TRAINING DAY 17 REPORT

23 July 2025

Understand Footprinting

Today, I learn some important concepts

Web Crawler

A **web crawler** (also known as a **spider** or **bot**) is a tool or program that automatically **browses the web** to **collect and index data** from websites. Web crawlers are essential for:

- Search engines (e.g. Googlebot)
- OSINT & footprinting
- Data mining & scraping
- SEO analysis
- Competitive intelligence

How Web Crawlers Work

1. **Start with a URL** (called a seed URL)
2. **Fetch the HTML content** of the page
3. **Parse the page** to extract data and links
4. **Follow extracted links** and repeat the process
5. **Store the collected data** in a structured format (JSON, database, CSV, etc.)

Wget Mirroring

wget is a command-line utility used to **download files** from the web. It can also **mirror entire websites recursively**, preserving the directory structure and converting links for offline use.

Basic Wget Mirror Command

```
wget --mirror --convert-links --adjust-extension --page-requisites --no-parent
https://example.com
```

Mirroring with Httrack

HTTrack is a powerful, free, and open-source tool for **mirroring websites** (i.e., downloading a complete copy for offline viewing). It's especially useful for **OSINT**, research, backups, and analyzing web structures without interacting with the live site beyond normal HTTP requests.

Temp mails

Temp mails (temporary email addresses) are **disposable email accounts** that can be used for **short-term or anonymous communication**. They're often used to:

- Avoid spam when signing up for a service
- Register on websites without giving your real email
- Test account creation processes or email functionality
- Stay anonymous in **OSINT investigations or red teaming**

WHOIS Lookup

WHOIS Lookup is a method to query databases that store **registration information about domain names** (and sometimes IP addresses). It reveals details like:

- Domain owner (registrant)
- Registrar (the company managing the domain)
- Registration and expiration dates
- Contact info (email, phone) — sometimes redacted for privacy
- Name servers
- Status of the domain (active, expired, locked, etc.)

DNS Resource Record

A **DNS Resource Record (RR)** is a fundamental data element in the Domain Name System (DNS). It defines information about a domain name, such as its IP address, mail server, or other attributes.

- Each RR consists of several fields that specify **type, value, and TTL** (time-to-live).
- RRs are stored in DNS zone files or databases.
- They tell DNS servers how to respond to queries.

Important Search Engines

Search Engine	Notes
Google	Most popular, vast index, advanced search operators
Bing	Microsoft’s search engine, different indexing, good for some regional searches
Yahoo! Search	Powered by Bing, alternative UI
DuckDuckGo	Privacy-focused, no user tracking
Ecosia	Privacy-respecting, plants trees per search
Yandex	Popular in Russia and CIS countries, good for Russian content

OSINT & Specialized Search Engines

Search Engine	Focus / Features
Shodan	Internet-connected devices (IoT, servers, webcams, etc.)
Censys	Internet-wide scanning, certificates, devices
ZoomEye	Similar to Shodan, global device and service discovery
Have I Been Pwned	Check if emails or accounts are compromised
Wayback Machine	Archive.org — archived web pages over time
Maltego (Transform)	OSINT framework with integrated search/transforms
IntelTechniques	OSINT search tools (social media, people, domains)
PublicWWW	Source code search engine (find scripts, tracking codes)
VirusTotal	Search URLs, files, and domains for malware and threats
Greynoise	Internet noise and threat intelligence

Academic and Data Search Engines

Search Engine	Focus
Google Scholar	Academic papers, theses, patents
Microsoft Academic	Academic articles and citations
Semantic Scholar	AI-powered academic research
Data.gov	US government open data
WolframAlpha	Computational knowledge engine

Social Media & People Search

Search Engine / Tool	Focus
Pipl	People search, social profiles
Spokeo	Aggregate social and public records
Social Searcher	Real-time social media search
Social Mention	Social media monitoring