails website. Among all the vulnerabilities, there are two, the CVE-2017-7668 and CVE-2017-3169, that, of course, were also shown by Shodan, that are really dangerous considering the CVSS of 7.5.

4	CVE-2017-7668	20	2017-06-20	2021-06-06	7.5	None	Remote	Low	Not required	Partial	Partial	Partial

The HTTP strict parsing changes added in Apache httpd 2.2.32 and 2.4.24 introduced a bug in token list parsing, which allows ap_find_token() to search past the end of its input string. By maliciously crafting a sequence of request headers, an attacker may be able to cause a segmentation fault, or to force ap_find_token() to return an incorrect value.

5	CVE-2017-3169	476	2017-06-20	2021-06-06	7.5	None	Remote	Low	Not required	Partial	Partial	Partial

In Apache httpd 2.2.x before 2.2.33 and 2.4.x before 2.4.26, mod_ssl may dereference a NULL pointer when third-party modules call ap_hook_process_connection() during an HTTP request to an HTTPS port.

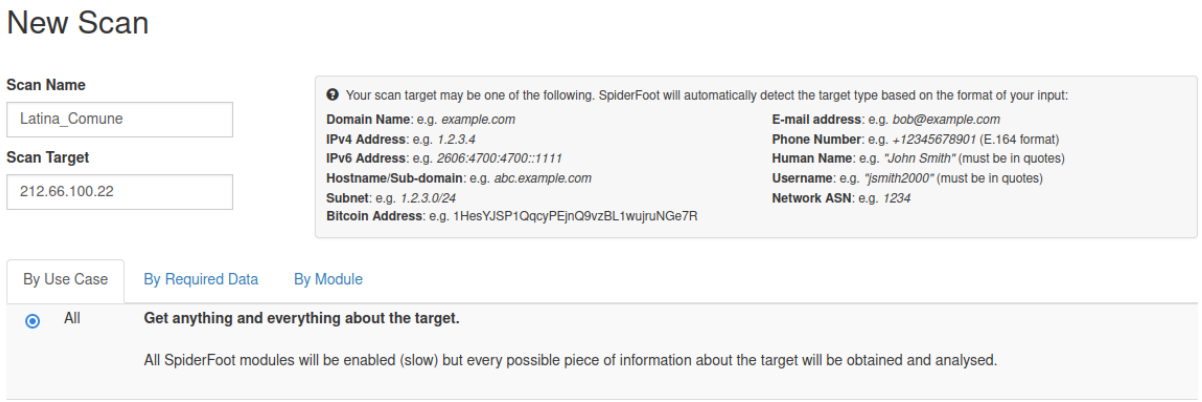
Step 3:

About the Google Dork procedure, several queries have been performed to Google with the condition “site:.*comune.latina.it” which states that every query must be performed just in the subdomains of *comune.latina.it*. Some examples of queries performed are the following:

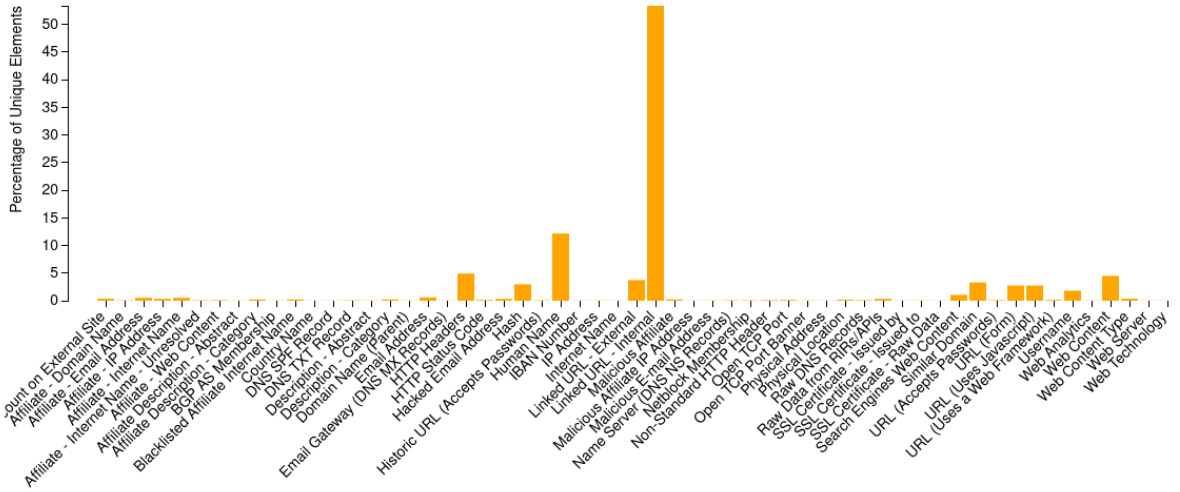
- site:.*comune.latina.it filetype:sql
  To find backup copies stored on the server, changing the filetype also with bak, tar.gz, sql. No results.
- site:.*comune.latina.it *“warning” “error”*
  The text of the error may contain various data about the app’s system components. No results.
- site:.*comune.latina.it filetype:txt
  This query allows us to find files with txt extensions that are commonly used to store usernames and passwords, and the result is on this webpage. The fact that on the webserver is stored a “robot.txt” file, that prevents the search engines to reach and index the folders “/storico-sito/” and “/latina-80/” means that the developers were aware of the danger of the Google Dorking procedure and, for this reason, they have tried to mitigate it.

Step 4:

The use of SpiderFoot is pretty straightforward; after starting the service spawned on a local server, with the command “ spiderfoot -l 127.0.0.1:5001 “, on localhost, a useful WUI has been spawned with the service running.
Now it is time to perform a scan inserting the IP address of the target and selecting “all” in the modalities to get all the relevant pieces of information that spiderfoot gets



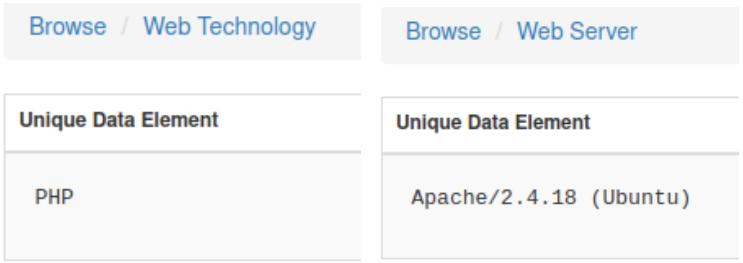
After approximately 9 minutes, Spiderfoot terminated the search on the target, and, in this particular case, it got 5320 results of which 2925 were unique.



At this point, the hard part is to extrapolate all the useful and relevant pieces of information for our needs, and this is what I got:

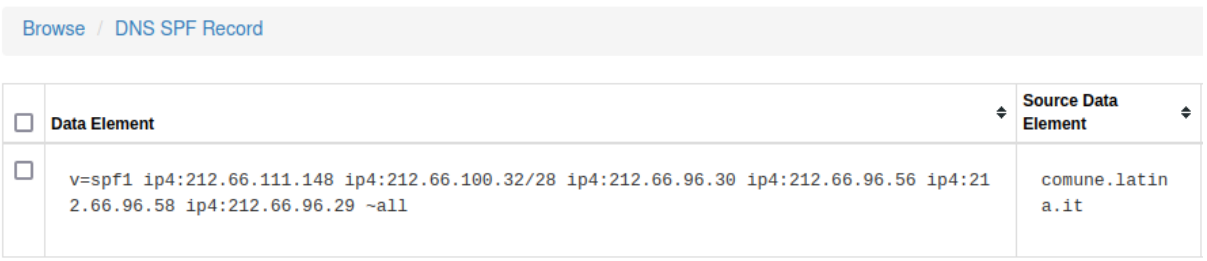
- Thanks to the virusTotal extension, we got that the IP address is marked as secure

- The technologies used, found by spiderfoot, confirms the ones found before, i.e., Apache server 2.4.18 and the web scripting technology used is PHP



- A very interesting discovery is that Spiderfoot was able to find also mail addresses that have been hacked in previous data breaches, such as the Cit0Day and the OnlineSpamBot (of course the mails have been obscured for privacy reasons):

- A DNS SPF Record has been identified that states how just the IP addresses in the subdomain of the main server are permitted to send an email on behalf of the main domain.



Step 5:

From the analysis just performed, it is possible to get some important aspects about the web infrastructure of the target:
First of all, considering the period of when the infrastructure has been created, traceable from the institutional papers found, that is coherent with the Apache version found (the Apache version 2.4.18 has been released at the beginning of 2016), we can definitely conclude that the webserver where the service is hosted is not secure and absolutely needs a refresh and an update of it. This certainty is due to all the vulnerabilities found before, especially with the two vulnerabilities with a really high CVSS that leads to an infrastructure prone to really dangerous attacks, potentially destructive.
Thanks to the documents published and then confirmed by SpiderFoot which showed the presence of the folder “wp-content” on the webserver, I have found that the CMS, or content management system, used is WordPress, thanks to this information, a list of potential vulnerabilities related to it can be discovered here.
A good job has been found related to the defense against the Google Dorking attacks that makes it really hard to locate valuable data or hard-to-find content.
The SSL certificate will expire in a month, but a contract to extend them has already been approved as seen before.
One major problem that SpiderFoot found is the possibility to get, via the webserver, some email addresses that have been pawned in previous data breaches. A remotion of those mail addresses from the website can be suggested.
Overall the website has been proven fairly secure, without any major issues, with some minor ones that can be overcome by simply updating them, with the only main issue is that the contract with the IT company expires at the end of 2023 so it is difficult to imagine an update in the short-term.

OSINT Search Second Target provincia.ragusa.it

Step 1:

On this webpage, it is possible to check every public announcement done by the Province of Ragusa from 2017 onwards. As before, I have searched for many keywords that could lead to relevant values, and the only document that I have found, referring to the website and its infrastructure is FNN20170033079.PDF, a pdf document that states that the most recent change in the website goes back to 2017 when the website has been refreshed with a more modern style and also the backbone of the infrastructure, has been changed.

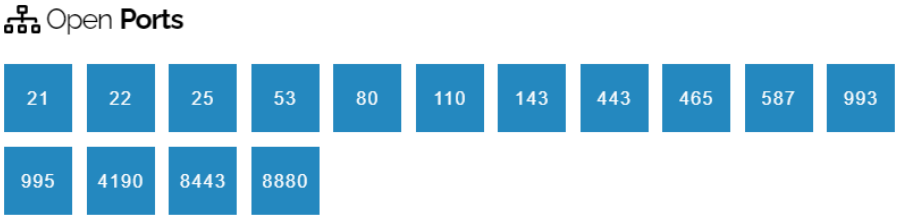
> allo stato attuale il sito web dell'Ente necessita di una profonda operazione di restyling non solo grafico, ma anche ai fini dell'adeguamento alle ultime norme di legge in materia di accessibilità, usabilità, trasparenza e diffusione di informazioni;

Translation: "At the actual state, the website of the institution needs a deep restyling operation, not just on the graphic side, but also for the purposes of adaptation to the last laws about accessibility, usability, transparency, and spreading of information."
Another interesting document that can be found on the platform is the following:
"Adoption of the document defining the "Procedure for the management of IT incidents and violations of personal data (data breach)""
which, as the name suggests, is a document that outlines some procedures in case of a data breach, that varies from risk management to the steps that must be followed in case of a data breach, all this to adequate to the European GDPR.

Step 2:

About the control using the Shodan web service, first of all, it is required to find the public IP address of *https://www.provincia.ragusa.it/*, which is found via, and it is “51.210.10.162 ”. After inserting it on Shodan, many pieces of information have been displayed, and the most interesting ones are the following:

- There are many open ports on the webserver:



  It is not a common behavior to let all those ports open on a web server, but before they are marked as dangerous, it should be found what is done with those open ports at the system level, and what are the services exposed and apps running on those ports. So, if a port has been found open, it can become a real threat if the services that are running on them aren’t properly hardened from a network, operating system, and app point of view.
- The SSL certificate is up to date and valid until 28/06/2022
- Considering that Shodan does not provide information about the version of the Apache server, the focus is on the SSH port open (the 22), in which Shodan identifies the server running in OpenSSH v. 7.6p1. After a search on the web, the most important and relevant vulnerability is the CVE-2017-15906 with a CVSS of 5.0.

Step 3:

As the Use-case performed before, also in this one there is mitigation against Google Dorking, because, if the following query is performed on Google Search “site:.*provincia.ragusa.it filetype:txt”, the following results appear:
https://www.provincia.ragusa.it/robots.txt
The robots.txt is basically a text file which is the first thing that is indexed by crawlers when crawling through a website. It is like setting permissions to a crawler, like restricting it from not indexing any section of a website that may contain sensitive information (login pages sensitive directories, and files), making useless the Google Dorking procedure.

Step 4:

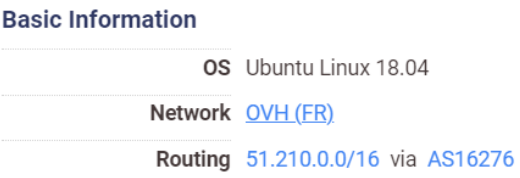
The search with SpiderFoot, at this time, does not provide many useful informations or add other useful insight. The only relevant addition is the following:

- It has been found a malicious affiliate:

Step 4 Extra:

This step has been performed with Censys. Unlike Shodan which captures the data in banners, Censys is built upon the Zmap. Zmap is a faster alternative to nmap that can scan the entire 4 billion IPv4. This enables it to have an almost real-time update on every IP address.
After inserting the IP address of the website on the search engine, those are the most interesting details:

- The website is hosted via OVH, which is the biggest web hosting company in Europe



- The platform provided also the information that on this particular webpage, that is a form with username and password, a wrong certificate was issued, as matter of fact, also Chrome identifies the webpage as not secure and suggests not inserting sensible pieces of information:

Step 5:

As we have seen in the previous step, the presence of potentially dangerous behavior, such as the presence of many open ports on the webserver, does not lead for sure to a detection of a potential flaw in the infrastructure using OSINT techniques, for sure this does not assure that they can not be found using more dangerous tools.
This website has been chosen to prove precisely this point, that is, that the OSINT procedures do not always produce a relevant result, but, in many modern scenarios, the developer or the security officers, knowing the danger of it, try in every possible way to mitigate the possibility to the attackers to get relevant pieces of information about the web infrastructure.

Conclusions

As technology and the amount of data increases day by day, the need for fast and specific information gathering arises and so it increases the need for OSINT procedures or techniques. In the upcoming years, OSINT will become the basic need, for the first approach, of the organization whether it's private or government. By using OSINT we can get important information in just a couple of minutes which, otherwise, would be only possible by deep analysis in newspapers, magazines, published institutional documents, etc…
The OSINT procedures have proved to be very effective when applied to those contexts where there is the public availability of documents, such as in the case of institutional systems, but, as we have seen before in the second target, those pieces of information could lead to potentially no conclusions if the web infrastructure has been built with the knowledge of such danger, considering an attacker side, which OSINT represents. Different scenario, represented by the first target, when just a small portion of the infrastructure has been protected against OSINT techniques, where many vulnerabilities, and therefore many dangers, have been found by just searching on the superficial part of the infrastructure.
The most useful and common tool to perform OSINT is for sure SpiderFoot, but the main problem of the program is exactly the huge amount of data that it produces, which is so much, that a not-so-skilled user can feel really lost.
Another useful tool is Shodan, which in many cases can produce relevant results immediately, but in others must be integrated with other programs. The Google Dorking procedure is really a fun procedure, but, nowadays, the web infrastructures are aware of this methodology and try to mitigate it, which is possible in very simple ways.
In the end, it can be said that the OSINT techniques are a good first approach in the vulnerability detection of the web infrastructure, but must be performed alongside other techniques to get a full big picture of the dangers and threats that an attacker can exploit.