Die WayBackMachine von Archive.org

Stand: 14.10.2017

Inhaltsverzeichnis

Die WayBackMachine direkt nutzen 1

Statistiken einer Webseite anzeigen lassen 2

Captures einer Webseite anzeigen lassen 2

Die WayBackMachine auf Google-Cache.de nutzen 4

Backup/Capture erstellen 4

Gespeicherte Webseiten anzeigen 4

Deine Webseite von der WayBackMachine ausschließen 4

Die robots.txt erweitern 5

Die WayBackMachine direkt nutzen

Der erste Schritt ist das Aufrufen der Webseite https://web.archive.org/ und dort auf der Startseite die gewünschte URL in die Suchleiste einzugeben.

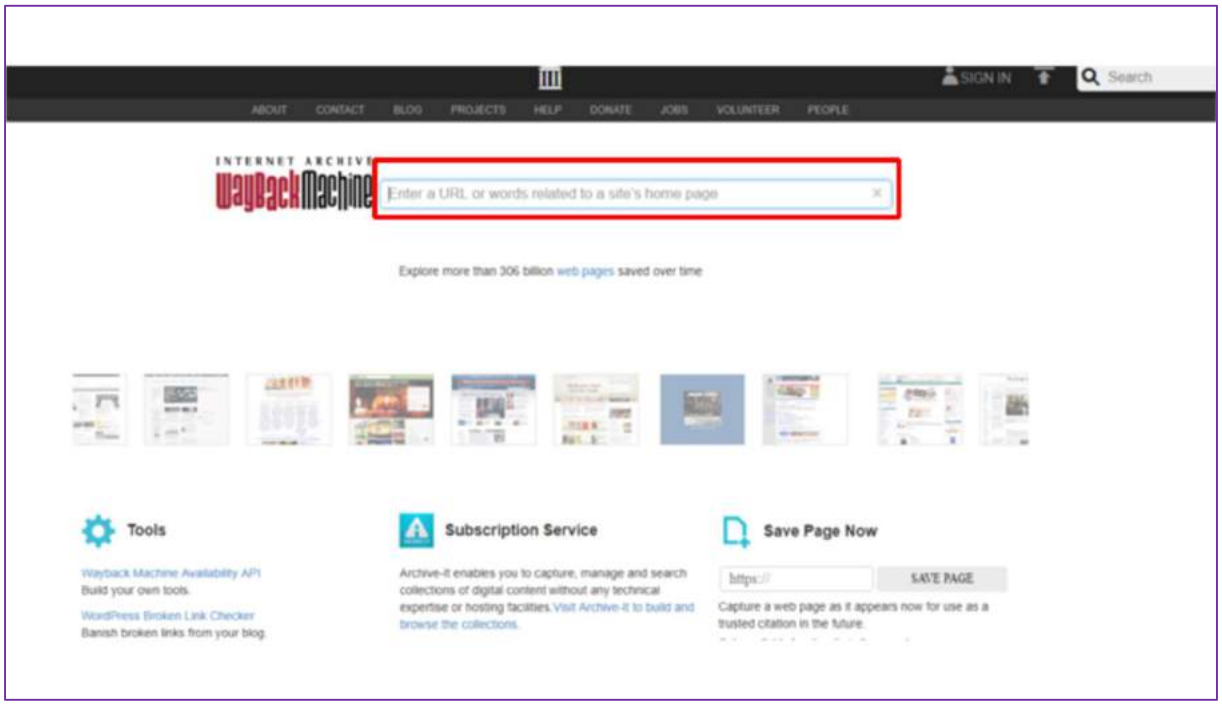

Abbildung 1 – Startseite der WayBackMachine

Die entsprechenden Daten werden automatisch angezeigt. Sollte dies nicht der Fall sein, kann auch die Enter-Taste betätigt werden. Sollte die Webseite existieren, wird die Kalenderübersicht des aktuellen Jahres angezeigt.

Abbildung 2 - Kalenderübersicht einer gesuchten Webseite

Statistiken einer Webseite anzeigen lassen

Auf der Kalenderübersicht der WayBackMachine ist es möglich sich einige Statistiken anzeigen zu lassen. Hierfür muss einfach auf den Link „Summary of Domainname“ geklickt werden.

So lassen sich zum Beispiel zu der gesuchten Webseite die insgesamt gecrawlten HTML- und CSS-Dateien anzeigen. Auch eine Filterung der Statistiken auf unterschiedliche Jahre ist möglich.

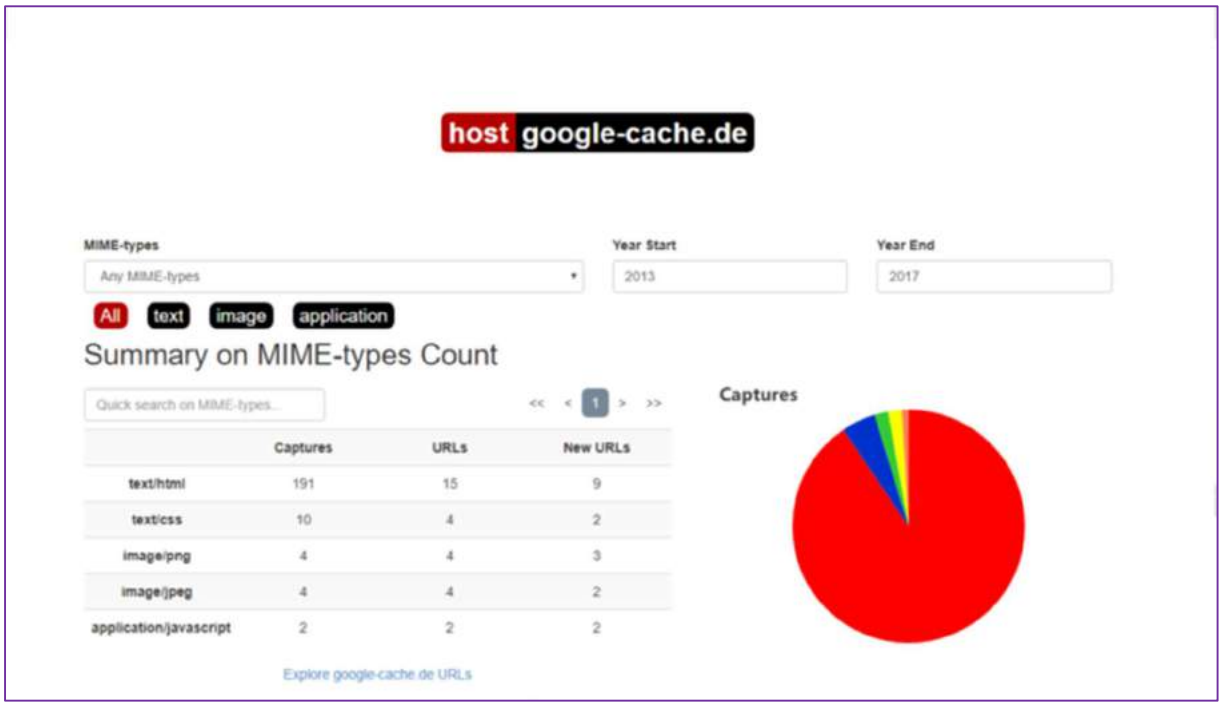

Abbildung 3 - Statistiken der WayBackMachine

Captures einer Webseite anzeigen lassen

Auf der Hauptseite der WayBackMachine von Archive.org lassen sich die verschiedenen Captures in einem Kalender anzeigen. Es ist immer das aktuellste Jahr ausgewählt. Mit einem Klick auf das gewünschte Jahr wechselt die Kalenderübersicht zu dem gewünschten Jahr.

In der Kalenderübersicht werden die Captures durch einen blauen oder einen grünen Kreis gekennzeichnet. Je mehr Captures für den jeweiligen Tag vorhanden sind, desto größer ist der entsprechende Kreis. Mit dem Halten der Maus über den gewünschten Kreis/Tage, werden die Captures für den jeweiligen Tag angezeigt. Mit einem Klick auf den Kreis wird das erste Capture des Tages aufgerufen.

Die Unterscheidung in den verschiedenen Farben tritt durch einen Redirect auf. Zum Beispiel bei einem 301-Redirect von http auf https oder von der WWW-Version auf die Version ohne WWW (Bsp. http://www.google-cache.de auf https://www.google-cache.de oder https://google-cache.de auf https://www.google-cache.de). Wird die

zur Archivierung eingegebene URL einer Domain mit einem solchen Redirect weitergeleitet, wird der Kreis grün gekennzeichnet.

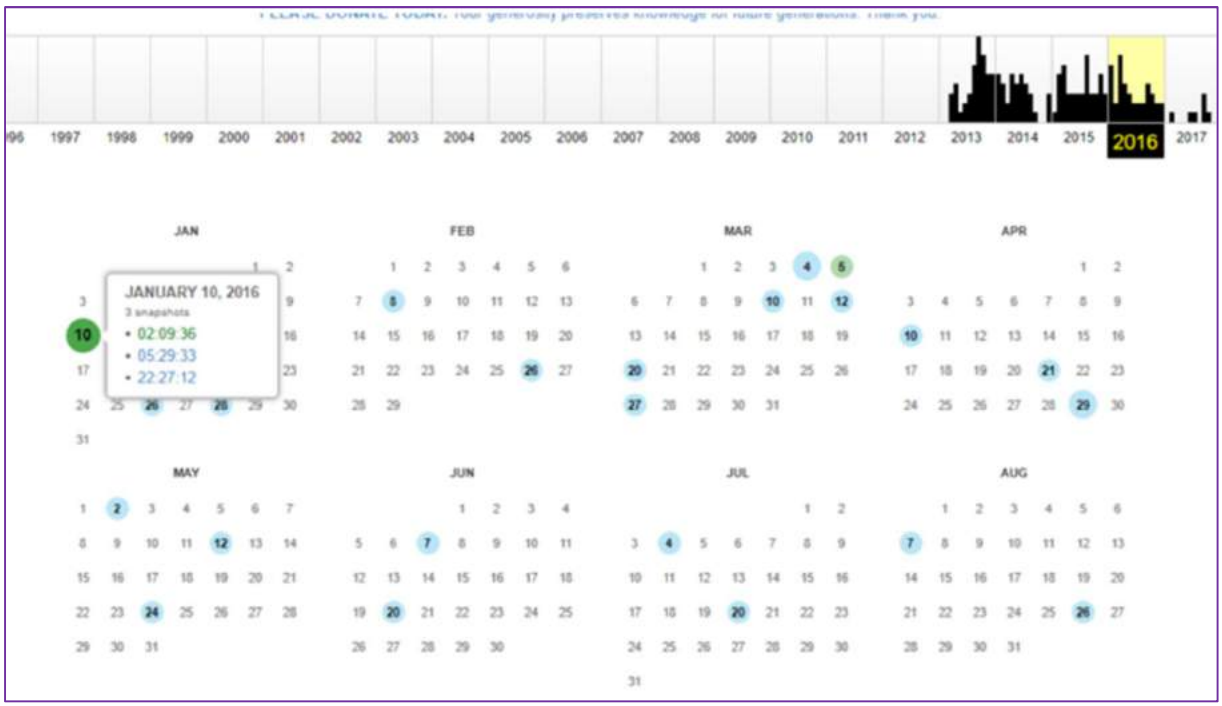

Abbildung 4 - Kalendereinträge in der Kalenderübersicht

Ist ein Capture erst einmal aufgerufen, werden die gespeicherten Daten angezeigt. Des Weiteren wird eine Navigationsleiste angezeigt, über die die verschiedenen Tage und die verschiedenen Captures ausgewählt werden können. Auch ein Sprung über mehrere Jahre ist möglich.

Abbildung 5 - Capture-Ansicht mit Kalendernavigation

Die WayBackMachine auf Google-Cache.de nutzen

Backup/Capture erstellen

Zum Erstellen eines Captures einfach die entsprechende URL in das Suchfeld eingeben und im Dropdown-Menü den Punkt „Backup“ auswählen. Danach auf „LOS!“ klicken und schon wird die gewünschte Seite durch die WayBackMachine erfasst.

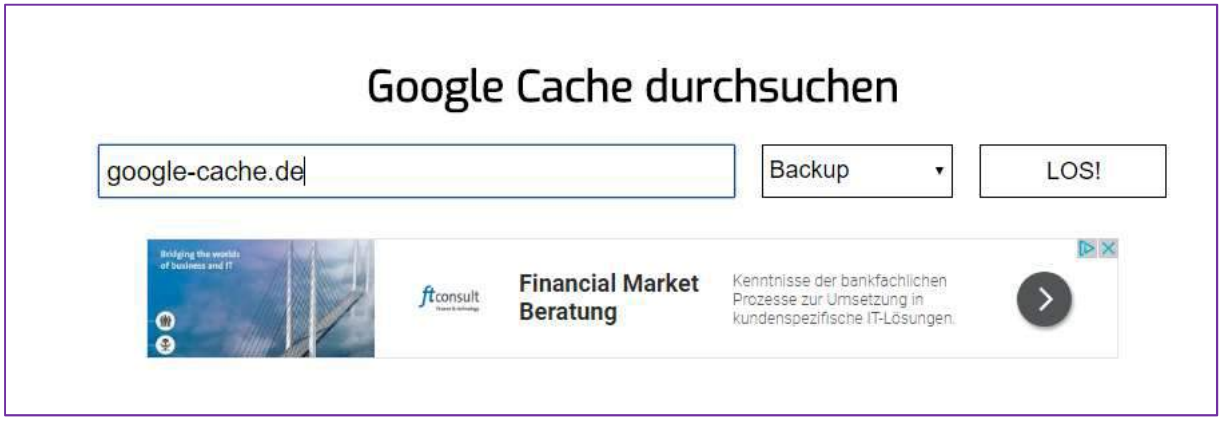

Abbildung 6 - Capture erstellen lassen

Das Capture ist sofort in der Kalenderübersicht zu sehen und auswählbar.

Gespeicherte Webseiten anzeigen

Die gespeicherten Webseiten einer Domain lassen sich ebenso anzeigen. Hierfür im Suchfeld die gewünschte URL eintragen und im Dropdown-Menü den Punkt „Archive.org“ auswählen. Danach auf „LOS!“ klicken und schon wird die Kalenderübersicht der gespeicherten Seite angezeigt, sofern Captures zu der gewünschten Seite vorhanden sind.

Deine Webseite von der WayBackMachine ausschließen

Früher war es möglich, den Crawler diverser Webarchive immer einfach über die robots.txt der entsprechenden Webseite auszusperren. Seit Mitte April 2017 ignoriert die WayBackMachine jedoch diese Sperre immer stärker.

Quelle: https://www.heise.de/newsticker/meldung/Archivierung-des-Internets-Internet-Archive-ignoriert-kuenftig-robots-txt-3693558.html

In einem Selbstversuch habe ich festgestellt, dass es jedoch möglich ist, den Crawler der WayBackMachine auszusperren. Der Crawler benötigt jedoch ein wenig Zeit, bevor die Aktualisierung der robots.txt bemerkt wird.

Abbildung 7 - Durch robots.txt blockierte Webseite

Die robots.txt erweitern

Damit der Web-Crawler der WayBackMachine eine Webseite nicht mehr archiviert, muss in der robots.txt der entsprechenden Webseite einfach folgendes vermerkt werden:

User-agent: ia_archiver
Disallow: /