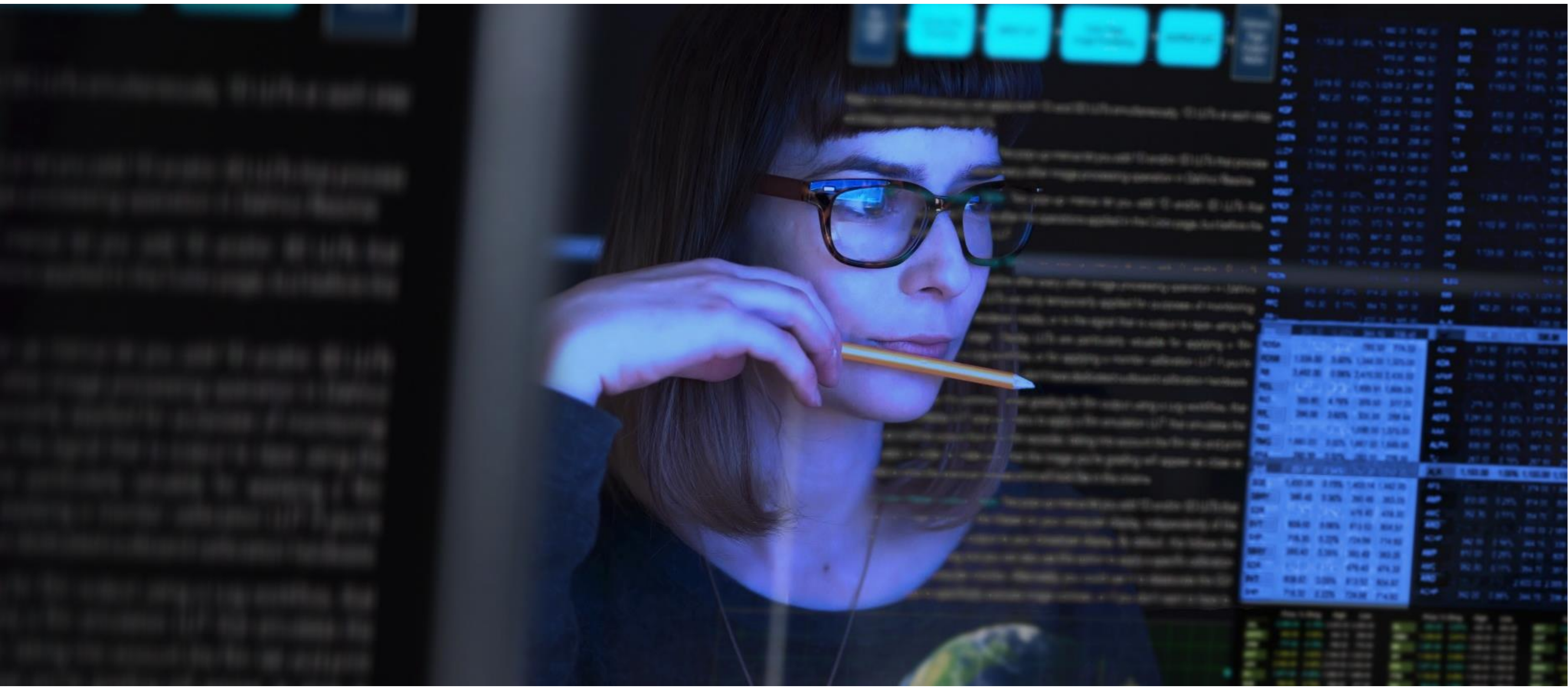
How to Check Your Web Archive for Dark Web Contraband



Web archives, such as the Internet Archive's Wayback Machine, focus on saving and replaying web pages. However, these archives can inadvertently be used to host and share dark web contraband. This guide is designed to help archivists and librarians safely identify and remove contraband material from their archives.

Dark web content in the surface web archives

As part of MITRE's independent reseach and development program, the Venom Project is working to understand how to best archive the non-illicit portions of the dark web using the tools, techniques, and practices sourced from surface web archiving (e.g., by institutions such as the Internet Archive). We developed this quick introduction to help archivists and librarians identify the contraband being hosted in their archives from the dark web to limit its spread **without requiring them to view the contraband**.

This how-to guide is motivated by our discovery of dark web content in the Wayback Machine. As part of a broader research project, we confirmed that dark web content appeared in the Wayback Machine, some of which was illicit content that the Wayback Machine does not wish to host. Consequently, the platform removed access to the content and is preventing further illicit content from being archived and served to the surface web. We want to share guidance on how other internet memory institutions can similarly protect their archives.

VENOM

Resources

The Venom Project is creating tools, techniques, and recommendations for archiving the dark web.

This work is funded by the not-for-profit MITRE Corporation and its independent R&D program.

MITRE

How to Check Your Web Archive for Dark Web Contraband

How does contraband get into the web archives?

Dark web proxies, such as tor2web, provide a way to browse the dark web from the surface web. Users of tor2web and similar proxies enter a `.onion` URL into a form. The proxy service requests the dark web page using a Tor browser and presents it to the user. A bad actor needs to perform three steps to get contraband into an archive.

First, the bad actor needs to identify the `.onion` URL they wish to save. **Second**, the bad actor needs to identify a dark web proxy that will connect the surface web to the dark web site. **Third**, the bad actor asks the archive to capture it (e.g., via a service such as the Wayback Machine's "Save Page Now" feature).

While this method will allow institutional archiving of undesirable dark web content, it can also be used to inject undesirable content from the surface web in much the same way since Tor (and, therefore, dark web proxies) can be used to access surface *and* dark web resources.

Finding onions in archives

To check archives for dark web content (i.e., `.onion` URLs), archivsts and librarians should follow a few steps.

First, the archivists should identify a list of dark web proxies to investigate. These proxies are available at StackOverflow and other resources discoverable by searching on Google or another surface web search engine.

Second, the archivists should search for content using the wild card URL look-up feature in the archive's replay mechanism (e.g., Wayback Machine). For example, an archivist can search for tor2web content by entering the following URL Into the Wayback Machine search form: `tor2web.onionsearchengine.com/*`

The replay mechanism will return a list of the proxy-provided material present in the archives. Depending on the server, archivists may need to add a suffix to the domain and manually search instead, such as:

`*.onion.ly`

Alternatively, if the replay mechanism uses a CDX index, an archivist can search the index for the archived proxy URLs using the same wildcard search:

```
http://web.archive.org/search/cdx?
url=proxy.onionsearchengine.com/*
```

Or

```
http://web.archive.org/search/cdx?
url=*.onion.ly/
```

At no point should an archivist click on or otherwise view the URLs listed since they may contain illicit contraband.

Third, the archivist should extract the hostname from the returned proxy URLs. Depending on the proxy service, the `.onion` hostnames may be a prefix, post-fix, or a URL parameter.

Fourth, the archivist should look up the hostname in lists of known contraband-serving URLs. See the next section for details.

Identifying Contraband

Archivists and librarians should not view dark web content to confirm it is contraband! Instead, archivists should continue from Step 4. to look up the extracted list of dark

How to check your archive for dark web content:

1. Start with a list of dark web proxies, such as tor2web
2. For each proxy, use the wildcard search function in the replay mechanism to get an index of archived proxy URLs
3. DO NOT click on any discovered URLs
4. Extract the `.onion` hostname and URLs from the archived proxy URLs; these will be prefixes, postfixes, or URL parameters in the proxy URL
5. Look up the `.onion` hostname and/or URL in a list of URLs known to serve contraband
6. Prevent access to the contraband material *but do not delete the WARCs*
7. Notify the appropriate authorities, e.g., (https://report.cybertip.org/)

How to Check Your Web Archive for Dark Web Contraband

web hostnames to determine if they serve contraband content. Ahmia is a dark web search engine that has a list of banned hosts stored as an MD5 hash. The hash is available here at `https://ahmia.fi/blacklist/banned`. Archivists should take the MD5 hash of the hostname and check the Ahmia (or other) list of banned hosts. *If the hash appears in the list of banned URLs, it is likely that dark web contraband exists in your archive*.

Fifth, if contraband-serving hosts are detected, the archivist should immediately restrict access to that archived content by preventing public web access of the mementos.

Sixth, as required by U.S. law, the archivist should notify the appropriate authorities that there may be contraband material on the archive's servers. Contact the 24-hour toll line at `https://www.missingkids.org/footer/contactus`. The authorities will provide official guidance and will work to verify whether contraband material has been captured.

Recommendations for archivists/librarians

First and foremost, we want to reiterate that archivists should not independently verify contraband content by clicking on and viewing the material. Archivists should remove public access to suspected contraband in the archives (but not delete the content!) and contact the relevant authorities for help.

Archives should also consider preventing crawl, archiving, and serving content from dark web proxies.

While the Venom team supports dark web archiving, the practice of archiving the dark web carries inherent risk, policy challenges, and technical challenges. As such, surface web archives should not host dark web content. The practice of dark web archiving should be left to specialized dark web archives with access to archived material limited to the dark web (i.e., dark web archives should only be available at their own `.onion` URLs).

If you have **questions** about checking your archives for dark web material or wish to understand more about dark web archiving, please contact the Venom Research Team at venom@mitre.org.

Recommendations:

1. Do not verify, click on, or view dark web content in your archive, especially suspected contraband
2. Remove public web access but do not delete suspected contraband
3. Contact authorities immediately upon identifying suspected contraband
4. Consider disallowing the archiving and viewing of content via dark web proxies
5. Consult the Venom Research Team with questions about dark web archiving

MITRE's mission-driven teams are dedicated to solving problems for a safer world. Through our public-private partnerships and federally funded R&D centers, we work across government and in partnership with industry to tackle challenges to the safety, stability, and well-being of our nation.

For information about MITRE's web archiving and dark web expertise and capabilities, please contact the Venom Research Team, venom@mitre.org.

MITRE