Automatisierte Überprüfung von Webauftritten auf Fremdinhalte

Konstantin Knorr[1], David Müller[2]

Abstract: Die öffentlichen Webauftritte von Unternehmen und Behörden sind heute oft sehr umfangreich und entwickeln sich dynamisch weiter. In der dazugehörigen Datenschutzerklärung sollte der potentielle Besucher laut Artikel 13 der DS-GVO u.a. über die Einbindung externer Inhalte wie Fonts, Videos oder Werbung sowie über die dahinterstehenden Parteien informiert werden. Meist geschieht dies in Form einer einzigen Datenschutzerklärung für den gesamten Webauftritt. Insbesondere bei rasch wachsenden und sich dynamisch verändernden Webaufritten ist es schwierig, die Datenschutzerklärung aktuell zu halten. Es kommt leicht zu einer Über- oder Unterdeckung bei der Angabe der externen Parteien. Schlimmstenfalls drohen Bußgelder oder Abmahnungen bspw. durch Konkurrenten. Das Papier beschreibt ein Tool, das automatisiert in einer Tiefensuche die externen Verbindungen eines Webauftritts identifiziert und mit den manuell aus der Datenschutzerklärung ausgelesenen Parteien abgleicht. Nicht gelistete Fremdparteien verletzen die Informationspflicht und entbehren einer Rechtsgrundlage. Der ca. 32.000 Seiten umfassende Webauftritt einer Hochschule wurde mit dem Tool getestet. Die Ergebnisse des Tests werden beschrieben und diskutiert. Ferner werden Empfehlungen für das Vorgehen zur Beseitigung von Abweichungen gegeben. Die Autoren hoffen, damit den für die Datenschutzerklärung Verantwortlichen und Datenschutzbeauftragten eine Hilfestellung für die Erstellung und Pflege der Datenschutzerklärung ihres Webauftritts bieten zu können.

Keywords: Browser Automation, Datenschutzerklärung, HTTP, Python, Scrapy, Selenium, Web

1 Einleitung

Durch Aufruf einer Webseite in einem Browser werden personenbezogene Daten, wie zum Beispiel die IP-Adresse des verwendeten Rechners, an einen Server geschickt. Ist in dieser Webseite z.B. ein YouTube-Video eingebettet, so werden diese Daten außerdem an einen YouTube-Server gesendet. Verwendet die Webseite zusätzlich Google Fonts, werden die Daten auch noch an einen Google-Server weitergeleitet, vgl. Abbildung 1. Durch die Verwendung des World Wide Webs werden eine Vielzahl von Daten gesammelt und verarbeitet. Diese Daten können persistent gespeichert und unter anderem zum Web-Tracking [MM12] und zur Profilbildung genutzt werden. Die meisten Anwender sind sich dieser Verarbeitung ihrer Daten nicht bewusst. Individualisierte Werbung, die sich den Interessen des Nutzers anpasst, ist meist die einzige ersichtliche Auswirkung des Web-Trackings. Wie personenbezogene Daten genutzt werden dürfen, muss dem Nutzer zwar

[1] Hochschule Trier, Fachbereich Informatik, Am Schneidershof, D-54208 Trier, knorr@hochschule-trier.de
[2] Hochschule Trier, Fachbereich Informatik, Am Schneidershof, D-54208 Trier, muelled@hochschule-trier.de

mitgeteilt werden, es schützt aber letztlich nicht vor Missbrauch.

Mayer und Mitchell geben einen Überblick der zum Web-Tracking eingesetzten Technologien [MM12]. Eine umfassende Studie mit mehr als einer Million untersuchten Seiten haben Englehardt und Narayanan 2016 präsentiert [EN16]. Ermakova et al. [EBFK18] analysieren in einer Literaturrecherche den aktuellen Forschungsstand bzgl. Web-Tracking. Wambach, Knorr und Schulte haben 2015 und 2016 die Start-Seiten von Hochschulen und Krankenhäusern auf externe Inhalte untersucht [WK15, WSK16]. Eine Diskussion der Ergebnisse im Vergleich zu dieser Studie findet sich in Abschnitt 5.

Der Beitrag beschreibt das Tool *wirechuck*, das automatisiert die eingebundenen Drittparteien eines kompletten Web-Auftritts durch eine Tiefensuche ermittelt und gegen die in der Datenschutzerklärung (DSE) aufgelisteten Fremdparteien abgleicht. Die DSE eines Webauftritts kann somit geprüft und ggf. angepasst werden, was insbesondere für sich dynamisch weiterentwickelnde Webauftritte regelmäßig getan werden sollte. Eine DSE sollte (1) über die Fremdparteien informieren und (2) die Rechtsgrundlage dazu liefern. Bei einer nicht angegebenen Fremdpartei sind beide Punkte verletzt. Die Prüfung der Rechtsgrundlage verlangt juristische Expertise und erfolgt daher in der Regel manuell. Neben dem hier beschriebenen quell-offenen Ansatz zur Analyse von Webseiten und deren DSE gibt es kommerzielle Ansätze wie z.B. https://www.alfright.eu/, deren Funktionsweise allerdings nicht publiziert ist.

Der Rest des Artikels hat den folgenden Aufbau: Abschnitt 2 beschreibt die Grundlagen zu HTTP, URI und zum Datenschutz, insbesondere den Inhalt einer DSE von Webseiten. Das verwendete Tool *wirechuck* [wich] wird in Abschnitt 3 beschrieben. Die Studie des Webauftritts der Hochschule Trier folgt in Abschnitt 4. Abschließend diskutiert Abschnitt 5 die Ergebnisse, setzt sie in Bezug zu verwandten Arbeiten, spricht Empfehlungen aus und endet mit einem Ausblick bzgl. einer Weiterentwicklung des Tools.

2 Grundlagen

2.1 Technische Grundlagen: HTTP und URI

Das Hypertext Transfer Protocol (HTTP) ist ein vom World Wide Web Consortium standardisiertes Protokoll zur Übertragung von Daten auf der Anwendungsschicht. Das Protokoll wird vor allem zum Laden von Webseiten in einem Browser verwendet. HTTP setzt auf dem Transmission Control Protocol (TCP), neuerdings auch auf User Datagram Protocol (UDP) und dem Internet Protocol (IP) auf. Dies bedeutet, dass zwischen einem Client und einem Server eine Verbindung aufgebaut wird, über welche Daten ausgetauscht werden können. Die vom Client verwendete IP-Adresse gilt hierbei im Sinne des Datenschutzes als personenbezogenes Datum.

Beim Abruf einer Webseite in einem Browser ist im Body der Antwort vom Server in der

Regel eine HTML-Datei enthalten. Diese HTML-Datei beschreibt den Aufbau der Webseite und wird vom Browser interpretiert und dargestellt. Falls die Datei Bilder, Videos, Fonts oder sonstige Inhalte referenziert, müssen diese wiederum durch eine HTTP-Anfrage an denselben oder einen anderen Server nachgeladen werden. Dieser Vorgang ist in Abbildung 1 dargestellt. Bei populären Webseiten können schnell mehrere hundert Objekte von unterschiedlichen Servern geladen werden. Das vollständige Laden der Webseite dauert oft mehrere Sekunden.

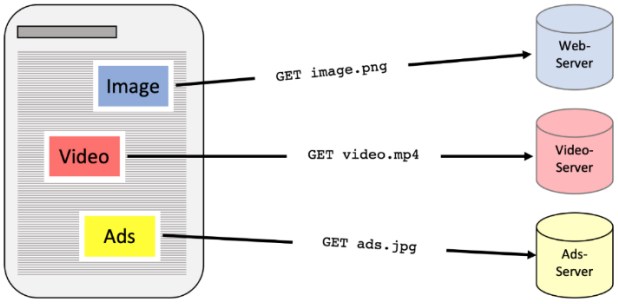

Abbildung 1: Einbindung externer Objekte auf einer Webseite.

Von besonderer Bedeutung beim Nachladen von Inhalten von Drittparteien ist hierbei der HTTP-Header. Dieser spezifiziert eine Anfrage oder Antwort und kann beispielsweise den Domain-Namen der Seite enthalten, die die benötigte Ressource referenziert hat (=Referer). Der Referer ist zwar eine freiwillige Angabe des Browsers, wird aber von den meisten Browsern in der Standard-Einstellung verwendet. Dies bedeutet, dass eine Drittpartei durch das Nachladen von Inhalten sowohl die IP-Adresse eines Clients über den IP-Header als auch die von ihm ursprünglich besuchte Webseite über den Referer im HTTP-Header erfahren kann. Parteien, die externe Objekte auf vielen Webseiten einbinden, erhalten so eine umfassende Übersicht der besuchten Webseiten des Clients. Es ist durch Einstellungen in der Datenschutz-Konfiguration des Browsers oder über Plug-Ins möglich, manche dieser Anfragen zum Nachladen von Inhalten zu unterdrücken. Firefox beispielsweise stellt dem Nutzer einen sog. Strict-Mode zur Verfügung, welcher bekannte Tracker blockiert und den Nutzer somit schützt.

Webseiten im Sinne dieses Artikels sind spezielle Uniform Resource Identifier (URI), die z.B. die folgende Form haben:

https://public.hochschule-trier.de/~knorr/AES/index.php?xy=fb

Dabei ist *https* (http über das kryptologische Protokoll TLS) das genutzte Protokoll. *public.hochschule-trier.de* ist die Domain, genauer eine Third Level Domain bzw. ein Rechnername in der Second Level Domain *hochschule-trier.de*. In der Studie in Abschnitt 4 wird auch der Begriff *Subdomain* verwendet. *~knorr/AES/* ist der Pfad zur Datei *index.php*. Hinter dem Fragezeichen wird an die Variable *xy* der Wert *fb* übergeben. Die Second Level Domain ist für die folgende Untersuchung fix. Jede Änderung in der Third Level Domain, im Pfad, in der referenzierten Datei und in den übergebenen Parametern ergibt eine neue URI / Webseite, die eigene externe Objekte einbinden kann. Daher ist die Anzahl der untersuchten Webseiten schnell sehr groß.

2.2 Datenschutz

Unter dem Begriff Datenschutz wird der Schutz natürlicher Personen bei der Verarbeitung ihrer personenbezogenen Daten verstanden. Die Verarbeitung personenbezogener Daten ist in der Datenschutz-Grundverordnung, kurz DS-GVO, geregelt. Als personenbezogene Daten werden alle Informationen bezeichnet, die sich auf eine identifizierte oder identifizierbare Person beziehen (Artikel 4 DS-GVO). Bei Anwendbarkeit der DS-GVO sind Webseiten-Betreiber in Deutschland dazu verpflichtet, Angaben zu der Verarbeitung personenbezogener Daten und insbesondere Angaben zu der Weitergabe von personenbezogenen Daten an Dritte zu machen. Zu diesen personenbezogenen Daten zählt auch die IP-Adresse – alleine oder in Kombination mit dem Referer, da sich diese durch den entsprechenden Internetanbieter eindeutig einer Person zuordnen lässt (identifizierbare Person) [WSK16]. Im Allgemeinen dürfen nur Daten erhoben werden, die für festgelegte, eindeutige und legitime Zwecke genutzt werden. Diese Daten dürfen auch nur so lange gespeichert werden, wie es für diese Zwecke unbedingt notwendig ist (Artikel 5 DS-GVO).

Werden personenbezogene Daten verarbeitet, gilt nach Artikel 13 der DS-GVO eine Informationspflicht, der für Webseiten i.d.R. in Forum einer sog. DSE nachgekommen wird. Artikel 13 fordert u.a. Informationen über die folgenden Punkte:

- Namen und die Kontaktdaten des Verantwortlichen
- Kontaktdaten des Datenschutzbeauftragten
- Zwecke und Rechtsgrundlage der Verarbeitung
- Empfänger der Daten
- Übermittlung der Daten in ein Drittland oder an eine internationale Organisation
- Dauer der Speicherung
- Rechte der Betroffenen: Auskunft, Berichtigung, Löschung, Einschränkung, Widerspruch, Datenübertragbarkeit, Beschwerderecht, ggf. Widerrufsrecht

Bei der Einbindung von externen Objekten in Webseiten sind also die Empfänger (Firma und Firmensitz statt IP-Adresse) zu nennen. Übermittlungen in Drittländer müssen dabei gesondert berücksichtigt werden. Artikel 6 der DS-GVO listet die folgenden Möglichkeiten für die Rechtsgrundlage: Einwilligung, Erfüllung eines Vertrags, Erfüllung einer rechtlichen Verpflichtung, lebenswichtige Interessen, Wahrnehmung einer Aufgabe im öffentlichen Interesse und berechtigtes Interesse. Die Einwilligung als Rechtsgrundlage für Webseiten wird im §25 des neuen TTDSG (Gesetz über den Datenschutz und den Schutz der Privatsphäre in der Telekommunikation und bei Telemedien) explizit genannt. Eine Einwilligung muss der Verarbeitung vorausgehen. Sie wird i.d.R. über Checkboxen oder Zwei-Klick-Lösungen umgesetzt. Sie ist allerdings bzgl. der Verwaltung der Einwilligungen und möglichen Widerrufe aufwändig umzusetzen. Für Hochschulen mit Ihrem öffentlichen Bildungsauftrag kommt ebenso Artikel 6 Abs. 1 lit. c) („Wahrnehmung einer Aufgabe im öffentlichen Interesse“) in Betracht. Das „berechtigte Interesse“ ist laut DS-GVO für Behörden bzw. staatliche Hochschulen nicht anwendbar.

3 Das eingesetzte Tool *wirechuck*

Die Untersuchung eines Webauftritts auf externe Inhalte läuft in zwei Schritten ab. Der erste Schritt wird im Folgenden als *Crawling* (dt. „Kriechen / Krabbeln“) bezeichnet und dient der schnellen Identifizierung aller verlinkten Webseiten. Der zweite Schritt wird als *Scraping* (dt. „Kratzen“) bezeichnet und analysiert alle im Crawling identifizierten Seiten im Detail auf Fremdinhalte. Die Aufteilung in diese zwei Schritte erlaubt es, die Schritte unabhängig voneinander durchzuführen und unterschiedlich zu parametrisieren (z.B. für das Threading und die Delays). Diese Vorgehensweise erleichtert zudem die wiederholte Untersuchung eines Webauftritts, da die im ersten Schritt gefundenen Unterseiten wiederverwendet werden können. Für die Implementierung des Tools wurde die Programmier-Sprache Python verwendet.

Der Crawler erhält zwei Eingaben: (1) eine Start-URI und (2) eine zu durchsuchende Domain. Zu Beginn wird an die Start-URI eine HTTP-Anfrage gesendet. Die in der Antwort enthaltene HTML-Datei wird analysiert und auf Links (genauer `<a href> … </a>` im HTML-Code) zu weiteren Webseiten untersucht. Links auf Webseiten, die sich nicht in der angegebenen Domain befinden, werden ignoriert. An alle anderen Links wird wiederum eine HTTP-Anfrage gesendet. Dieser Vorgang wird solange wiederholt, bis keine weiteren Unterseiten mehr gefunden werden. Jedes Mal, wenn der Crawler eine HTTP-Antwort erhält, wird die in der dazugehörigen Anfrage enthaltene URI in eine Datei geschrieben (*<domain>.txt*). Wenn der gesamte Vorgang abgeschlossen ist, erhält der Nutzer als Ausgabe eine Text-Datei mit sämtlichen erreichbaren Unterseiten des durchsuchten Webauftritts. Für das Crawling wird das Python-Framework Scrapy (https://scrapy.org/) verwendet, welches für das Durchsuchen von Webseiten auf sogenannte Spider-Klassen (dt. „Spinne“) zurückgreift. Von besonderer Bedeutung für den Nutzer sind die Auszüge aus den Dateien settings.py und crawl_spider.py, die in Listing 1 und 2 abgebildet sind.

```
22. ROBOTSTXT_OBEY = True
…
30. DOWNLOAD_DELAY = 0.1
```

Listing 1: Auszug aus der Datei settings.py

```
11. rules = (
12.     Rule(LinkExtractor(allow=(), deny=('<Insert here>')),
    callback='parse_items', follow=True),
13. )
```

Listing 2: Auszug aus der Datei crawl_spider.py

ROBOTSTXT_OBEY legt fest, ob auf die robots.txt Datei des zu untersuchenden Webauftritts Rücksicht genommen werden soll und DOWNLOAD_DELAY gibt an, wie lange zwischen dem Herunterladen der einzelnen HTTP-Antworten gewartet werden soll. In Zeile 12 von Listing 2 kann ein Nutzer an Stelle von *<Insert here>* beliebige Strings angeben, die bei der Suche ignoriert werden sollen. Ein Beispiel hierfür wäre eine Subdomain, die nicht mit untersucht werden soll.

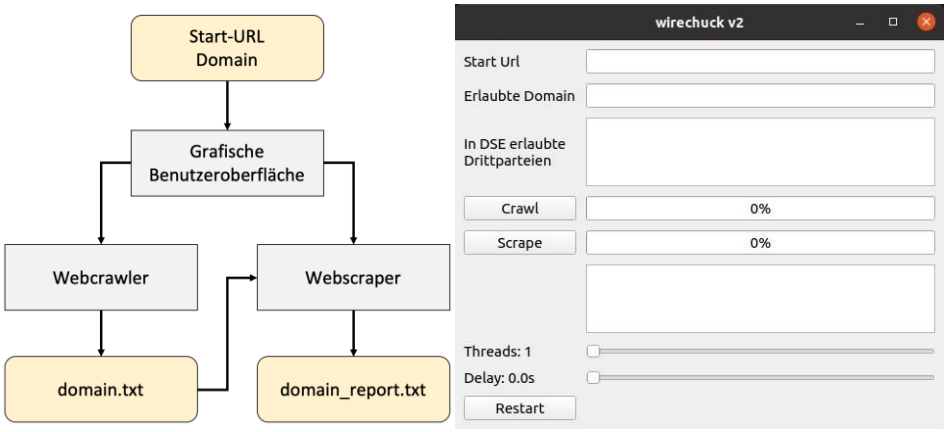

Abbildung 2: *wirechuck* Architektur (links) und Bedienoberfläche (rechts)

Der Scraper benötigt insgesamt vier Eingaben. (1) Die Domain wird verwendet, um den Namen der Datei *<domain>.txt* abzuleiten und die Datei auszulesen. Die in dieser Datei enthaltenen URIs werden anschließend der Reihe nach in einem Browser aufgerufen. Um den hierbei entstehenden Netzwerkverkehr aufzuzeichnen, wird ein Proxy-Server verwendet. Dies bedeutet, dass die HTTP-Anfragen vom Browser an diesen Server und anschließend vom Proxy-Server an die Ziel-Adresse gesendet werden. Jedes Mal, wenn eine dieser Webseiten vollständig geladen ist, werden alle entstandenen externen Verbindungen ausgelesen und gegen (2) eine Liste von erlaubten Drittparteien abgeglichen. Alle Verbindungen, die nicht explizit erlaubt sind, werden in eine Text-Datei mit dem Namen *<domain>_report.txt* geschrieben. Dieser Schritt dauert viel länger als das Crawling, da gewartet werden muss, bis alle Webseiten, inklusive externer Inhalte, vollständig geladen sind. Daher werden an dieser Stelle mehrere Threads verwendet, wobei in jedem Thread ein eigenes Browser-Fenster geöffnet wird. (3) Die Anzahl der zu verwendenden Threads sowie (4) die Dauer, die nach Öffnen einer Webseite gewartet werden soll (Delay), können vom Nutzer selbst festgelegt werden. Für die Implementierung des Scrapers wird das Framework Selenium (https://www.selenium.dev) verwendet, welches es dem Nutzer ermöglicht, Eingaben in einem Browser, in diesem Fall das Öffnen eines Browsers und der Aufruf einer bestimmten URI in diesem, zu automatisieren. Webdriver dafür werden von fast allen herkömmlichen Browsern angeboten, wobei hier Google Chrome verwendet wird.

Die Architektur und die Benutzeroberfläche des Tools sind in Abbildung 2 dargestellt. Der Source Code und Installationshinweise stehen unter [wich] zur Verfügung. Müller [Mü22] beschreibt das Tool im Detail.

4 Studie an der Hochschule Trier

Die Hochschule Trier (HST) hat aktuell ~7.000 Studierende und ~750 Beschäftige. Sie hat mit Trier, Birkenfeld und Idar-Oberstein mehrere Standorte. Die HST nutzt zwei Second Level Domains: (1) hochschule-trier.de für die Standorte Trier und Idar-Oberstein und (2) umwelt-campus.de für den Standort Birkenfeld. Es gibt zwei Rechenzentren, eines in Trier, eines in Birkenfeld. Die Rechenzentren betreiben die Server für die Web-Infrastruktur selbst. Die DSE für die Webseiten der HST stammt vom Mai 2018, vgl. [DSEHST]. Darin werden keine externen Parteien als Empfänger von personenbezogenen Daten aufgeführt.

Die Untersuchung des öffentlichen Teils des Web-Auftritts der Hochschule Trier wurde im Zeitraum 21.-24.02.2022 von einer externen IP-Adresse außerhalb des Hochschulnetzes durchgeführt. Als Start-URI wurde https://www.hochschule-trier.de und als Allowed-Domain hochschule-trier.de verwendet. Subdomains wie fsi.hochschule-trier.de und public.hochschule-trier.de wurden daher ebenfalls untersucht. Die Rechenzentren wurden vorab über die Untersuchung informiert. Für die Untersuchung des Webauftritts der Hochschule Trier wurde eine virtuelle Maschine mit 4GB RAM, 2 Prozessor-Kernen und einem Ubuntu 20.04 Image verwendet.

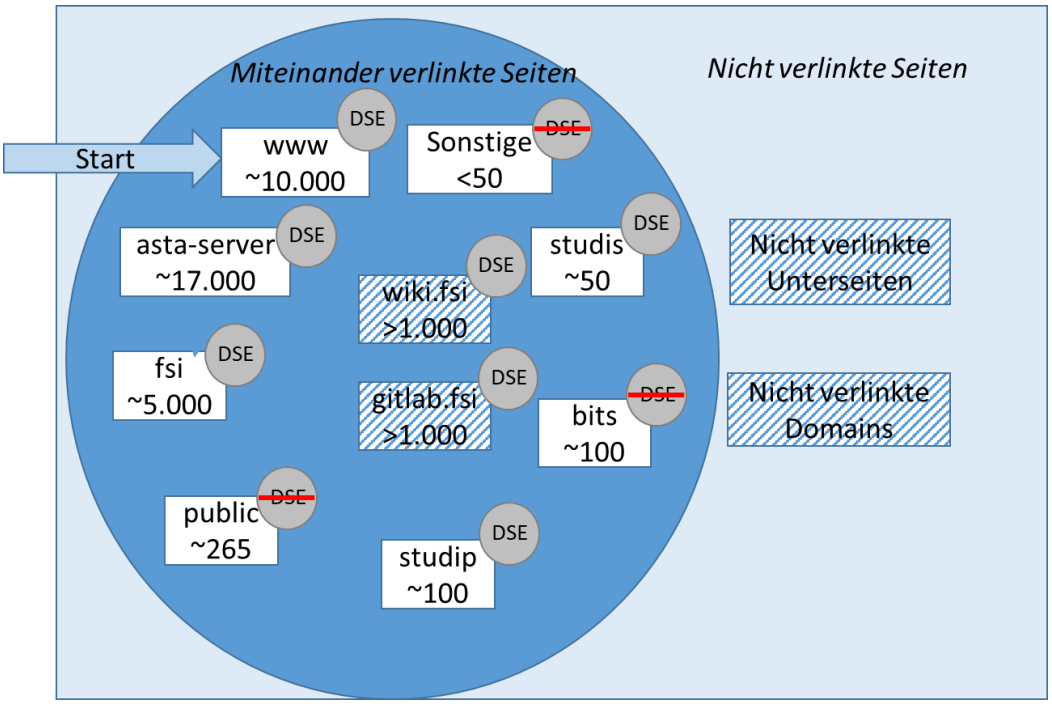

Abbildung 3: Öffentlich erreichbare Webseiten und deren Anzahl unter hochschule-trier.de. In der Studie exkludierte Unterseiten sind schraffiert dargestellt. Zusätzlich ist dargestellt, ob die Unterseiten eigene DSEn besitzen oder nicht.

4.1 Ergebnisse des Crawlings

Insgesamt wurden 32.574 Unterseiten in 2 Stunden und 26 Minuten gefundenen. Das entspricht ~220 gefundener Unterseiten pro Minute. Für das Crawling wurde auf die Vorgaben der *robots.txt* Datei Rücksicht genommen und es wurde ein Download-Delay von 0,1 Sekunden gewählt. Dies bedeutet, dass maximal 10 Anfragen pro Sekunde gesendet wurden. Abbildung 3 gibt einen Überblick der untersuchten Unterseiten. Alle Subdomains mit mehr als 50 Links sind aufgeführt.

4.2 Ergebnisse des Scrapings

Die Überprüfung der Unterseiten auf externe Verbindungen hat 32 Stunden und 2 Minuten gedauert. Hierfür wurden zwei Threads und ein Delay von 5 Sekunden verwendet. Pro Minute wurden also ~17 Unterseiten überprüft. Es wurden 69 Unterseiten mit externer Verbindung entdeckt. Die dazugehörigen externen Domains sind in Tabelle 1 dargestellt.

Domain	Anzahl Anfragen	Anzahl in Prozent
panopto-cache.zdv.net	1175	81,65%
www.youtube-nocookie.com	63	4,38%
www.gstatic.com	57	3,96%
fonts.gstatic.com	25	1,74%
www.youtube.com	18	1,25%
www.google.com	15	1,04%
f.vimeocdn.com	12	0,83%
i.vimeocdn.com	12	0,83%
i.ytimg.com	9	0,63%
yt3.ggpht.com	9	0,63%
fonts.googleapis.com	8	0,56%
fresnel.vimeocdn.com	8	0,56%
idp.fh-trier.de	7	0,49%
googleads.g.doubleclick.net	4	0,28%
player.vimeo.com	4	0,28%
static.doubleclick.net	2	0,14%
unpkg.com	2	0,14%
vimeo.com	2	0,14%
www.alfresco.com	2	0,14%
cdn.embed.ly	1	0,07%
cdnjs.cloudflare.com	1	0,07%
code.jquery.com	1	0,07%
ajax.googleapis.com	1	0,07%
dugie.de	1	0,07%

Tabelle 1: Externe Verbindungen nach Domain

5 Diskussion

Abbildung 3 zeigt eine Übersicht der über das Crawling erreichten Webseiten. *wirechuck* erkennt sehr zuverlässig alle verlinkten Seiten. Dies wurde bei der Entwicklung auch über einen eigens dafür erstellten Webauftritt geprüft. Das Tool kann allerdings nur die verlinkten Seiten finden. Nicht verlinkte Unterseiten oder nicht verlinkte Subdomains können nicht entdeckt werden. Subdomains kann es sehr viele geben. Für einen umfassenden Test müsste eine Liste mit allen Subdomains als Startadresse durchgetestet werden.

Es hat sich gezeigt, dass es einige Wikis und Webseiten zur Bereitstellung von Programm Code wie z.B. Gitlab gibt, die untereinander stark vernetzt sind. Hier genügt oft die Untersuchung der Startseite, da alle Seiten ähnlich aufgebaut sind. Bei den ersten Versuchen mit dem Tool stellte sich heraus, dass insbesondere die beiden Seiten fsi.hochschule-trier.de/wiki und gitlab.fsi.hochschule-trier.de mehr als 100.000 bzw. 60.000 Unterseiten besitzen. Aus Performanzgründen wurden diese beiden Seiten exkludiert. Während der Untersuchung wurden 17.000 Seiten unter asta-server.hochschule-trier.de untersucht. Auch hier ist ein enthaltenes Wiki die Ursache für die hohe Zahl an Seiten. Insgesamt sucht *wirechuck* also gründlich und teilweise zu sorgfältig, so dass einzelne Unterdomains ausgeschlossen werden müssen.

Bei der Untersuchung der Subdomains ist wichtig zu wissen, ob die Seiten eine eigene DSE aufweisen oder die Standard DSE der HST [DSEHST] verwenden. Das ist leider automatisiert noch nicht über das Tool feststellbar und muss manuell getan werden (vgl. Abbildung 3). Wird eine eigene DSE gefunden, muss diese gegen die entdeckten externen Verbindungen überprüft werden.

Beim Scraping zeigte sich, dass erfreulicherweise über 99% der untersuchten Webseiten keine externen Inhalte einbinden. Die DSE muss daher für diese Seiten nicht angepasst werden. Tabelle 1 zeigt die gefundenen externen Domains und die Häufigkeit ihres Aufrufs. panopto-cache.zdv.net ist die häufigste externe Adresse. Es handelt sich dabei um eine zentral für die Hochschulen in Rheinland-Pfalz vom Zentrum für Datenverarbeitung der Universität Mainz zur Verfügung gestellte Videoplattform. Diese wurde während der Pandemie-Jahre für die Aufzeichnung und Bereitstellung von Screencasts und Videos zu Lehrveranstaltungen intensiv genutzt. Mit der Universität Mainz besteht ein Vertrag zur Auftragsverarbeitung. Die nächsten fünf Domains sowie einige weitere Einträge wie i.ytimg.com und static.doubleclick.net kommen aus dem Google-Umfeld und sind im Hochschulkontext durchaus kritisch zu sehen [WK15]. Unter idp.fh-trier.de ist der Shibboleth-Identity-Provider zu finden, der vom Rechenzentrum der HST betrieben wird und aus technischen Gründen noch die alte Second Level Domain fh-trier.de statt hochschule-trier.de verwendet.

Etwa 10.000 der 32.000 Seiten werden über das freie Content-Management-System Typo3 (https://typo3.org/) erstellt. Dafür gibt es an der HST ein professionelles Team von Web-Redakteuren, die in der Vergangenheit hinsichtlich Datenschutz entsprechend sensibilisiert und geschult wurden. Es gibt klare Verantwortlichkeiten und Prozesse für die Erstellung und Pflege der Typo3-Webseiten. Die meisten externen Objekte wurden auf Webseiten eingebunden, die nicht über Typo3 erstellt wurden. Das sind z.B. Webseiten

für Forschungsprojekte, Konferenzen und Online-Berechnungstools, wie sie für Hochschulen und andere Forschungseinrichtungen typisch sind.

Die Überprüfung und ggf. die Beseitigung der eingebetteten Drittparteien ist nur eine Facette der zahlreichen Aufgaben, die aus Datenschutzsicht bzgl. eines Web-Auftritts anfallen. Der rasche Wechsel der eingesetzten Technologien im Webumfeld erschwert die Arbeit und macht eine kontinuierliche Überprüfung der Inhalte und Sensibilisierung der beteiligten Personen notwendig. Andere Themen im Webumfeld an der Hochschule in den letzten Jahren umfassen z.B. die analytische Auswertung der Webseitenbesuche, die Einbindung von Social Media in den Webauftritt und den Umgang mit Cookies.

5.1 Vergleich zu verwandten Arbeiten

Knorr, Wambach und Schulte haben in [WK15, WSK16] die Start-Webseiten von Krankenhäusern und Hochschulen untersucht. Damals vor der Anwendung der DS-GVO betteten 50% der Seiten externe Inhalte ein. 15% der untersuchten Webseiten hatten entweder keine oder eine fehlerhafte DSE. Bei den eingebetteten Drittparteien konnte eine starke Dominanz amerikanischer Unternehmen wie Google / Alphabet festgestellt werden. Dieser Trend wird auch von Englehardt und Narayanan in [EN16] bestätigt: “All of the top 5 third parties, as well as 12 of the top 20, are Google-owned domains. In fact, Google, Facebook, Twitter, and AdNexus are the only third-party entities present on more than 10% of sites.“ Auch diese Studie bestätigt diesen Trend. Im Unterschied zu den angeführten Publikationen ist die Anzahl der Fremdparteien einbettenden Seiten bei der Tiefensuche in unserer Fallstudie geringer. Dies liegt daran, dass in den Publikationen lediglich die Startseiten eines Webauftritts untersucht wurden und dass seit 2018 die DS-GVO angewendet wird.

5.2 Handlungsempfehlungen

Vorgelagert ist die Identifizierung der für die Webseiten verantwortlichen Personen. Da oft die Hochschulleitung im Impressum angegeben ist, kann über die Registrierungsdaten für die Subdomain, die beim Rechenzentrum hinterlegt sind, auf die Verantwortlichen geschlossen werden. Die Empfehlung an die Verantwortlichen ist: entweder (1) auf die Einbindung externer Objekte zu verzichten, (2) diese durch Alternativen zu ersetzen oder (3) die Erstellung einer eigenen DSE, die auf die externen Parteien hinweist.

Ad (1) und (2): Die Einbindung von externen Videos und Fonts wird dem Ersteller von Webseiten leichtgemacht. Meist geschieht dies in Anleitungen über ein Copy&Paste von entsprechendem HTML-Code. Auf die Datenschutzprobleme wird dabei oft unvollständig hingewiesen. Im Gespräch mit den Verantwortlichen stellte sich die fehlende Information als häufigster Grund für die nicht dokumentierte Einbettung von externen Objekten heraus. Insbesondere die Verwendung externer Fonts kann in der Regel durch die lokale Bereitstellung von Fonts ersetzt werden. Für die Einbindung von Videos kann z.B. ein

Bild mit hinterlegtem Link auf das externe Video (am besten mit Hinweis auf die externe Quelle) statt einer direkten Einbindung des Videos oder eine Zwei-Klick-Lösung, die die Einwilligung des Nutzers einholt, verwendet werden.

Ad (3): Die Erstellung einer eigenen DSE wird vom Datenschutzbeauftragten der Hochschule durch Text-Vorlagen und ein Schulungsvideo unterstützt. Die Erarbeitung einer eigenen DSE erhöht das Bewusstsein und das Verständnis für den Datenschutz und zeigt den Verantwortlichen zusätzlich das Spannungsfeld zwischen aktuellen Web-Technologien und dem Datenschutz deutlich auf.

5.3 Ausblick

Die automatische Identifikation von Fremdinhalten eines gesamten Web-Auftritts erleichtert die Prüfung der DSE für die verantwortlichen Parteien wie die Datenschutzbeauftragten enorm. Verstöße werden unmittelbar anzeigt. Eine regelmäßige Prüfung ist möglich. Das vorgestellte Tool kann eine Über- oder Unterdeckung zuverlässig feststellen. Für die Zukunft sind folgende Verbesserungen denkbar:

- Die Überprüfung des internen Webauftritts (Intranet), der oft nur nach vorheriger Authentisierung z.B. über Shibboleth erreichbar ist, wird noch nicht unterstützt. Die Herausforderungen liegt hierbei darin, die Anmeldung in Shibboleth zu automatisieren. Hierzu wird ein zusätzlicher Server, der sog. Identity Provider, genutzt. Ähnlich herausfordernd ist der Umgang mit dynamisch von Webseiten eingeblendeten Pop-Up-Fenstern, die z.B. Einwilligungen zu Cookies einfordern.
- Im Zuge einer regelmäßigen Überprüfung eines Web-Auftritts wäre die Verwendung einer Datenbank statt Text-Dateien zur Speicherung der Ergebnisse wünschenswert. So könnten einzelne Seiten gezielt nachgeprüft werden und eine zeitliche Entwicklung der gefunden und gelösten Probleme aufgezeigt werden.
- Der Abgleich zwischen den Objekten, die extern eingebettet sind, und der in der DSE gelisteten externen Partner geschieht im vorgestellten Tool manuell (über Angaben zulässiger externer Domains). Das Auffinden und automatisierte Auslesen und Auswerten einer DSE ist eine Herausforderung. Ein standardisiertes, maschinenlesbares Format für eine DSE wäre dafür hilfreich. Insbesondere müsste für jede Fremdpartei mindestens das Tripel (1) Name / Anschrift der Partei, (2) genutzte Domain / IP-Adressen und (3) Rechtsgrundlage der Einbettung erfasst werden. Mit dem Projekt „Platform for Privacy Preferences“ (https://www.w3.org/P3P/) gab es dazu einen Vorschlag, der aber kaum genutzt und nicht mehr weiterentwickelt wird.

Literatur

[DSEHST] Datenschutzerklärung des Web-Auftritts der Hochschule Trier, https://www.hochschule-trier.de/datenschutz

[EBFK18] Ermakova, E., Bender, B., Fabian, B., Klimek, K.: “Web Tracking – A Literature

Review on the State of Research“, Proceedings of the 51st Hawaii International Conference on System Sciences, 2018

[EN16] Englehardt, E., Narayanan, A.: “Online Tracking: A 1-million-site Measurement and Analysis”, Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, pp. 1388–1401, https://doi.org/10.1145/2976749.2978313

[MM12] Mayer, J., Mitchell, J.: “Third-Party Web Tracking: Policy and Technology”, 2012 IEEE Symposium on Security and Privacy, 2012, pp. 413-427, https://doi.org/10.1109/SP.2012.47

[Mü22] Müller, D.: „Automatisierte Überprüfung von Datenschutzerklärungen bezüglich Fremdinhalten“, Bachelorarbeit, Hochschule Trier, März 2022

[wich] Source Code und Installationsanleitung zum Tool *wirechuck*: https://seafile.rlp.net/d/af0df4d9566a4b5caa59/

[WK15] Wambach, T., Knorr, K.: „Technische Prüfung der Datenschutzerklärungen auf deutschen Hochschulwebseiten“, Information Security Day (ISD), Würzburg, 2015.

[WSK16] Wambach. T., Schulte, L., Knorr, K.: „Einbettung von Drittinhalten im Web“, DuD - Datenschutz und Datensicherheit 40(8): 523-527 (2016)