Ing. Mag. Horst Greifeneder
Allgemein beeideter und gerichtlich zertifizierter Sachverständiger, Fachgebiet Informationstechnik, LG Wels
Computer-Forensik-Spezialist
Externer Datenschutzbeauftragter, CIPP/E, CIPM, CIPT

Forensische Spurensuche im Internetarchiv

„Mit neuen Perspektiven kommen neue Einsichten, was bedeutet, dass es nur eine Frage der Zeit ist, bis wir finden, was wir suchen."

Dr. Temperance „Bones" Brennan

1. Einleitung

Angesichts der enormen Bedeutung des Internets für das private und geschäftliche Leben, wundert es nicht, dass Onlinemedien und deren Inhalte immer häufiger Gegenstand von Gerichtsverfahren sind. Durch die kurze Lebensdauer von Websites kann es der Fall sein, dass verfahrensrelevante Daten zum Zeitpunkt der Befundaufnahme online nicht mehr verfügbar sind.

Eine Möglichkeit, gelöschte Websites oder nicht mehr abrufbare Internetinhalte wiederherzustellen, bietet der Onlinedienst Wayback Machine der Non-Profit-Organisation Internet Archive. Der Dienst wird unter https://web.archive.org kostenlos zur Verfügung gestellt.[1]

Auch wenn die Wayback Machine nicht primär für die rechtliche Verwendung entwickelt wurde, werden durch den Dienst gewonnene Beweismittel immer öfter in Gerichtsverfahren vorgelegt. In Österreich hielt bereits 2020 der OGH in einer Entscheidung fest, dass im Zuge eines Verfahrens der archivierte Inhalt von rekonstruierten Websites durch die Vorlage von Wayback-Screenshots bescheinigt werden kann.[2] Das Europäische Patentamt dokumentiert auf seiner Website eine Reihe von weiteren Entscheidungen, in denen Beweismittel aus Internetarchiven eine zentrale Rolle spielten.[3]

Der nachfolgende Beitrag beschäftigt sich zum einen mit der Funktionsweise des Internetarchivs und zum anderen mit Aspekten zur Vollständigkeit und Echtheit der verfügbaren Informationen.

2. Funktionen der Wayback Machine

2.1. Allgemeines

Die Wayback Machine ist ein seit 1996 verfügbarer Onlinedienst, welcher nach archivierten Webdateien durchsucht werden kann. Zur Erstellung des Archivs werden Webadressen (URLs)[4] über sogenannte Webcrawler[5] immer wieder abgerufen. Die gefundenen Inhalte werden auf eigenen Servern des Anbieters gespeichert und ermöglichen die gänzliche oder teilweise Wiederherstellung von Websites zu einem ausgewählten Datum. Der Gesamtumfang des Archivs betrug im Oktober 2022 mehr als 751 Mrd Websits.[6]

2.2. Suche nach archivierten Seiten

Die Kalenderansicht der Wayback Machine ist die zentrale Schnittstelle für die Suche und Anzeige von archivierten URLs (zB eine Webadresse). Falls Aufzeichnungen für eine URL verfügbar sind, werden in der Kalenderansicht die jeweiligen Treffer und verfügbaren Snapshots[7] angezeigt.

Anhand der Website des Hauptverbands sollen die Funktionen der Wayback Machine veranschaulicht werden. Nach Eingabe der URL https://www.gerichts-sv.at in die Suchmaske der Wayback Machine erscheint das in Abbildung 1 wiedergegebene Ergebnis in der Kalenderansicht.

Die Kalenderansicht zeigt in einer Zeitleiste an, wann und wie häufig eine einzelne URL über die Jahre gecrawlt wurde. Die Ansicht liefert keine Angaben darüber, wie oft eine Seite tatsächlich aktualisiert worden ist.

Die Frequenz, mit der einzelne Websites gecrawlt werden, ist unterschiedlich. Manche werden täglich, andere in unregelmäßigeren Abständen indiziert. Im Falle der Startseite des Hauptverbands wurde diese vom 7. 12. 2004 bis zum 5. 10. 2022 insgesamt 134 Mal erfasst, das heißt, im Schnitt wurde die Startseite etwa alle 49 Tage abgerufen.

In der Ergebnisansicht zeigen farbige Kreise jene Tage an, für die ein oder mehrere Snapshots der Suchadresse verfügbar sind. Blaue Punkte bestätigen einen erfolgreichen Seitenaufruf (2xx-Code). Grüne Kreise zeigen einen Redirect (3xx-Code) beim Abruf der ursprünglichen Seite an. Orange Kreise verweisen auf einen Client Error (4xx-Code). Rote Kreise belegen einen Server Error (5xx-Code). In der Gesamtschau liefert die Kalenderansicht ein erstes Indiz für die Verfügbarkeit von Websites.

Beim Überfahren eines Kreises mit der Maus erscheint ein Pop-up-Fenster, welches auf einzelne Snapshots nach Datum und Uhrzeit verlinkt. Durch Anklicken eines Links wird eine archivierte Webdatei (zB eine Website) aufgerufen und anhand der gespeicherten Daten dargestellt.

Zu Demonstrationszwecken soll im nächsten Schritt die älteste Version der Startseite des Hauptverbands vom 7. 12. 2004 aufgerufen und angezeigt werden (siehe Abbildung 2).

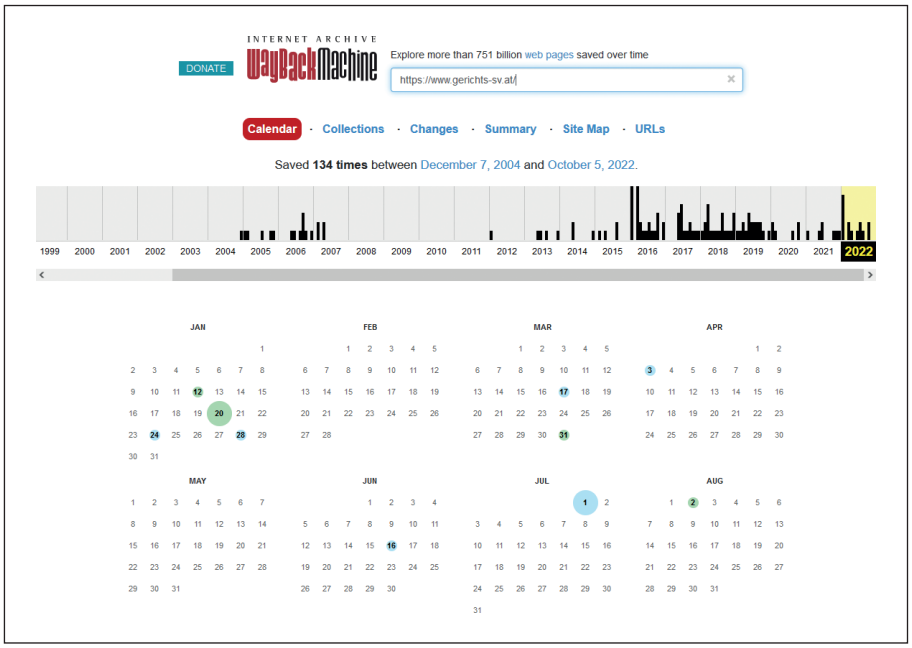

Abbildung 1: Ergebnisansicht der Wayback Machine (URL: https://www.gerichts-sv.at)

Abbildung 2: Startseite der Website des Hauptverbands (Snapshot vom 7. 12. 2004)

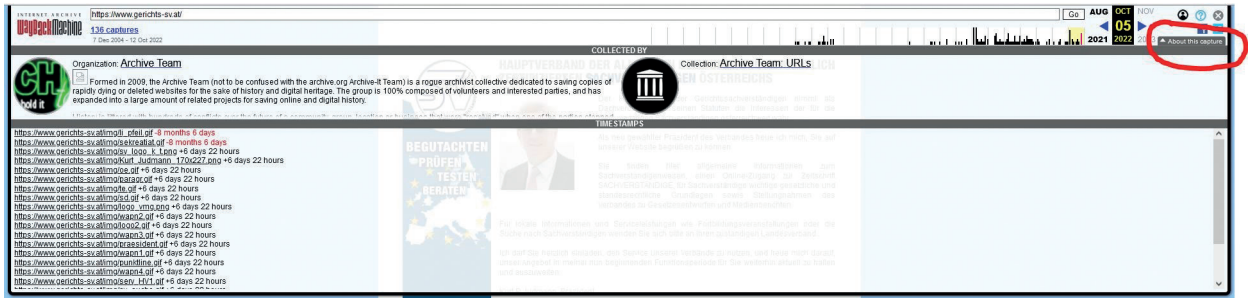

Abbildung 3: „About this Capture“ liefert Informationen zu den Speicherdaten der Elemente der Website

Der Zeitpunkt der Archivierung der angezeigten Website ist am oberen rechten Rand des Fensters ersichtlich. Anzumerken ist, dass der Zeitstempel (UTC) nur für die jeweils archivierte URL verbindlich ist.

An diesem Punkt stellt sich die Frage: Wird die gefundene Website historisch korrekt und vollständig angezeigt?

Im Falle einer Website, die aus einer Quelldatei und eingebetteten Dateien (zB Bildern, CSS, Videos) besteht, wurden einzelne Elemente meist zu unterschiedlichen Zeitpunkten archiviert. Tatsächlich werden angezeigte Website jeweils on-the-fly mit den verfügbaren Dateien, welche zum nächstgelegenen Datum archiviert worden sind, bestmöglich zusammengestellt.[8]

Für den Fall, dass ein Webelement überhaupt noch nicht archiviert worden ist, wird direkt auf der aktuellen Website nach dem Element gesucht und dieses – falls vorhanden – verwendet.

Forscher haben rausgefunden, dass etwa 38,7 % der zusammengesetzten Websites zeitlich kohärent sind. Höchstens 17,9 % (ungefähr 1 von 5) sind sowohl zeitlich kohärent als auch zu 100 % vollständig.[9]

Die Feststellung mag auf den ersten Blick beunruhigend erscheinen. In der Praxis hat sie sich aber als handhabbar erwiesen. Meist geht es bei Fragestellungen ja darum, ob bzw welche fallrelevanten Informationen zu einem bestimmten Zeitpunkt im Internet veröffentlicht waren oder nicht.

Genauere Informationen zum Zeitpunkt der Archivierung einzelner Elemente können über die Funktion „About this Capture“ gewonnen werden (siehe Abbildung 3).

Die Funktion listet die Zeitstempel aller Seitenelemente im Vergleich zu Datum und Uhrzeit der Basis-URL einer Seite auf.[10] Bei der Untersuchung (Abbildung 5) wird deutlich, dass einzelne Elemente der Website bereits Monate vor dem Speicherdatum der angezeigten URL archiviert worden sind. Andere Elemente wurden hingegen erst Tage später erfasst.

Es kann aber auch passieren, dass einzelne Websites gänzlich fehlen und archivierte Seiten nicht vollständig sind. Archivierte Webinhalte sind besonders dann nicht optimal aufbereitet, wenn die ursprüngliche Website dynamische Website-Inhalte beinhaltet.

2.3. Herkunft der Archivdaten

Durch Aufruf der Collections-Ansicht erhält man einen Einblick in die Herkunft der einzelnen Snapshots. Angezeigt werden Metadaten (wie Name und Ursprung des Crawls, Anzahl der enthaltenen URLs und Zeitpunkt des ursprünglichen Crawl-Vorgangs).

2.4. Vergleich von Webdokumenten

Die Ansicht „Changes“ beinhaltet ein interessantes Tool, mit dem Änderungen im Inhalt von archivierten Websites erkannt und angezeigt werden können.

Im Beispiel der Abbildung 4 wurde die archivierte Startseite vom 5. 10. 2022 mit der Version vom 3. 4. 2022 verglichen. Die Änderungen werden farblich hervorgehoben, wobei mit gelber Farbe unterlegte Elemente auf neue und blaue Hervorhebungen auf gelöschte Inhalte verweisen.

Der Autor musste leider feststellen, dass es bei einzelnen Seitenvergleichen immer wieder zu funktionalen Problemen des Dienstes gekommen ist.

Abbildung 4: Websites im Vergleich

2.5. Übersicht der erfassten Webdaten

Die Ansicht „Summary“ enthält Statistiken über die gecrawlten MIME-Types[11] einer Website (wie Text, Grafik, Audio, Video oder Applikation) in einem festgelegten Zeitraum.

In der Statistik der Abbildung 5 werden die Anzahl der Erfassungsvorgänge (Captures), URLs und neuen URLs für einzelne MIME-Typen der Website angezeigt. Die Grafik zeigt eine umfangreiche Website mit über 100.000 erfassten Elementen pro Jahr. Interessant ist auch die vergleichsweise große Anzahl neu archivierter URLs. Ein Hinweis auf eine Website mit häufig aktualisierten Inhalten. Für die letzten 10 Captures sind zusätzlich der Zeitpunkt, Statuscode, MIME-Type und die Dateigröße in Bytes verfügbar.

2.6. Dokumentation der Websitestruktur

Visuell interessant ist die „Site Map“-Ansicht, welche die erfassten Seiten einer Domain nach Jahren gruppiert und dann für jedes Jahr eine visuelle Sitemap in Form eines radialen Baumdiagramms erstellt. Ausgehend von der Starseite wird in Ringsegmenten die strukturelle Tiefe einer Website veranschaulicht. Den Mittelkreis bildet die Websitedomain. Aufeinanderfolgende Ringe, die sich von der Mitte aus bewegen, präsentieren einzelne Verzeichnisse und Seiten der Website.

2.7. Gesamtmenge der erfassten Daten

Die Ansicht „URLs“ liefert Informationen zur Gesamtanzahl der erfassten URLs einer Website und ermöglicht die Suche nach spezifischen Webinhalten anhand von Suchbegriffen und MIME-Types.

Angezeigt werden unter anderem die Anzahl der Captures, Duplikate und Unikate der angezeigten Webdateien. Mit dieser Funktion kann beispielsweise erhoben werden, ob eine Datei mit gleicher Internetadresse, aber unterschiedlichem Inhalt archiviert wurde.

3. Aspekte der Wayback-Nutzung

3.1. Vorbemerkung

Nachfolgend werden verschiedene Aspekte des Einsatzes der Wayback Machine bei der Befundaufnahme und Gutachtenserstellung erörtert.

3.2. Vollständigkeit der Websites

Die Wayback Machine sammelt nur Inhalte von öffentlich zugänglichen Websites. Es werden keine passwortgeschützten oder auf Formulareingaben basierende Websites sowie Seiten von sicheren Servern archiviert. Zudem kann es sein, dass Seiten aufgrund von Robots-Ausschlüssen[12] nicht archiviert oder Websites auf Antrag des Website-Eigentümers ausgeschlossen worden sind.

Enthält die archivierte Website dynamische Seiten[13] mit Formularen, JavaScript und anderen Elementen, die eine Interaktion mit dem ursprünglichen Host erfordern, kann es sein, dass die angezeigte Archivseite die Inhalte oder Funktionen der ursprünglichen Websit nicht richtig wiedergibt.

Manchmal führt dies dazu, dass Websites präsentiert werden, die als solche in der Vergangenheit gar nie existiert haben. In derartigen Fällen ist der Sachverständige gefordert, nachweislich festzustellen, welche Folgen eine ahistorische Rekonstruktion einer Website für Befund und Gutachten hat.

3.3. Echtheit der archivierten Daten

Um die Echtheit der archivierten Daten zu bestätigen, kann der Plattforminhaber auf Antrag eine eidesstattliche Erklärung für ausgedruckte Websites erstellen.[14] Dabei wird bestätigt, dass das ausgedruckte Dokument eine echte und korrekte Kopie der archivierten Daten darstellt.

Wie bereits dargestellt, bedeutet das aber nicht, dass es sich beim ausgedruckten Dokument um eine 100%ige Kopie der tatsächlichen Website zum angezeigten Zeitpunkt handeln muss.

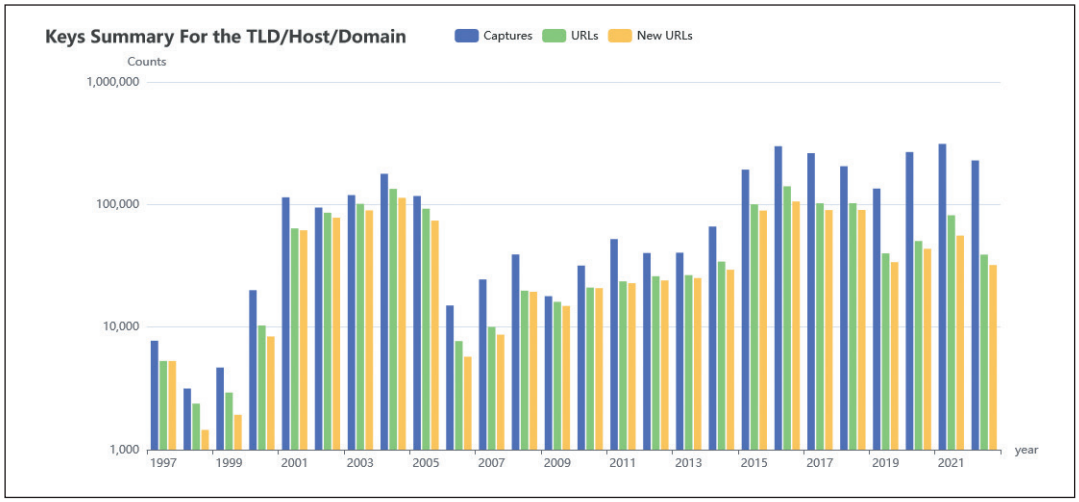

Abbildung 5: Übersicht der Captures einer Domain

3.4. JavaScript

JavaScript-Elemente sind oft schwer zu archivieren, insbesondere dann, wenn sie Links generieren, ohne dass der vollständige Name auf der Seite steht. Wenn JavaScript den ursprünglichen Server kontaktieren muss, um zu funktionieren, schlägt es außerdem fehl, wenn es archiviert wird.[15]

Falls JavaScript bei der Auswertung der archivierten Inhalte im Browser deaktiviert ist, stammen Bilder und Links von der aktuellen Website und nicht aus dem Archiv der Webdateien.

3.5. Einzelne Websites selbst sichern

Die Wayback Machine bietet unter https://web.archive.org/save die Möglichkeit, einzelne Websites für Beweismittelzwecke selbst zu sichern.

Die zu sichernde Seite wird in das Formularfeld eingetragen und gespeichert. Nach erfolgreicher Speicherung steht sofort eine permanente URL als vertrauenswürdige Referenz für die Seite zur Verfügung. Zu beachten ist dabei, dass diese Methode nur eine einzelne Seite speichert, nicht die gesamte Website.

Benutzer mit einem kostenlosen Wayback-Konto können darüber hinaus die archivierte Seite samt Screenshot in einem eigenen Webarchive abspeichern und externe Links (Outlinks) samt Inhalten durch die Wayback Machine archivieren lassen.[16] Das dazugehörige Protokoll wird einem dann noch per E-Mail übermittelt.

3.6. Wiederherstellung von Websites

Es gibt mehrere Dienste von Drittanbietern (zB Wayback Rebuilder[17] oder Wayback Machine Downloader),[18] die anbieten, eine Website aus den gespeicherten Wayback-Archiven neu zu erstellen. Über die Funktionalität und Vollständigkeit der wiederhergestellten Websites kann in diesem Beitrag wegen fehlender praktischer Erfahrungen keine Aussagen getroffen werden.

4. Zusammenfassung

Der Internetarchivierungsdienst Wayback Machine hat sich als ein wertvolles Werkzeug zur Rekonstruktion von gelöschten oder nicht mehr verfügbaren Webinhalten erwiesen.

Natürlich darf man nicht erwarten, dass im Archivierungsdienst die Inhalte einer Website vollständig gespeichert sind. Die Stärke des Archivs liegt in der Möglichkeit, eine Zeitreise in die Geschichte einer Website unternehmen und im Idealfall gutachtensrelevante Informationen zutage fördern zu können.

Der Umstand, dass die Inhalte archivierter Websites zeitlich nicht immer kohärent sind, hat in der Praxis meist geringere Auswirkungen als im ersten Moment befürchtet. In vielen Fällen ist primär der Text auf der Website von Interesse für die Befundaufnahme. Und dabei spielt die zeitliche Kohärenz anderer Elemente der Website keine so entscheidende Rolle.

In Untersuchungen, bei denen Unterschiede zwischen den Screenshots der Archivseite und den zeitlichen Eigenschaften eingebetteter Seitenelemente von Relevanz sind, sollte eine differenzierte Analyse der Ursachen und Folgen der Diskrepanzen die Aussagekraft der Feststellungen in der Befundaufnahme erhöhen.

Von zentraler Bedeutung für die Beweiskraft vorgelegter Wayback-Daten ist zudem eine schlüssige Darstellung der technischen Funktionsweise des Internetarchivs samt nachvollziehbaren Angaben zur Vollständigkeit und Kohärenz der gespeicherten Internetinhalte.

Anmerkungen:

1 Das Internet Archive in San Francisco ist ein gemeinnütziges Projekt, das 1996 von *Brewster Kahle* gegründet wurde und seit 2007 den offiziellen Status einer Bibliothek hat; siehe https://de.wikipedia.org/wiki/Internet_Archive.

2 OGH 23. 10. 2020. 15 Os 42/20w ua.

3 Seite https://www.epo.org/law-practice/legal-texts/html/caselaw/2019/d/clr_iii_g_4_2_3.htm.

4 Mit der URL wird eine Adresse bezeichnet, die eine Ressource (zB eine Webdatei) auf einem Server angibt.

5 Ein Webcrawler ist ein Computerprogramm, das automatisch das World Wide Web durchsucht und die Inhalte von Websites auswertet; siehe https://de.wikipedia.org/wiki/Webcrawler.

6 Siehe https://web.archive.org.

7 Ein Snapshot ist eine Darstellung der archivierten Inhalte einer Webdatei. Dabei kann es sich um ein einzelnes Bild bzw Textdokument oder um eine ganze Website handeln.

8 Siehe https://archive.org/legal/affidavit.php.

9 Siehe https://www.slideshare.net/ScottAinsworth/only-one-out-of-five-archived-web-pages-existed-as-presented.

10 Siehe https://blog.archive.org/2017/10/05/wayback-machine-playback-now-with-timestamps.

11 MIME-Type oder Content-Type klassifiziert die im Internet übermittelten Daten.

12 Mit einer Robots.txt-Datei können Admins steuern, auf welche Dateien Crawler auf der Website zugreifen können. Die Textdatei befindet sich im Stammverzeichnis der Website und kann – falls vorhanden – direkt abgerufen werden.

13 Als dynamische Website werden Seitendesigns bezeichnet, bei denen nicht alle eingeblendeten Elemente (wie Texte, Bilder oder Videos) fest in die Seite integriert sind. Stattdessen werden Website-Inhalte – zB abhängig von Benutzeraktionen, meist mittels JavaScript – dynamisch in die Struktur der Website eingebunden.

14 Siehe: https://archive.org/legal/faq.php#aff.

15 Siehe https://help.archive.org/help/using-the-wayback-machine.

16 Siehe https://blog.archive.org/2019/10/23/the-wayback-machines-save-page-now-is-new-and-improved.

17 Siehe https://waybackrebuilder.com.

18 Siehe https://waybackmachinedownloader.com.

Korrespondenz:
Ing. Mag. Horst Greifeneder, CIPP/E, CIPM, CIPT
FDS | Forensik Data Services
Tel.: 07242 / 777 15
E-Mail: office@fds.at