Automatisierte Überprüfung von Webauftritten auf Fremdinhalte

Konstantin Knorr, David Müller

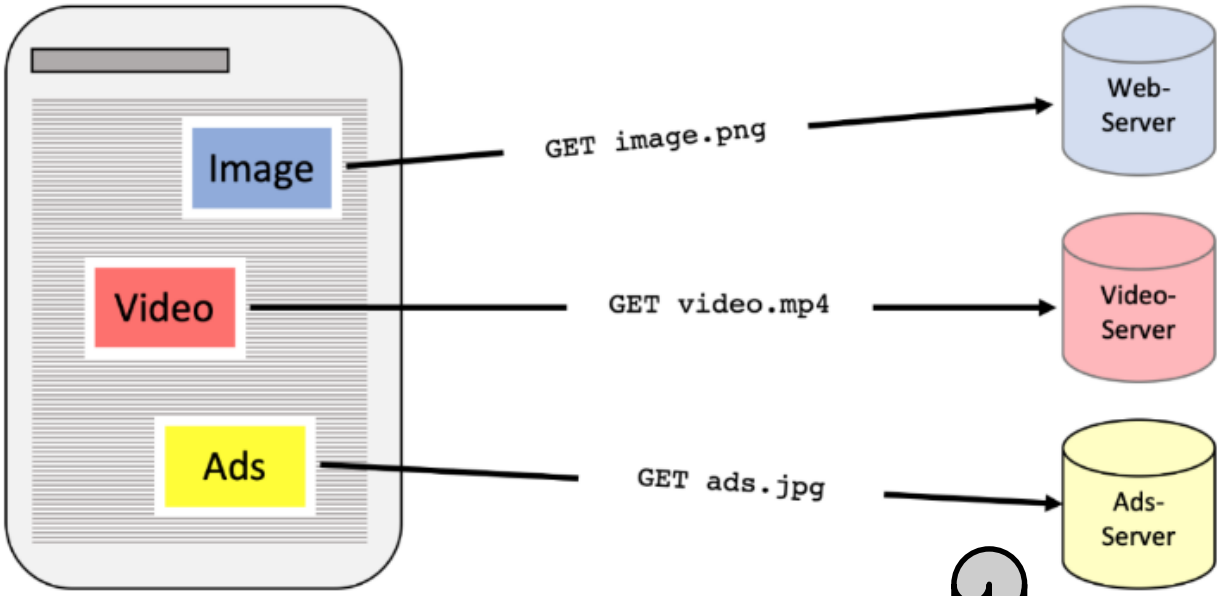
Workshop Recht und Technik:
Datenschutz im Diskurs (RuT)

Hamburg, 26.9.2022

Trier University
of Applied Sciences

HOCHSCHULE TRIER

Motivation und Ausgangslage



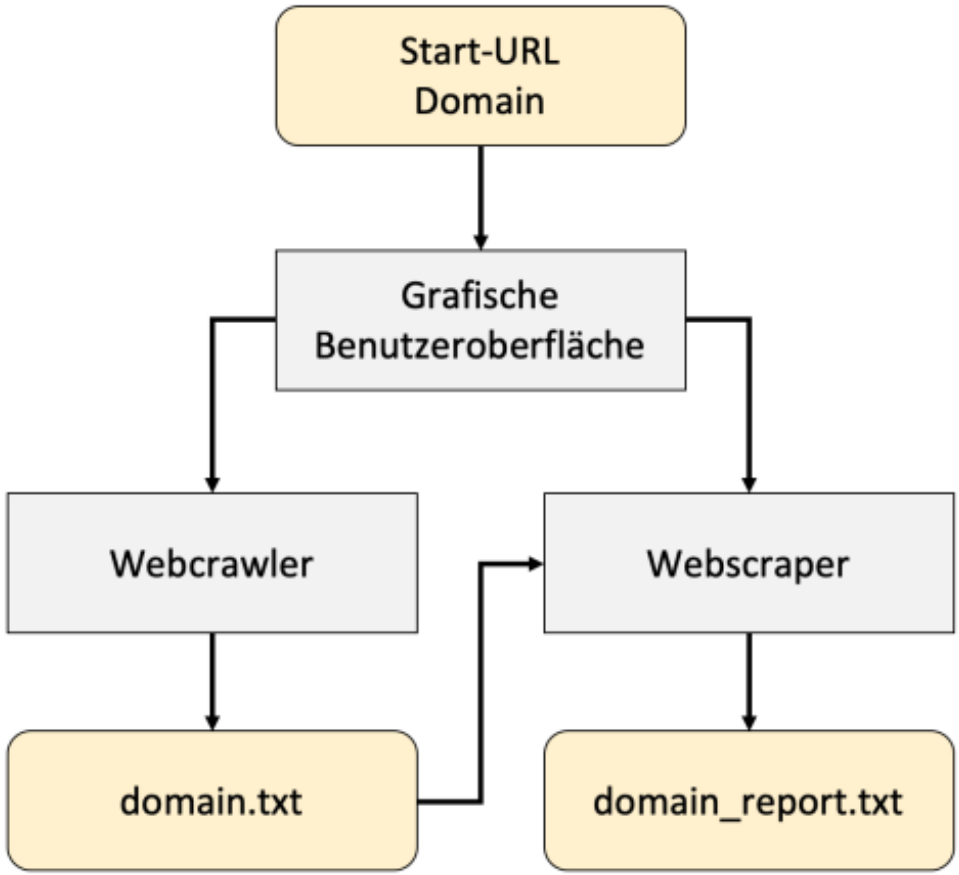
Datenschutzerklärung

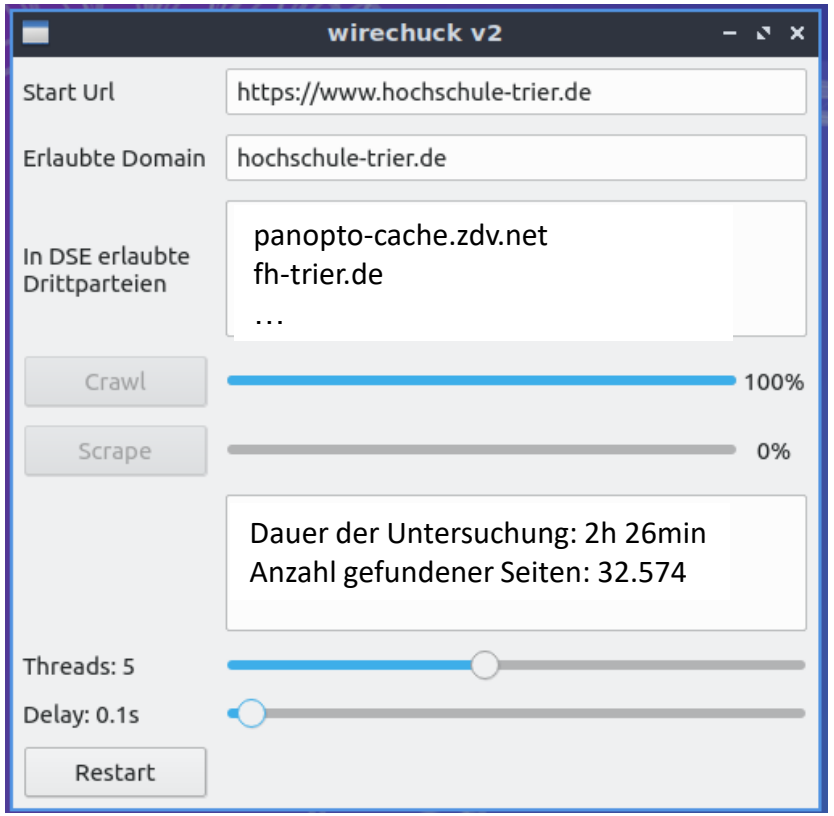
…

- „Fremdpartei“
- Zweck
- Rechtsgrundlage

…

Das eingesetzte Tool *wirechuck*



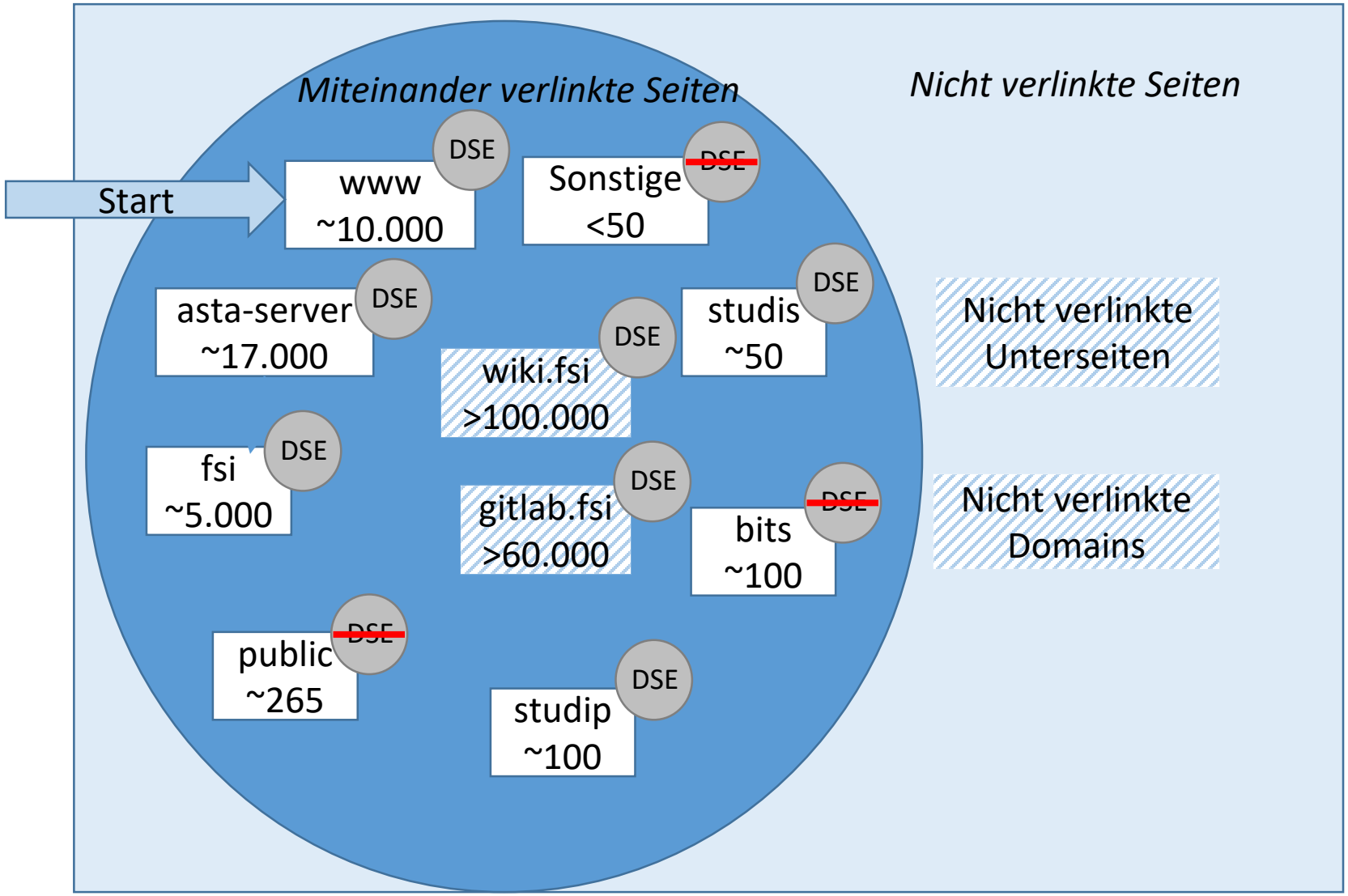


Download-Link: https://seafile.rlp.net/d/af0df4d9566a4b5caa59/

Studie an der Hochschule Trier

- Hochschule Trier: ~7.000 Studierende, ~750 Beschäftige
- Standorte: Trier, Birkenfeld, Idar-Oberstein
- Zwei Second Level Domains:
  - hochschule-trier.de
  - umwelt-campus.de
- Zeitraum der Untersuchung: 21.-24.02.2022
- Start-URI: www.hochschule-trier.de
- Verwendete *wirechuck* IP: außerhalb des Hochschulnetzes
- Verwendete HW und SW: virtuelle Maschine mit 4GB RAM, 2 Prozessor-Kernen und einem Ubuntu 20.04 Image
- Crawling: 32.574 Unterseiten in 2 Stunden und 26 Minuten (Delay=0,1 sec)
- Scraping: 32 Stunden und 2 Minuten (Delay=5 sec, 2 Threads)

Ergebnisse des Crawlings „Landkarte www.hochschule-trier.de“



Ergebnisse des Scrapings
Eingebundene Drittparteien

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%

Diskussion und Ausblick

- Automatisierung bei der hohen Anzahl von Seiten notwendig
- Nur ~2% (67 von 32.574) Seiten verwenden Fremdinhalte
- Wirechuck arbeitet sehr zuverlässig: Viele Unterseiten von Wikis, CMS und Gitlab / Github
- Gutes Abschneiden des „Standard"-Webauftritts mit Typo3
- Probleme bei Projekt- und Konferenzseiten
  ➔ Schulungsvideo, Templates für DSE, „Zulassungsprozess"
- *wirechuck*-Ergebnisse in DB speichern (Nachprüfung, Statistik)
- Intranet untersuchen, automatisierter Shibboleth-Login notwendig
- *wirechuck*-Erweiterung auf Cookie-Prüfung
- Manueller Abgleich mit DSE notwendig, standardisierte DSE wäre hilfreich inkl.
  - Name + Anschrift der Fremdpartei
  - Domain / IP-Adresse
  - Zweck
  - Rechtsgrundlage

Vielen Dank für die Aufmerksamkeit

Konstantin Knorr
knorr@hochschule-trier.de