Scientific Working Group on Digital Evidence

Best Practices for Acquiring Online Content

21-F-001-1.1

Disclaimer Regarding Use of SWGDE Documents

SWGDE documents are developed by a consensus process that involves the best efforts of relevant subject matter experts, organizations, and input from other stakeholders to publish suggested best practices, practical guidance, technical positions, and educational information in the discipline of digital and multimedia forensics and related fields. No warranty or other representation as to SWGDE work product is made or intended.

SWGDE requests notification by e-mail before or contemporaneous to the introduction of this document, or any portion thereof, as a marked exhibit offered for or moved into evidence in such proceeding. The notification should include: 1) The formal name of the proceeding, including docket number or similar identifier; 2) the name and location of the body conducting the hearing or proceeding; and 3) the name, mailing address (if available) and contact information of the party offering or moving the document into evidence. Subsequent to the use of this document in the proceeding please notify SWGDE as to the outcome of the matter. Notifications should be submitted via the SWGDE Notice of Use/Redistribution Form or sent to secretary@swgde.org.

From time to time, SWGDE documents may be revised, updated, or sunsetted. Readers are advised to verify on the SWGDE website (https://www.swgde.org) they are utilizing the current version of this document. Prior versions of SWGDE documents are archived and available on the SWGDE website.

Redistribution Policy

SWGDE grants permission for redistribution and use of all publicly posted documents created by SWGDE, provided that the following conditions are met:

1. Redistribution of documents or parts of documents must retain this SWGDE cover page containing the Disclaimer Regarding Use.
2. Neither the name of SWGDE nor the names of contributors may be used to endorse or promote products derived from its documents.
3. Any reference or quote from a SWGDE document must include the version number (or creation date) of the document and also indicate if the document is in a draft status.

Requests for Modification

SWGDE encourages stakeholder participation in the preparation of documents. Suggestions for modifications are welcome and must be submitted via the SWGDE Request for Modification Form or forwarded to the Secretary in writing at secretary@swgde.org. The following information is required as a part of any suggested modification:

a) Submitter's name
b) Affiliation (agency/organization)
c) Address

d) Telephone number and email address
e) SWGDE Document title and version number
f) Change from (note document section number)
g) Change to (provide suggested text where appropriate; comments not including suggested text will not be considered)
h) Basis for suggested modification

Intellectual Property

Unauthorized use of the SWGDE logo or documents without written permission from SWGDE is a violation of our intellectual property rights.

Individuals may not misstate and/or over represent duties and responsibilities of SWGDE work. This includes claiming oneself as a contributing member without actively participating in SWGDE meetings; claiming oneself as an officer of SWGDE without serving as such; claiming sole authorship of a document; use the SWGDE logo on any material and/or curriculum vitae.

Any mention of specific products within SWGDE documents is for informational purposes only; it does not imply a recommendation or endorsement by SWGDE.

Scientific Working Group on Digital Evidence

Best Practices for Acquiring Online Content

Table of Contents

1. Purpose

The purpose of this document is to provide best practices and considerations for acquiring and preserving digital evidence from online content.

2. Scope

For the scope of this document, online content refers to public facing data including, but not limited to, websites, streaming services, and communication platforms.

This document is not meant to address searches conducted pursuant to a warrant, or acquisition of data as described in *SWGDE Best Practices for Digital Evidence Acquisition from Cloud Service Providers*. Additionally, this document is focused on the acquisition for investigative purposes, not the identification or analysis of such data.

This document is directed at those needing to preserve internet content for future use in the context of legal, administrative, or similar proceedings, including forensic examiners, investigators, attorneys, etc. Those conducting such collections should have a working knowledge of basic information technology and foundational computer forensics principles.

3. Limitations and Considerations

The examiner should be prepared to explain the discovery of evidentiary online content, the acquisition process, and the reliability of the principles and methods implemented. The examiner should also have the ability to distinguish content available through legal means from an Electronic Service Provider ("ESP") or solely through live capture of privately hosted content. This document is intended to be tool agnostic; references to specific tools are for demonstrative purposes only and are not an endorsement by SWGDE. It is not intended as a step-by-step guide, nor should it be construed as legal advice.

3.1 Principles of Digital Evidence

Digital evidence is governed by three basic principles: relevance, reliability and sufficiency. It should be possible to demonstrate that all the material acquired is relevant to the investigation, containing information of value in assisting the investigation of the particular incident, and that there is a good reason for it to have been acquired. All processes mentioned here should be auditable and repeatable, so that the results of applying such processes should be reproducible. Reproducibility is established when the same test results are produced using the same measurement method, and using different tools / utilities on the same dynamic or ephemeral website, but can be reproduced at any time after the original test. Sufficient material should be collected to allow a proper investigation to be conducted.

3.2 Accessibility

Acquiring results can be limited by protected areas of a site, hidden URLs, dynamic content generated based upon browser identity, or software limitations due to complex site design. For restricted access sites, consider alternative investigative tactics in accordance with agency policies.

3.3 Supplemental Preservation

Preservation of web content from the Electronic Service Provider through legal process can verify captured content and supplement content using standard collection methods. Refer to *SWGDE Best Practices for Digital Evidence Acquisition from Cloud Service Providers*.

3.4 Evidence Contamination

Prior to executing methods or tools on a target site, the examiner should understand the effects live access and collection tools may have on the integrity of the data such as superficial web traffic and browser fingerprints, access to the content from a known network (e.g., government attributable network), allowing the target site to identify browser strings and MAC addresses, or using login credentials which may lead to detection or spoilage of the investigation.

3.5 Legal Authority

As with all digital evidence acquisitions, collectors must comply with their agency policies and procedures and ensure appropriate legal authority is established prior to conducting the acquisition. Examiners should consult with legal counsel to understand their legal authorities before the need for legal process arises. Advance development of procedures and use of working copies of legal process are valuable practices that save time, especially in exigent situations.

4. Preparations

The dynamic nature of web content and its constantly evolving platforms, protocols, technologies, and tools often requires multiple procedures for acquisition. Examiners should select the appropriate course of action based on the goal of the collection, available resources, and their knowledge and understanding of the circumstances. This document identifies the high-level considerations and steps for the collection of web content.

4.1 Configuration

- The device (i.e., hardware) and its installed applications (i.e., software), hereinafter "device," used to access the target website should be equipped with the tools necessary for acquiring online content and void of extraneous applications which may interfere with the acquisition process or alter the data acquired. For instance, an enabled browser plug-in may hinder a tool's ability to fetch data (e.g., pop-up blocked) or it may alter the data displayed within the web browser (e.g., translator).
- The device should be configured to mimic its target audience in accordance with the investigation. This can be accomplished by using regional 'settings' (e.g., keyboard layout, browser agent strings, IP address) to ensure targeted content is presented and acquired accordingly, with regard to the goals of the investigation and to avoiding detection.

- Use of a sanitized hard drive or virtual machine can avoid cross-contamination, such as extraneous software and pre-existing data, which may alter results.
- Web servers often log incoming Internet Protocol (IP) addresses and associated browser header information. This information is used to identify a user and can result in different content being displayed depending upon the user's location and software used to connect to the site. For this reason, the use of Virtual Private Networks ("VPNs") or proxy servers is recommended for the purpose of presenting a profile and location consistent with the goals of the investigation. In covert investigation cases, examiners should avoid using agency internet connections and take measures to obfuscate the origin of the investigator's access to the site.
- Consider any previous intelligence reports regarding what DNS server that the Person or Organization of Interest is using, and then resolve domain names with it before acquiring any online content, to help ensure the accuracy of the target websites.

4.2 Content Volatility

Multiple acquisitions may be necessary to document content changes due to volatility or tampering.

4.3 Tool Validation

Refer to *SWGDE Minimum Requirements for Testing Tools used in Digital and Multimedia Forensics* for additional information.

5. Goals of Acquisition

Collection should be driven by the intended use of the captured information including, but not limited to, the following:

- Identification of the web server location and system administrator.
- Tracking web traffic and identifying authenticated members.
- Preservation of Graphical User Interface ("GUI") content.
- Preservation of text and graphical content.
- Preservation of hidden configuration information and metadata.
- Attribution of public information to a subject.
- Documentation of embedded links for communication and associated social networks.
- Identification of malicious code or intent.

6. Categories

Online content should be identified in the following categories to establish priority and best acquisition methods to be implemented. This determination can be accomplished by identifying

if the content is constantly changing (i.e., dynamic) and if that content is available for a limited time (i.e., ephemeral).

6.1 Static Websites

Static websites contain fixed content and display the same content to every user. Static websites are commonly coded in HTML and are uniform across visitor credentials, web browser, location, date, and time.

6.2 Dynamic Websites

Dynamic websites can display various types of content (e.g., static, dynamic, ephemeral), provide user interaction, and are database-driven, and, therefore, the contents of these websites are often dependent on external factors including, but not limited to, visitor credentials, browser, location, date, and time. Dynamic websites are commonly found with social networking platforms. It should be noted that content that is known to violate the service's Terms of Service ("TOS") is likely to be removed by the ESP. Due to the unpredictability and rate of change for this type of web content, websites should be prioritized by their relative volatility.

6.3 Ephemeral Websites

Ephemeral content hosted by websites is a live event and by design not saved. Ephemeral content is commonly found on websites that host live streaming services and volatile communications. This content should be triaged with a higher priority, given its relative volatility. Please reference *SWGDE Guidelines for Video Evidence Canvassing and Collection* document for more in-depth guidelines.

7. Documentation

Beyond standard documentary procedures for acquisition of electronic evidence, the following methods are recommended for online content, according to the goal and category of content being acquired.

7.1 Documenting Screen Captures (Stills and Motion)

Screen captures preserve the GUI of a website and include both still images and video recordings of the website access. Screen captures can be used to provide additional context and a form of prima facie evidence. All available information displayed should be included in the screen capture such as URL and date and time. Preference should be given to video capture of the screen and can be used in conjunction with still images.

Still images should be captured at the shortest and most focused viable interval in order to create a seamless and complete record of the content displayed. The shorter the interval, the less likely that contents will be omitted. This method is beneficial when acquiring ephemeral websites and for use as demonstrative exhibits to present the information.

7.2 Format

Online content from a target site such as raw media, webpages, scripts, xrefs and similar links should be acquired in its native form, file formats, and language. Documentary evidence generated by the examiner during the acquisition process, to include extended image format metadata, should be preserved in standard formats.

7.3 Hashing

NIST-approved secure hash algorithms should be used to calculate digests to validate and uniquely identify the entire collection data set, as well as the individual content (files) acquired, including graphical contents, underlying browser data, and documentary evidence. Refer to *SWGDE Position on the Use of MD5 and SHA1 Hash Algorithms in Digital and Multimedia Forensics*.

7.4 Network Documentation

If deemed relevant, network traffic with the target site(s) should be logged during live access. This can be accomplished using the following methods:

7.4.1 Packet Captures

Use a packet analyzer, packet sniffer, or a web proxy to intercept multiple redirecting URLs and log traffic that passes over the local network, to identify and authenticate suspect content in network traffic. By capturing packets and identifying DNS queries, IP addresses, port numbers, timestamps, digital certificates, and HTTP headers, this information can record activity that is not visually apparent in the operation of the site or interaction with the content, as well as provide an account of content that is gathered from the target site(s). Packet capture can also help ascribe discovered online content to a particular author, remote host, or person.

7.4.2 Open network connections

Open network connections can be logged (e.g., use of the netstat command) to determine additional sources of embedded online content. Associated network connections, routing tables, interface statistics, masquerade connections, and multicast memberships should be printed to better document network video streams.

Complementing a passive packet capture is the optional use of a reliable tool for active remote host exploration, using raw IP packets to determine what services (application name and version) a host is offering, what operating system (and OS versions) it is running, what type of packet filters / firewalls are in use, monitor the host or service uptime, and dozens of other characteristics. The expected output should include the port numbers (i.e., from 0 to 65535) and protocol (e.g., TCP, UDP, SCTP), service name, and state (e.g., open, filtered, closed, or unfiltered). Additionally, the use of traceroute (e.g., UDP, TCP, ICMP) may determine or guess the physical location and network structure of the target host.

7.5 Collection Documentation

The following components should be documented during the acquisition process where possible:

- Uniform Resource Locator (URL) to include the protocol (blue), domain (orange), subdomains (green), subpages (purple), path (red), and session information when available which can include valuable information, such as timestamps and parameters, to display the site. It is important to capture the sequence of URLs and ensure it is not truncated.

  https://mail.google.com/mail/u/0/#inbox

- Domain registration including, but not limited to, creation date, expiration date, status, nameservers, and contact information. This information can be captured using ICANN.org.
- History of website, if applicable, from archival resources such as Archive.org and archive.today.
- Document physical location, access IP address, browser, dates, timestamps, and local time zone of website access.

8. Acquisition

8.1 Methods

In prioritizing collection or acquisition of potential digital evidence, it is imperative to understand the reason that the potential digital evidence is being collected or acquired. As a general principle, one should attempt to maximize the amount of data preserved by collection and acquisition actions. However, it may be necessary to prioritize items by volatility, relevance, or potential evidentiary weight. Items of high relevance or potential evidence value are those that are most likely to contain data relating directly to the incident under investigation.

Prioritization by volatility is only applicable if the specific circumstances of the case being investigated require this. Collection of potential evidence can be categorized by three methods: the use of "Utilities," collection through "Web Browser / Plug-ins / Extensions," and the simple use of "Screenshots."

The acquisition of online content should be accomplished using a combination of the following methods, in order of most inclusive. The level of analysis required depends on the request and the specifics of the investigation. Each acquisition level has its own corresponding skill set, tool set, and risk. The levels are:

8.1.1 Utilities

Application Programming Interfaces

Many websites, particularly social media platforms, have Application Programming Interfaces (APIs) available for developers to communicate with the service. APIs are often the most inclusive method of acquiring online content and can be used to search, aggregate, and extract content outside of the standard GUI. A service API call can capture what is being seen on a webpage in plain view as well as critical metadata that are not available through web pages and screen captures. Some APIs require paid licenses and free versions may be limited in their capabilities.

As this methodology can frequently discover the most amount of evidence needed, as well as provide a reasonable amount of guarantee on reliability of evidence, this methodology is preferred over the "Web Browser Capture" or "Screenshot" methods that are also recommended in this document.

Native Operating System Utilities

Most operating systems include native utilities such as curl, wget, or Invoke-WebRequest, for downloading web hosted content. These utilities can be configured with options for presenting specific browser agent strings and other settings to emulate desired user profiles and can be used to download entire websites or specific content on a site. The tools can be scripted against repeated collections or collections of multiple sites.

Disclaimer: Due to the fast-development of utilities (i.e., hardware and software tools), please refer to your agency approved open-source and commercially available tool list.

8.1.2 Web Browser / Plug-ins / Extensions

Web browsers used to access websites also contain native HTML archive acquisition capabilities (i.e. Save As) which can be used to combine a webpage into a single file. This file format is represented in the RFC standard 2557. (reference: https://www.rfc-editor.org/info/rfc2557). Web browser acquisitions are typically best suited for static websites. Dynamic content captured via this method may be limited to the current user's view at that time (e.g., video served off-site via JavaScript). Due to the disparity in technical implementations and diversity of web content, not all extensions will be reliable. For example, plug-ins like BatchLink or Vimeo Downloader will not work for videos embedded in other ways, or in ways that frustrate the method employed by the plug-in.

Directory Traversal

Directories on the target web server can be discovered by iterating through common directory and file names. This can be automated or done manually by making web requests continually through such a list. This can be useful information when discovering available data that is not referenced on known pages, or for matching web servers to web sites following additional discovery.

Fusking

The process of inferring the name of images or known website artifacts (e.g., site maps, media) by trial and error using common naming conventions is often referred as "brute-force guessing" and colloquially known as "fusking." For instance, camera manufacturers establish naming conventions to prevent naming collisions which can be used to identify online content. Since they are consistent and contiguous, other images may exist with adjacent names. For example, an image found online with the name DSC_0004.jpg not only implies the image was captured with a Kodak digital camera (of a make post establishing of that convention), it also implies the existence of an image named DSC_0001.jpg, DSC_0002.jpg, DSC_0003.jpg and potentially DSC_0005.jpg and beyond. Making requests for those images may allow for their collection.

8.1.3 Screenshots

A screenshot is a digital image that shows the contents of a computer display, preserved by electronic means, kept in the regular course or conduct of an official activity, all of which are shown by the testimony of the custodian or other qualified witnesses, and is excepted from the rule on hearsay evidence. Compared to "Utilities" and "Web Browser Capture" methodologies, this is the least forensically sound manner but may be the only way to capture all content displayed to the user (e.g., live streaming) and to understand the context of the content. This method should always be used in conjunction with the methods above.

9. Preservation

The examiner should preserve all online content acquired from the target website, in addition to the documentary evidence generated by the examiner during that process, to a forensically sound image (e.g., .Lx01, .ad1) or other archive (e.g., zip, gzip) using industry standard procedures. The forensic image should include documentation of examiner name, acquisition date and time, and evidence descriptions.

For additional information regarding evidence archiving, refer to *SWGDE Best Practices for Archiving Digital and Multimedia Evidence, v1.0*.

For discovery purposes and in cases where the target website is still active, a working copy of the online content acquired should be preserved by disabling live links to avoid inadvertent access.

10. History

Revision	Issue Date	History
1.0 DRAFT	6/17/2021	Initial draft created and voted by SWGDE for release as a Draft for Public Comment.
1.0	1/13/2022	Published
1.1 DRAFT	10/13/2023	Made changes and released for public comment.
1.1	3/7/2024	SWGDE voted to approve as Final Approved Document. Formatted for release as a Final Approved Document.