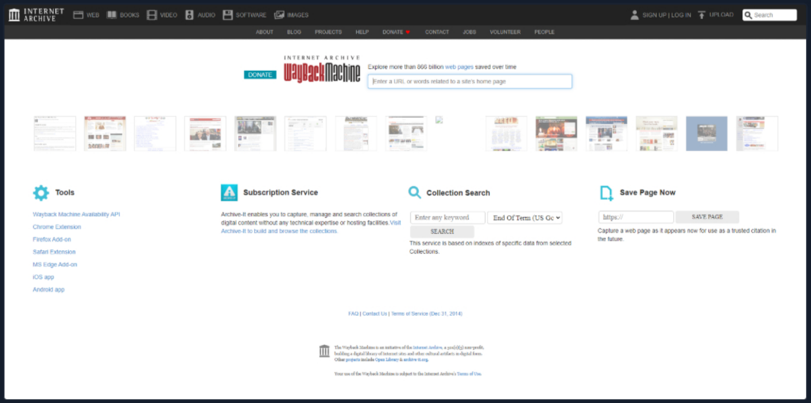
Web Archives

In the fast-paced digital world, websites come and go, leaving only fleeting traces of their existence behind. However, thanks to the Internet Archive's Wayback Machine, we have a unique opportunity to revisit the past and explore the digital footprints of websites as they once were.

What is the Wayback Machine?


INTERNET ARCHIVE
WEB
BOOKS
VIDEO
AUDIO
SOFTWARE
IMAGES
SIGN UP | LOG IN
UPLOAD
Search
ABOUT
BLOG
PROJECTS
HELP
DONATE
CONTACT
JOBS
VOLUNTEER
PEOPLE
DONATE
INTERNET ARCHIVE WaybackMachine
Explore more than 866 billion web pages saved over time
Enter a URL or words related to a site's home page
Tools
Wayback Machine Availability API
Chrome Extension
Firefox Add-on
Safari Extension
MS Edge Add-on
iOS app
Android app
Subscription Service
Archive-It enables you to capture, manage and search collections of digital content without any technical expertise or hosting facilities. Visit Archive-It to build and browse the collections.
Collection Search
SEARCH
This service is based on indexes of specific data from selected Collections.
Save Page Now
https://
SAVE PAGE
Capture a web page as it appears now for use as a trusted citation in the future.
FAQ | Contact Us | Terms of Service (Dec 31, 2014)

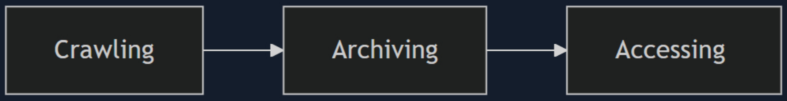
`The Wayback Machine` is a digital archive of the World Wide Web and other information on the Internet. Founded by the Internet Archive, a non-profit organization, it has been archiving websites since 1996.

It allows users to "go back in time" and view snapshots of websites as they appeared at various points in their history. These snapshots, known as captures or archives, provide a glimpse into the past versions of a website, including its design, content, and functionality.

How Does the Wayback Machine Work?

The Wayback Machine operates by using web crawlers to capture snapshots of websites at regular intervals automatically. These crawlers navigate through the web, following links and indexing pages, much like how search engine crawlers work. However, instead of simply indexing the information for search purposes, the Wayback Machine stores the entire content of the pages, including HTML, CSS, JavaScript, images, and other resources.

The Wayback Machine's operation can be visualized as a three-step process:



1. `Crawling`: The Wayback Machine employs automated web crawlers, often called "bots," to browse the internet systematically. These bots follow links from one webpage to another, like how you would click hyperlinks to explore a website. However, instead of just reading the content, these bots download copies of the webpages they encounter.
2. `Archiving`: The downloaded webpages, along with their associated resources like images, stylesheets, and scripts, are stored in the Wayback Machine's vast archive. Each captured webpage is linked to a specific date and time, creating a historical snapshot of the website at that moment. This archiving process happens at regular intervals, sometimes daily, weekly, or monthly, depending on the website's popularity and frequency of updates.
3. `Accessing`: Users can access these archived snapshots through the Wayback Machine's interface. By entering a website's URL and selecting a date, you can view how the website looked at that specific point. The Wayback Machine allows you to browse individual pages and provides tools to search for specific terms within the archived content or download entire archived websites for offline analysis.

The frequency with which the Wayback Machine archives a website varies. Some websites might be archived multiple times a day, while others might only have a few snapshots spread out over several years. Factors that influence this frequency include the website's popularity, its rate of change, and the resources available to the Internet Archive.

It's important to note that the Wayback Machine does not capture every single webpage online. It prioritizes websites deemed to be of cultural, historical, or research value. Additionally, website owners can request that their content be excluded from the Wayback Machine, although this is not always guaranteed.

Why the Wayback Machine Matters for Web Reconnaissance

The Wayback Machine is a treasure trove for web reconnaissance, offering information that can be instrumental in various scenarios. Its significance lies in its ability to unveil a website's past, providing valuable insights that may not be readily apparent in its current state:

1. `Uncovering Hidden Assets and Vulnerabilities`: The Wayback Machine allows you to discover old web pages, directories, files, or subdomains that might not be accessible on the current website, potentially exposing sensitive information or security flaws.
2. `Tracking Changes and Identifying Patterns`: By comparing historical snapshots, you can observe how the website has evolved, revealing changes in structure, content, technologies, and potential vulnerabilities.
3. `Gathering Intelligence`: Archived content can be a valuable source of OSINT, providing insights into the target's past activities, marketing strategies, employees, and technology choices.

Cheat Sheet

? Go to Questions

Table of Contents

Introduction

WHOIS

DNS & Subdomains

Fingerprinting

Crawling

Search Engine Discovery

Web Archives

Automating Recon

Skills Assessment

My Workstation

OFFLINE

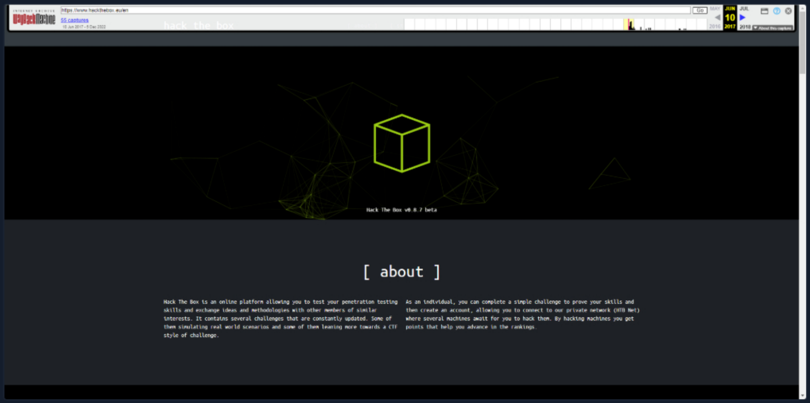
Start Instance

∞ / 1 spawns left

4. Stealthy Reconnaissance: Accessing archived snapshots is a passive activity that doesn't directly interact with the target's infrastructure, making it a less detectable way to gather information.

Going Wayback on HTB

We can view the first archived version of HackTheBox by entering the page we are looking for into the Wayback Machine and selecting the earliest available capture date, being `2017-06-10 @ 04h23:01`



https://www.hackthebox.eu/en

55 captures

10 Jun 2017 - 5 Dec 2022

Hack The Box v0.8.7 beta

[about]

Hack The Box is an online platform allowing you to test your penetration testing skills and exchange ideas and methodologies with other members of similar interests. It contains several challenges that are constantly updated. Some of them simulating real world scenarios and some of them leaning more towards a CTF style of challenge.

As an individual, you can complete a simple challenge to prove your skills and then create an account, allowing you to connect to our private network (HTB Net) where several machines await for you to hack them. By hacking machines you get points that help you advance in the rankings.

Connect to Pwnbox

Your own web-based Parrot Linux instance to play our labs.

Pwnbox Location

UK 182ms

Terminate Pwnbox to switch location

Start Instance

∞ / 1 spawns left

Waiting to start...

Enable step-by-step solutions for all questions

Questions

Answer the question(s) below to complete this Section and earn cubes!

Cheat Sheet

+ 1 How many Pen Testing Labs did HackTheBox have on the 8th August 2018? Answer with an integer, eg 1234.

ns.inlanefreight.htb

Submit

+ 0 How many members did HackTheBox have on the 10th June 2017? Answer with an integer, eg 1234.

2

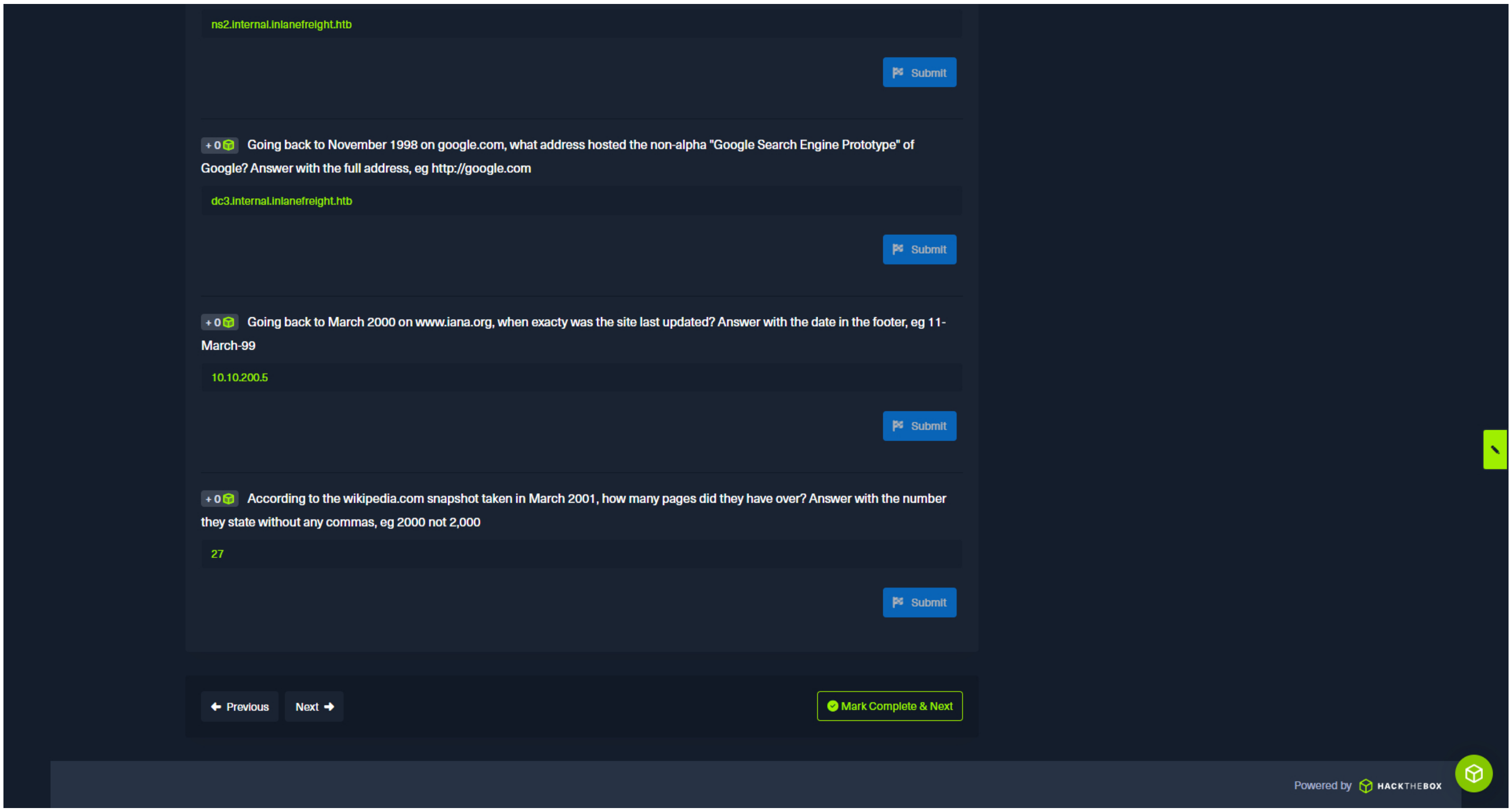
Submit

+ 0 Going back to March 2002, what website did the facebook.com domain redirect to? Answer with the full domain, eg http://www.facebook.com/

ZONE_TRANSFER{87o2z3cno7zsoiedznxoi82z3o47xzhoi}

Submit

+ 0 According to the paypal.com website in October 1999, what could you use to "beam money to anyone"? Answer with the product name, eg My Device, remove the ™ from your answer.



ns2.internal.inlanefreight.htb

+ 0 Going back to November 1998 on google.com, what address hosted the non-alpha "Google Search Engine Prototype" of Google? Answer with the full address, eg http://google.com

dc3.internal.inlanefreight.htb

+ 0 Going back to March 2000 on www.iana.org, when exacty was the site last updated? Answer with the date in the footer, eg 11-March-99

10.10.200.5

+ 0 According to the wikipedia.com snapshot taken in March 2001, how many pages did they have over? Answer with the number they state without any commas, eg 2000 not 2,000

27