Web Information Gathering Mindmap

Automating reconnaissance enhances efficiency, scalability, consistency, and coverage. It speeds up tasks, reduces human error, and allows for comprehensive data collection across multiple targets. Automation also integrates easily with other tools for a streamlined process.

Eth3rn5l$./finalrecon.py --headers --whois --url http://target.com

tools

- FinalRecon
- Recon-ng
- theHarvester
- SpiderFoot
- OSINT Framework

automation

Web Archives

The Wayback Machine is a digital archive that lets users view past versions of websites. Since 1996, it captures snapshots of websites through automated crawlers, storing content like HTML, images, and scripts. Users can access these archived versions by entering a URL and selecting a date, offering a glimpse into how websites looked in the past.

Search Engine Discovery (OSINT): Using search engines to gather publicly available data for web reconnaissance, uncovering sensitive info, vulnerabilities, and other details about websites, organizations, or individuals.

Search Operators: Special commands to refine searches (e.g., site:, inurl:, filetype:, intitle:, etc.) for targeted data extraction.

Google Dorking: Using search operators to find hidden or sensitive information like login pages, exposed files, and database backups on websites.

Search Engine Discovery

Crawlers Extract Data for Reconnaissance:

Links: Internal and external links map website structure and uncover hidden pages.

Comments: Users may reveal sensitive information or vulnerabilities.

Metadata: Page details (titles, descriptions, keywords) offer context.

Sensitive Files: Crawlers find exposed files (backups, config, logs) containing confidential data.

crawling

robots.txt

A file that tells bots which areas of a site they can or cannot crawl.

Directives:

Disallow: Blocks access to specific paths.

Allow: Permits access to specific paths.

Crawl-delay: Sets delay between requests.

Purpose:

Prevents server overload.

Protects sensitive data.

Ensures legal compliance.

Web Recon: Reveals hidden paths, site structure, and potential vulnerabilities.

.well-known:

A standardized directory for critical website metadata, defined in RFC 8615. Accessible via the /.well-known/ path, it centralizes configuration files and security information. This simplifies discovery for browsers, apps, and security tools, allowing automatic access to files like security policies (e.g., https://example.com/.well-known/security.txt).

tools

- Burp Suite Spider
- OWASP ZAP (Zed Attack Proxy)
- Scrapy (Python Framework)
- Apache Nutch (Scalable Crawler)

ReconSpider.py http://target.com

After running ReconSpider.py, the data will be saved in a JSON file, results.json. This file can be explored using any text editor. Below is the structure of the JSON file produced:

results.json

Code: json

```
{
  "emails": [
    "lily.floid@inlanefreight.com",
    "cvs@inlanefreight.com",
    ...
```

Certificate Transparency (CT) logs are public, append-only ledgers tracking SSL/TLS certificate issuance. They help detect rogue certificates, hold Certificate Authorities accountable, and strengthen web security by ensuring transparency and oversight.

tools

- crt.sh
- Censys

Certificate Transparency Logs

Eth3rn5l$ curl -s "https://crt.sh/?q=facebook.com&output=json" | jq -r '.[] | select(.name_value | contains('dev')) | .name_value' | sort -u

Fingerprinting Techniques

Banner Grabbing: Analyzing server banners for software details.

HTTP Headers: Extracting info like server software from headers.

Probing Responses: Sending crafted requests to reveal technologies.

Page Content Analysis: Looking for clues in page structure and scripts.

fingerprint

Tools

- Wappalyzer
- BuiltWith
- WhatWeb
- Nmap
- Netcraft
- wafw00f

Eth3rn5l$ nikto -h target.com -Tuning b

Types of Virtual Hosting

Name-Based: Uses the HTTP Host header to distinguish websites, cost-effective, and easy to set up, but limited with certain protocols like SSL/TLS.

IP-Based: Assigns a unique IP to each site, offering better isolation but requiring multiple IPs, making it costly and less scalable.

Port-Based: Differentiates websites by port numbers on the same IP, useful when IPs are scarce but less user-friendly.

VHOST

tools

- gobuster
- Feroxbuster
- ffuf

Steps for Virtual Host Brute-Forcing:

Identify Target IP via DNS lookups or reconnaissance.

Prepare a Wordlist of potential virtual hostnames (e.g., SecLists).

Run Gobuster:

gobuster vhost -u http://<target_IP> -w <wordlist> --append-domain

-u: Target URL

-w: Wordlist file

--append-domain: Ensures full virtual hostnames

Useful Flags:

-t: Increase threads for faster scanning

-k: Ignore SSL/TLS errors

-o: Save output to a file

www.target.com

Data

whois

- Domain Name
- Registrar
- Registrant Contact
- Administrative Contact
- Technical Contact
- Creation and Expiration Dates
- Name Servers

DNS Zone Transfer

Digging DNS

A DNS zone transfer is a full copy of all DNS records for a domain and its subdomains from one name server to another, ensuring consistency and redundancy.

A DNS zone transfer (AXFR) involves a secondary server requesting a full zone copy from the primary server. The primary responds with the SOA record, followed by all DNS records. Once the transfer is complete, the secondary server acknowledges receipt.

exploiting zone transfer:

This command asks dig to perform a full zone transfer (AXFR) from the target.com DNS server. If misconfigured, it will reveal all DNS records, including subdomains.

Eth3rn5l$ dig axfr @nsztm1.digi.ninja target.com

tools

- dig
- nslookup
- dnsenum
- dnsrecon
- fierce
- theHarvester

Subdomain Enumeration

Wordlist Selection

- General-Purpose
- Targeted
- Custom

dnsenum --enum target.com -f worlist.txt -r

-> Iteration and Querying

-> DNS Lookup

-> Filtering and Validation

Web Information Gathering Mindmap

Automating reconnaissance enhances efficiency, scalability, consistency, and coverage. It speeds up tasks, reduces human error, and allows for comprehensive data collection across multiple targets. Automation also integrates easily with other tools for a streamlined process.

Eth3rn5l$./finalrecon.py --headers --whois --url http://target.com

tools

- FinalRecon
- Recon-ng
- theHarvester
- SpiderFoot
- OSINT Framework

automation

Search Engine Discovery

Web Archives

The Wayback Machine is a digital archive that lets users view past versions of websites. Since 1996, it captures snapshots of websites through automated crawlers, storing content like HTML, images, and scripts. Users can access these archived versions by entering a URL and selecting a date, offering a glimpse into how websites looked in the past.

Search Engine Discovery (OSINT): Using search engines to gather publicly available data for web reconnaissance, uncovering sensitive info, vulnerabilities, and other details about websites, organizations, or individuals.

Search Operators: Special commands to refine searches (e.g., site:, inurl:, filetype:, intitle:, etc.) for targeted data extraction.

Google Dorking: Using search operators to find hidden or sensitive information like login pages, exposed files, and database backups on websites.

Certificate Transparency / Logs

Certificate Transparency (CT) logs are public, append-only ledgers tracking SSL/TLS certificate issuance. They help detect rogue certificates, hold Certificate Authorities accountable, and strengthen web security by ensuring transparency and oversight.

tools

- crt.sh
- Censys

```
Eth3rn5l$ curl -s "https://crt.sh/?q=facebook.com&output=json" | jq -r '.[]
| select(.name_value | contains("dev")) | .name_value' | sort -u
```

crawling

Crawlers Extract Data for Reconnaissance:

Links: Internal and external links map website structure and uncover hidden pages.

Comments: Users may reveal sensitive information or vulnerabilities.

Metadata: Page details (titles, descriptions, keywords) offer context.

Sensitive Files: Crawlers find exposed files (backups, config, logs) containing confidential data.

robots.txt

A file that tells bots which areas of a site they can or cannot crawl.

Directives:

Disallow: Blocks access to specific paths.

Allow: Permits access to specific paths.

Crawl-delay: Sets delay between requests.

Purpose:

Prevents server overload.

Protects sensitive data.

Ensures legal compliance.

Web Recon: Reveals hidden paths, site structure, and potential vulnerabilities.

.well-known: A standardized directory for critical website metadata, defined in RFC 8615. Accessible via the /.well-known/ path, it centralizes configuration files and security information. This simplifies discovery for browsers, apps, and security tools, allowing automatic access to files like security policies (e.g., https://example.com/.well-known/security.txt).

tools

- Burp Suite Spider
- OWASP ZAP (Zed Attack Proxy)
- Scrapy (Python Framework)
- Apache Nutch (Scalable Crawler)

ReconSpider.py http://target.com

After running ReconSpider.py, the data will be saved in a JSON file, results.json. This file can be explored using any text editor. Below is the structure of the JSON file produced:

results.json

Code: json

```
{
"emails": [
"lily.floid@inlanefreight.com",
"cvs@inlanefreight.com",
...
```

Fingerprint

Fingerprinting Techniques

Banner Grabbing: Analyzing server banners for software details.

HTTP Headers: Extracting info like server software from headers.

Probing Responses: Sending crafted requests to reveal technologies.

Page Content Analysis: Looking for clues in page structure and scripts.

Tools

- Wappalyzer
- BuiltWith
- WhatWeb
- Nmap
- Netcraft
- wafw00f

Eth3rn5l$ nikto -h target.com -Tuning b

www.target.com

VHOST

Types of Virtual Hosting

Name-Based: Uses the HTTP Host header to distinguish websites, cost-effective, and easy to set up, but limited with certain protocols like SSL/TLS.

IP-Based: Assigns a unique IP to each site, offering better isolation but requiring multiple IPs, making it costly and less scalable.

Port-Based: Differentiates websites by port numbers on the same IP, useful when IPs are scarce but less user-friendly.

tools

- gobuster
- Feroxbuster
- ffuf

Steps for Virtual Host Brute-Forcing:

Identify Target IP via DNS lookups or reconnaissance.

Prepare a Wordlist of potential virtual hostnames (e.g., SecLists).

Run Gobuster:

```
gobuster vhost -u http://<target_IP> -w <wordlist> --append-domain
```

-u: Target URL

-w: Wordlist file

--append-domain: Ensures full virtual hostnames

Useful Flags:

-t: Increase threads for faster scanning

-k: Ignore SSL/TLS errors

-o: Save output to a file

Data

whois

- Domain Name
- Registrar
- Registrant Contact
- Administrative Contact
- Technical Contact
- Creation and Expiration Dates
- Name Servers

DNS Zone Transfer

A DNS zone transfer is a full copy of all DNS records for a domain and its subdomains from one name server to another, ensuring consistency and redundancy.

A DNS zone transfer (AXFR) involves a secondary server requesting a full zone copy from the primary server. The primary responds with the SOA record, followed by all DNS records. Once the transfer is complete, the secondary server acknowledges receipt.

exploiting zone transfer:

This command asks dig to perform a full zone transfer (AXFR) from the target.com DNS server.
If misconfigured, it will reveal all DNS records, including subdomains.

```
Eth3rn5l$ dig axfr @nsztm1.digi.ninja target.com
```

Digging DNS

tools

- dig
- nslookup
- dnsenum
- dnsrecon
- fierce
- theHarvester

Subdomain Enumeration

Wordlist Selection

- General-Purpose
- Targeted
- Custom

```
dnsenum --enum target.com -f wordlist.txt -r
```

-> Iteration and Querying

-> DNS Lookup

-> Filtering and Validation