**Information gathering**

[Introduction 3](#_Toc32603025)

[Business side 3](#_Toc32603026)

[Infrastructure side 3](#_Toc32603027)

[Output of information gathering 3](#_Toc32603028)

[Information gathering technique 3](#_Toc32603029)

[Passive 3](#_Toc32603030)

[Active 3](#_Toc32603031)

[Business side 4](#_Toc32603032)

[Search 4](#_Toc32603033)

[Web presence 5](#_Toc32603034)

[Search engine 6](#_Toc32603035)

[Google dork 6](#_Toc32603036)

[LinkedIn 6](#_Toc32603037)

[Inspy tool 6](#_Toc32603038)

[U.S Government 7](#_Toc32603039)

[Partners and third parties 8](#_Toc32603040)

[Job posting 8](#_Toc32603041)

[Financial information 9](#_Toc32603042)

[Harvesting 10](#_Toc32603043)

[Cached and archival sites 10](#_Toc32603044)

[Social media 11](#_Toc32603045)

[Infrastructure 13](#_Toc32603046)

[IPs, domains and subdomains enumeration 14](#_Toc32603047)

[WHOIS 14](#_Toc32603048)

[NSLOOKUP 15](#_Toc32603049)

[dig 15](#_Toc32603050)

[Netcraft 15](#_Toc32603051)

[MSN ping 15](#_Toc32603052)

[Netblock enumeration 15](#_Toc32603053)

[Live host enumeration 16](#_Toc32603054)

[virtual host enumeration 16](#_Toc32603055)

[Domains and Subdomains enumeration 16](#_Toc32603056)

[Fingerprint the webserver 17](#_Toc32603057)

[Mitigation fingerprint the webserver webserver 17](#_Toc32603058)

[Fingerprinting webserver modules 17](#_Toc32603059)

[Mitigation fingerprint webserver modules 17](#_Toc32603060)

[Fingerprinting frameworks and applications within webserver 17](#_Toc32603061)

[Fingerprinting custom applications 18](#_Toc32603062)

[Finding Hidden Paths and files within webserver 18](#_Toc32603063)

[Information through Misconfigurations within webserver 18](#_Toc32603064)

[Enumerating user accounts from webserver 18](#_Toc32603065)

[Automate information gathering 19](#_Toc32603066)

[DNSEnum tool 19](#_Toc32603067)

[dnsmap 19](#_Toc32603068)

[Shodan HQ 19](#_Toc32603069)

[Tips and tricks 20](#_Toc32603070)

[Resources 20](#_Toc32603071)

[Google dork 20](#_Toc32603072)

# Introduction

Must follow a methodical, organized and controlled process in order to effectively review targets and keep the penetration tester safe from consequences if issues arise.

Must use different techniques in order to obtain information. Being this detail driven will allow the tester to record only the needed data on the intended target.

The information gathering phase is focused on two essential aspects of all targets: business and infrastructure.

## Business side

Information gathering deals with collecting information regarding the type of business, its shareholders, assets, products, services, employees and generally non-technical information.

## Infrastructure side

The organization will probably operate its business purpose through infrastructures such as networks, systems, domains, IP addresses and so on.

## Output of information gathering

At the end of the information gathering process you should at least have the following important information about the target:

|  |  |
| --- | --- |
| Infrastructure | Business |
| Network maps | Documentation |
| Network blocks | Web presence (Domains) |
| IP addresses | Physical locations |
| Ports | Employees / Departments |
| Services | Emails |
| DNS | Partners and third parties |
| OS | Press/ new releases |
| Alive machines | Financial information |
| System | Job posting |

## Information gathering technique

### Passive

Passive or OSINT (Open source intelligence) is gathering as much information about your target (network, system…) without exposing our presence. Gather information such as web presence, partners, financial info., physical plants.

### Active

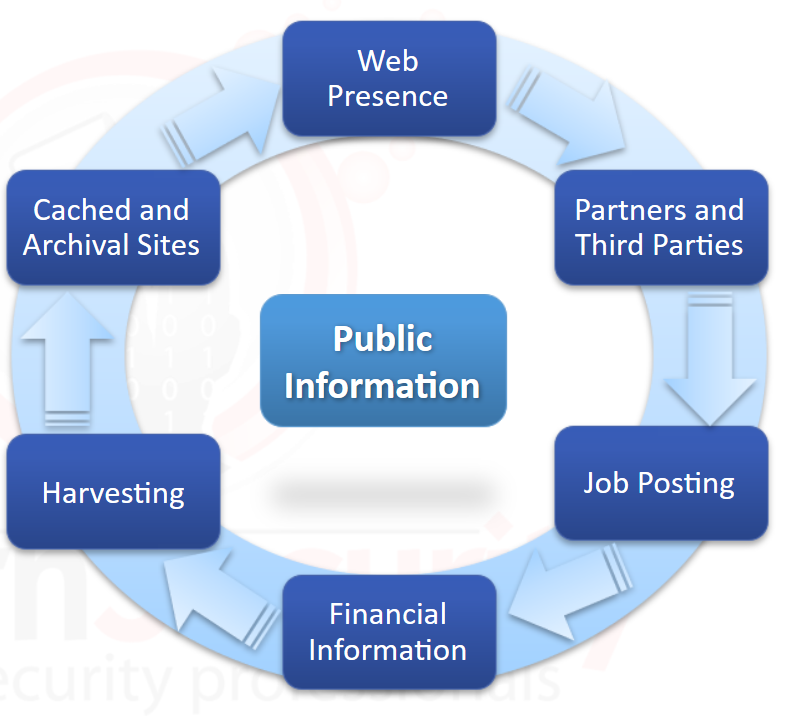
Interact directly with the target system, in this phase, we will gather information about ports, services, running system, net blocks and so on.

**Note:** IDS or servers logs could be existed so caution should be taken to prevent this.

# Business side

## Search

During the business-related information gathering phase, there is a great deal of diverse research conducted are as the following:



### Web presence

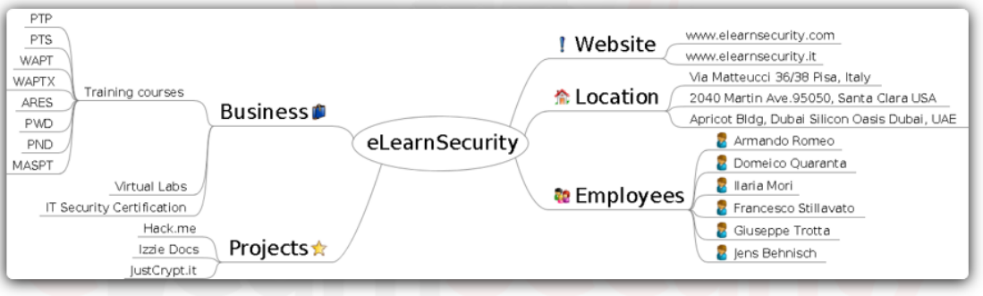
In this phase, you will learn more about your target including:

* What they do?
* What is their business purpose?
* Physical and logical locations.
* Employees and departments.
* Email and contact information.
* Alternative web sites and sub-domains.
* Press release, news, comments, opinions.

eLearn security as example:



Each time you find something new on the target company, jot it down in your mind mapping tool



### Search engine

#### Google dork

Using Google’s sophisticated search operators for our information gathering purposes, to uncover using Google to find misconfigured web servers, sensitive information left on a server (that was crawled by google bots), password files, log files, directory listings and many others.

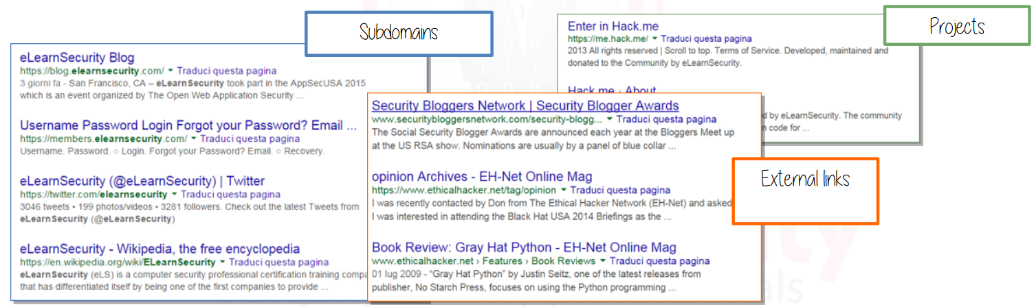
Through Google Hacking, we may able to detect:

* Error Messages that may contain valuable information.
* Sensitive Files and directories.
* Server or application vulnerabilities.
* Pages that contain login portals.
* Much more.

For example:

|  |  |
| --- | --- |
| cache | cache: www.website.com, will show the cached content of website.com |
| Link | link: www.website.com, will display websites that have links to the specific website. |
| site | Google dork site:www.website.com |
| filetype | Google dork filetype:pdf |

Kind of information that can retrieve using search engine



#### LinkedIn

extract LinkedIn users based on the organization, job description or email address.

Tools for crawling:

* Inspy tool.

##### Inspy tool

The objective of this Information Gathering tool is to extract LinkedIn users based on the organization, job description or email address. InSpy was written in python by gojhonny.

Multiple Functionalities:

* TechSpy.
* EmpSpy.

TechSpy: Crawls LinkedIn job listings for technologies used by the provider company. It attempts to identify technologies by matching job descriptions to keywords from a new line delimited file.

EmpSpy: Crawls LinkedIn for employees working at the provided company. It searches for employees by title and/or departments from a new line delimited file. It may also create emails for the identified employees if the user specifies an email format.

Installation link: git clone <https://github.com/gojhonny/InSpy.git>

## U.S Government

Organizations that operate globally and have a desire to sell to the U.S government or government agencies, are required to possess two codes useful to us:

1. DUNS number (DUNS and Bradstreet).
2. CAGE code (or NCAGE for a non-U.S. business).

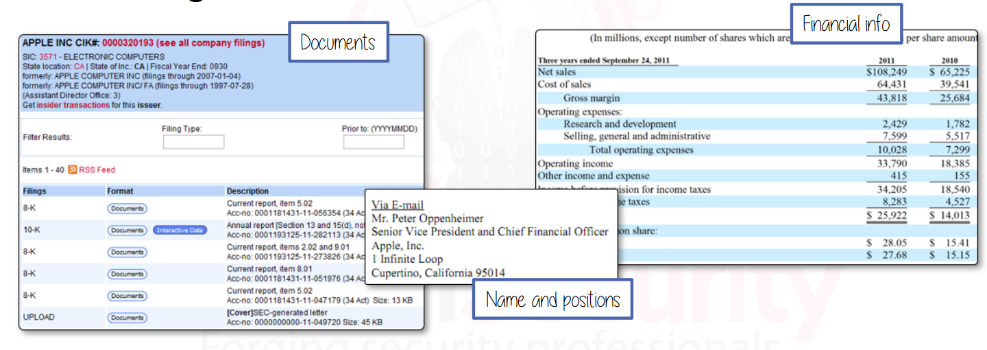
The two codes allow us to retrieve even more information such as contacts, products, lists, active/inactive contracts with the government and much more.

SAM link: <https://www.sam.gov/SAM/pages/public/searchRecords/search.jsf>

**Note:** organizations belonging to different industries can be investigated through search in different publicly available databases. Compliance and regulations might force companies to publish different kind of the information publicly.

For example: traded companies that have to file their financial document to sec database.

* To find financial document database, can use the EDGAR (Electronic Data Gathering, Analysis, and Retrieval system).
  + Link: <http://www.sec.gov/edgar.shtml>



## Partners and third parties

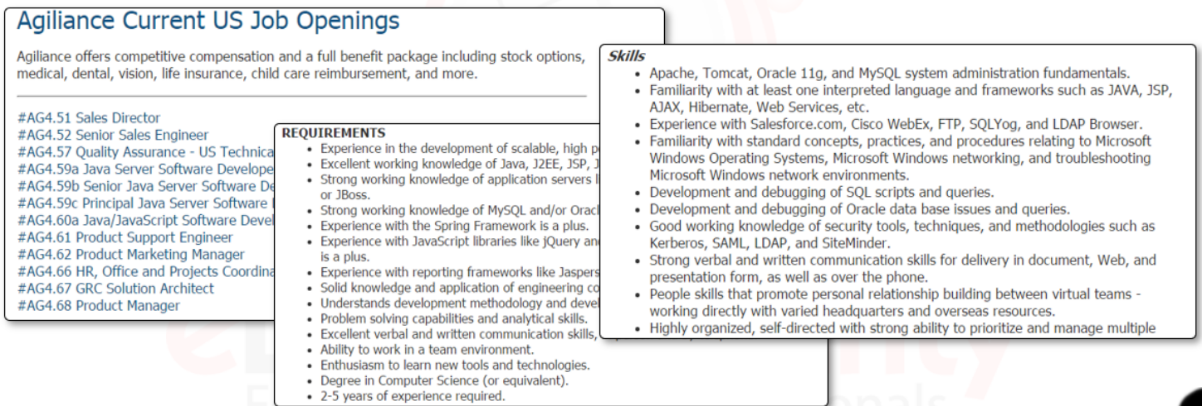
Other useful information that can gather about company are mergers and acquisitions, partnerships, third parties…

We can deduce what type of technologies and systems they use internally. You can take advantage of this information in later phases.

## Job posting

Many organizations have a web site section including open positions and career opportunities.

Investigator can deduce internal hierarchies, vacancies, projects, responsibilities, weak departments, financed project, technology implementation and more.



Search engine for job vacancies:

* Indeed: <https://eg.indeed.com/?r=us>
* LinkedIn: <https://www.linkedin.com/feed/>
* Monster: <https://www.monster.com/>
* Careerbuild: <https://www.careerbuilder.com/>
* Glassdoor: <http://www.glassdoor.com/>
* Simplyhired: <http://www.simplyhired.com/>
* Dice: <http://www.dice.com/>

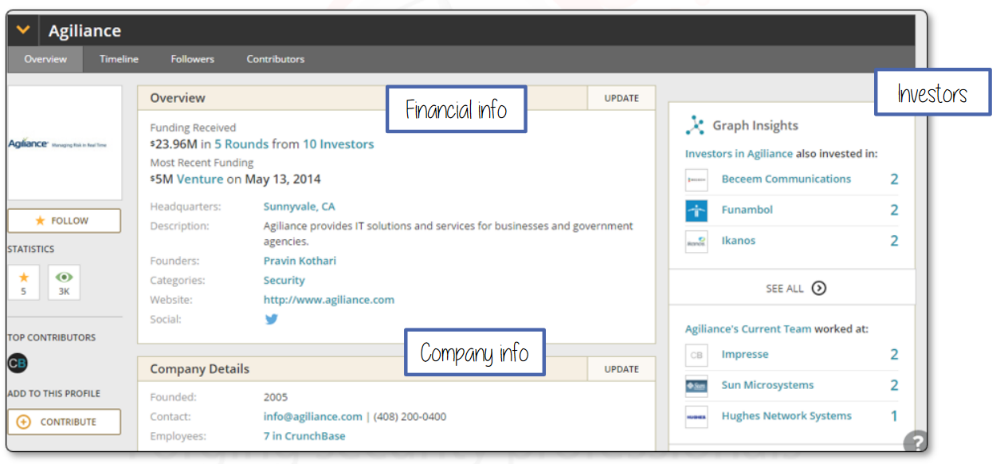
## Financial information

More useful information can be acquired from financial details about the organization.

* Is going to invest in a specific technology.
* Might be subject to a possible merge with another organization.
* Has critical assets and business services.

How we can collect financial information?

* CrunchBase: <http://www.crunchbase.com/>
  + Is a database where you can find information about:
    - Companies.
    - People
    - Investors and financial information.
  + The power of CrunchBase is grounded on the concept of anyone able to edit information in it.



* INC: <http://www.inc.com/>
  + Focuses its attention on growing companies and provides advice, resources and information to companies.



* Google finance
* EDGAR
* Yahoo financial

## Harvesting

We unpack methods for gathering company documents such as charts (detailing the corporate structure), database files, diagrams, papers, documentations, spreadsheets and so on.

Kindly be noted when document is created, it automatically stores information (metadata) like who created it, date and time of creation, software used, computer name and so on.

Tools could be useful:

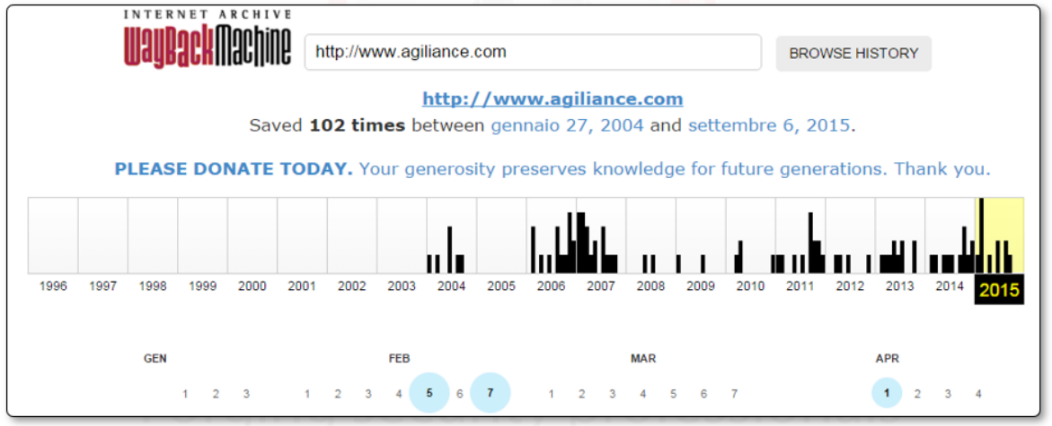
* FOCA (Fingerprinting Organizations with Collected Archives).
  + is a tool used mainly to find metadata and hidden information in the documents its scans. These documents may be on web pages and can be downloaded and analyzed with FOCA.
* TheHarvester tool:
  + A tool for gathering e-mail accounts and subdomain names from public sources.

## Cached and archival sites

Since the information on the web changes so quickly, sometimes seeking an older version of a site could prove useful to our cause. For example, pervious job post.

Useful tool:

* Archive: <http://www.archive.org/index.php>



* Google dork
  + cache:URL



## Social media

The spread of social networks has made information gathering extremely important and effective.

With social media we can gather employee’s personal information such as: phone numbers, addresses, history, CV, opinions, responsibilities, projects and so on.

Social media useful in the following ways:

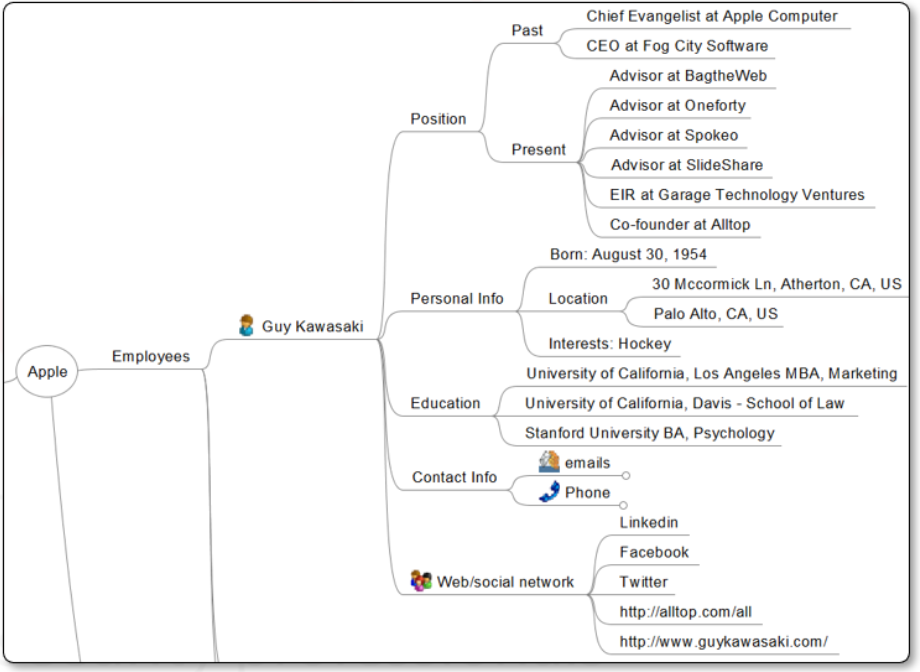
* learn about corporate culture, hierarchies, business processes, technologies, applications.
* Build a network map of people (relationships).
* Select the most appropriate target for a social engineering attack.

Once you get list of people, you can start collecting personal information on them. You can use the following:

* pipl: <http://www.pipl.com/>
* spokeo: <http://www.spokeo.com/>
* peoplefinders: <http://www.peoplefinders.com/>
* CrunchBase: <http://www.crunchbase.com/>

At this point of our information gathering phase we already know some information seen below:

* Age
* Phone number
* Business
* Addresses
* Occupation
* Interests
* Email addresses
* Website owned
* Related documents
* Financial information



**Note:** Usenet is a world-wide distributed discussion system. It consists of a set of newsgroups with names that are classified hierarchically by subject. We can also find additional information by searching for individuals’ name or email in google group. This may lead us to further sensitive data shared by the target company and its employees.

# Infrastructure

The main goal here is to retrieve data such as:

* Domains.
* Netblocks or IP addresses.
* Mail servers
* ISP’s used
* Any other technical information.

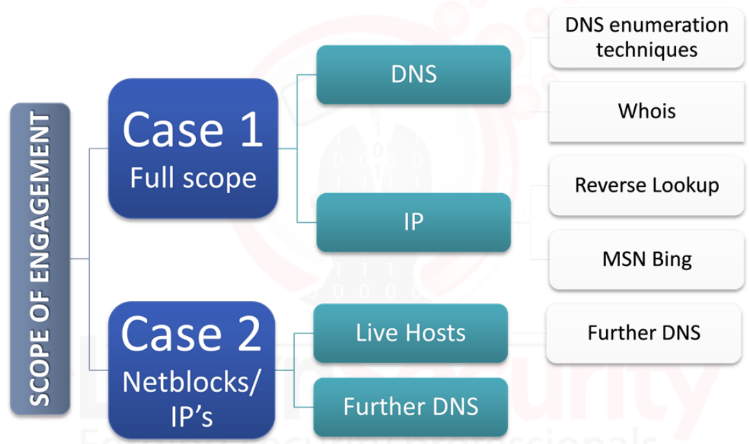
The infrastructure behind a web application is what supports it and allows it to function. This includes the web server that is directly involved in the execution of any web application. The two common web servers used on the internet today are Apache and Microsoft IIS.

Uncovering both the web server type and version will give us enough information to mount many different attacks against its components (during stages of the test).

IIS components, usually called ISAPI extensions, work as dynamic libraries, extending the functionalities of the web server and performing different tasks for the web server.

These include: URL rewriting, load balancing, script engines (like: PHP, Python or Perl) and many others.

URL Rewriting: changes ugly web application URLs such as “new.php?id=12” to more search-engine-friendly URL like “news/12.html” or route like “news/12”.



Let’s first consider a full scope engagement.

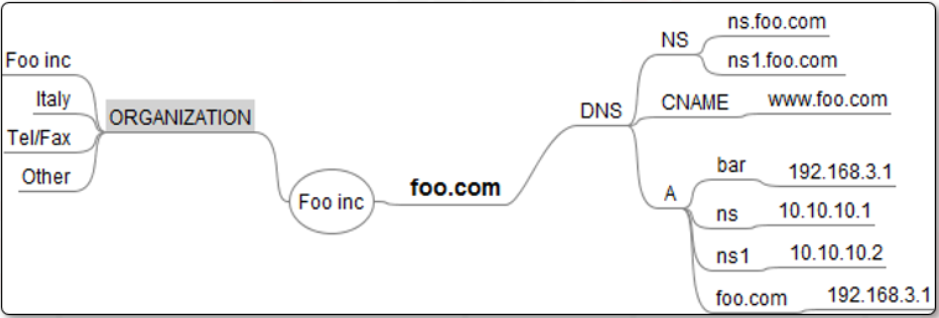
In this case your engagement is similar to how a malicious hacker would attack. Indeed, the hacker only knows the target organization name at the beginning and then, he tries to derive as much information from that.

## IPs, domains and subdomains enumeration

This process aims to collect all the hostnames related to the organization and the relative IP addresses.

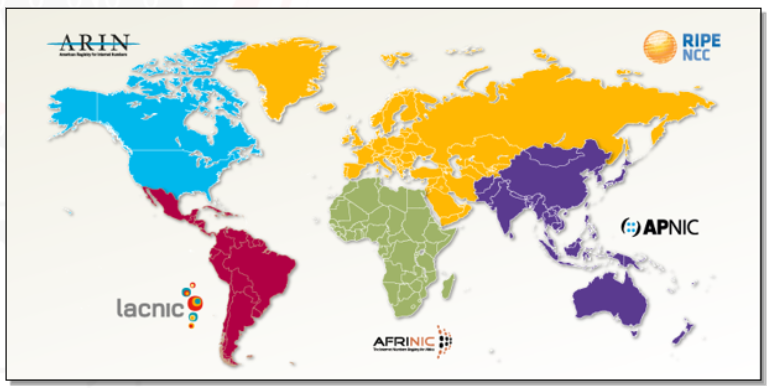
This process ends when we obtain the following information:

* Domains
* DNS servers in use
* Mail servers
* IP addresses

[output simple – target mapping]

### WHOIS

This is public database and should be the first step in any investigation on infrastructure related information.

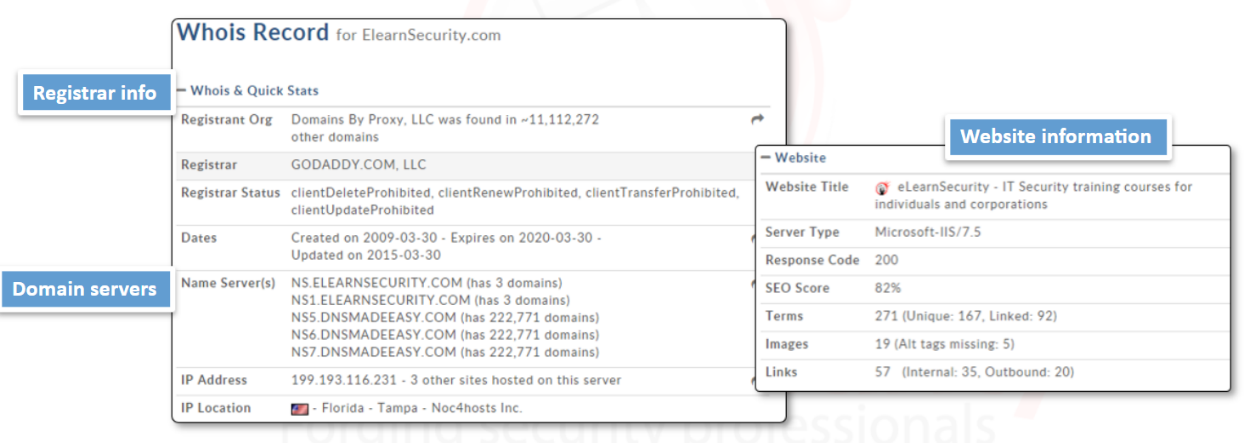


Is a query/response protocol runs TCP port 43, widely used for querying an official domain registrar’s database, in order to determine:

* The owner of a domain name.
* IP address or range.
* Autonomous system
* Technical contacts
* Expiration date of the domain

For perform WHOIS lookup:

* <http://who.is>
* <http://whois.domaintools.com>
* <http://bgb.he.net/>
* <http://networking.ringofsturn.com/Tools/whois.php>
* <http://www.networksolutions.com/whois/index.jsp>
* <http://www.betterwhois.com/>



**Note:** name servers are the most valuable information due to the NS store all the DNS related information (records) about the domain.

### NSLOOKUP

is another very handy tool that lets you translate hostnames to IP addresses and vice versa.

What we get?

* NS record
* A record
* AAA record
* MX record
* CNAME’s record
* Etc.

### dig

more powerful tool and alternative for nslookup.

### Netcraft

has a wealth of information and it performs the same functions of NSlookup and WHOIS.

### MSN ping

Bing search engine offer a query filter that return all the websites hosted on a given IP address.

You can also use the following tools:

* <http://www.tcpiputils.com/domain-neighbors>
* <http://reverseip.domaintools.com/>
* <https://www.robtex.com/>

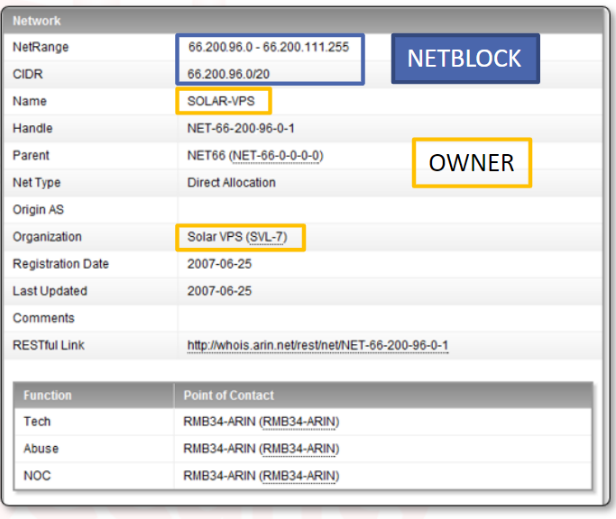
### Netblock enumeration

A netblock is a range or set of IP addresses, usually assigned to someone and has both a starting and an ending IP address, for example 192.168.0.0-192.168.255.255

What is autonomous system?

Is a made of one or more net blocks under the same administrative control. Big corporations and ISP’s have an autonomous system, while smaller companies will barely have a netblock.

We can use WHOIS to discover netblock as the following:



### Live host enumeration

There is different method that one can use to identify live hosts. The common is the ICMP ping sweep.

It consists ICMP ECHO requests sent to multiple hosts. If a given host is alive, it will return ICMP ECHO reply.

You can perform it through the following tools:

* fping
* nmap
* hping

### virtual host enumeration

A virtual host is simply a website that shares an IP address with one or more other virtual hosts. These hosts are domains and subdomains, this is very common in a shared hosting environment where a multitude of websites share the same server/IP address.

### Domains and Subdomains enumeration

The enumeration exercise stars by mapping all available subdomains within a domain name. this will widen out attack surface and sometimes reveal hidden management backend panels or intranet web applications that the network administrators intended to protect through the old disgraced method of security through obscurity

There are lots of ways to enumerate subdomains:

* netcraft.
* Google
* Crawling.
* Brute force.
* Zone transfer.

# Fingerprint the webserver

One of simplest way to retrieve the web server version along with other useful information. This information is leaked through the HTTP headers in response to a trivial HTTP request to the web server.

Cookies are also an interesting resource that may reveal useful information, each technology has its default cookies names. Such as:

|  |  |
| --- | --- |
| Server | Cookie |
| PHP | PHPSESSID=XXXX |
| .NET | ASPSSESSIONIDYYYY=XXXXX |
| JAVA | JESSION=XXXX |

## Mitigation fingerprint the webserver webserver

you canmanually have obfuscated or using mod\_security module to change HTTP response information.

# Fingerprinting webserver modules

Although these same lines of how we fingerprinted the web server version, we can fingerprint what modules are installed and in use on the server.

Modules we are looking for are ISAPI modules (for IIS) or apache modules that may interfere with or alter our test results.

Changing ugly URL to URL-friendly done on apache with the **mod\_rewrite** module of **.htaccess**, while on IIS it is handled by **Ionic lsapi rewrite** or **Helion lsapi rewrite**.

## Mitigation fingerprint webserver modules

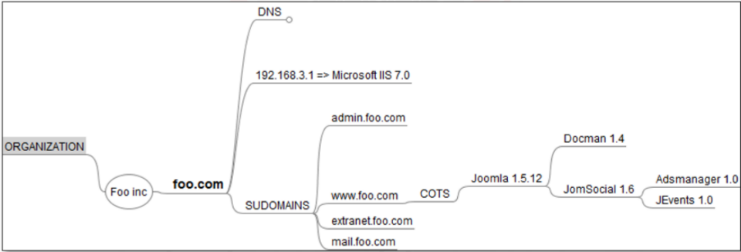
Searching engine friendly URLs are not a security feature at all.

# Fingerprinting frameworks and applications within webserver

Once we have a list of subdomains, we will apply the techniques that follow in this module to all of them.

Common applications may be:

* forums (Ex: phpBB, vBulletin).
* CMS’s (Ex: Joomla or Drupel).
* CRM’s, blogging platform (Ex: WordPress or Movable types).
* Social networking scripts and a number of other applications.



In this stage we determine common applications and which add-on could be already used.

## Fingerprinting custom applications

When you are not in front of a commonly available application, in this case will be to consider the overall scope of the application as the following:

* What is it for?
* Does it allow user registration?
* Does it have an administration panel?
* Does it take input from the user?
* What kind of input?
* Does it accept file upload?
* Does it use JavaScript or Ajax or Flash? And so on.

These questions can be answered by just visiting the website and taking notes of anything. Spidering (Crawling) the application is addressed too.

# Finding Hidden Paths and files within webserver

**DirBuster** is a mature OWASP project that allows us to crawl a website and also mount a directory or brute force discovery attack on the remote file system by probing each call and identifying a correct guess through the HTTP response code or web page content.

**Burp Suite Spider** is crawling (spidering) tool that built in burp suite.

**Note:** we need to pay attention to files and directories that were not retrieved by Burp suite spider.

# Information through Misconfigurations within webserver

Sometimes we find that the best way to retrieve relevant information about web application is to look for potential mistakes in web server configuration.

A quick and very common way to gather information, files, source code and misconfigurations is by looking for open directory listings.

# Enumerating user accounts from webserver

Among the resources we can enumerate in a website, usernames are another important bit of information that may turn up useful information when we have to audit an authentication mechanism.

A badly designed system can reveal sensitive information even if wrong credentials have been inserted.

For example, a web application could reveal information about existence of a user.

We can use the following tools to enumerate valid username:

* burp suite
* patator

# Automate information gathering

## DNSEnum tool

The purpose is to gather as much information possible about a domain. The tool currently performs the following operation:

* Get the host’s addresses (A record).
* Get the name servers (Threaded).
* Get the MX record (Threaded).
* Preform AXFR queries on name server (threaded).
* Get extra names and subdomain via google dorks
  + allinurl: -www site:domain
* brute force sub domains from file, can also perform recursion on domain that have NS records (all threaded).
* Calculate C class domain network ranges and perform WHOIS quires on them (threaded).
* Perform reverse lookup on net ranges (C class or/and WHOIS ranges) (threaded).

## dnsmap

is very old tool, still works great when it comes to sub domains enumeration and brute forcing.

DNS map uses the primary domain that we provide as a target and then forces all the subdomains by using:

* A directory file that comes with the tool.
* A word list file that the user makes.

## Shodan HQ

Similar to Google Hacking, great search engine that will be useful for information gathering process.

Shodan is a computer search engine that uses different approach from other search engine like Google, Yahoo, Bing, etc. indeed instead of crawling data in web pages, Shodan scans the entire internet and interrogates ports in order to gather information from the banners.

Shodan searches includes the following protocols:

* HTTPS/HTTP
* RDP
* SSH
* SNMP
* MySQL/MondoDB
* FTP
* Telnet
* More…

We can use it to search for:

* Devices with default username and password.
* Viewing the configuration of a device.
* Detect server versions, and more…

# Tips and tricks

* Information gathering is not a linear process but actually a cyclical process, when you find new organization projects, websites and subdomains, you have to repeat the whole investigation process for each of them
* Social engineering, if your target is Bob and you know that bob trusts Adam therefore, you can get to bob through Adam.

# Resources

## Google dork

* [Exploit-DB (Google Hacking Database)](https://www.exploit-db.com/google-hacking-database)
* [Google Search Operators](http://www.googleguide.com/advanced_operators_reference.html)
* [The Google Hacker’s Guide Understanding and Defending Against the Google Hacker](http://pdf.textfiles.com/security/googlehackers.pdf)