URNs oder Urnen?

WHOIS

Frederik Stey

Angewandte Informatik
Projektleiter & Full-Stack Entwickler im Bereich der Industrieautomation

MAGRATHEA LABORATORIES
CREATORS OF NEW WORLDS

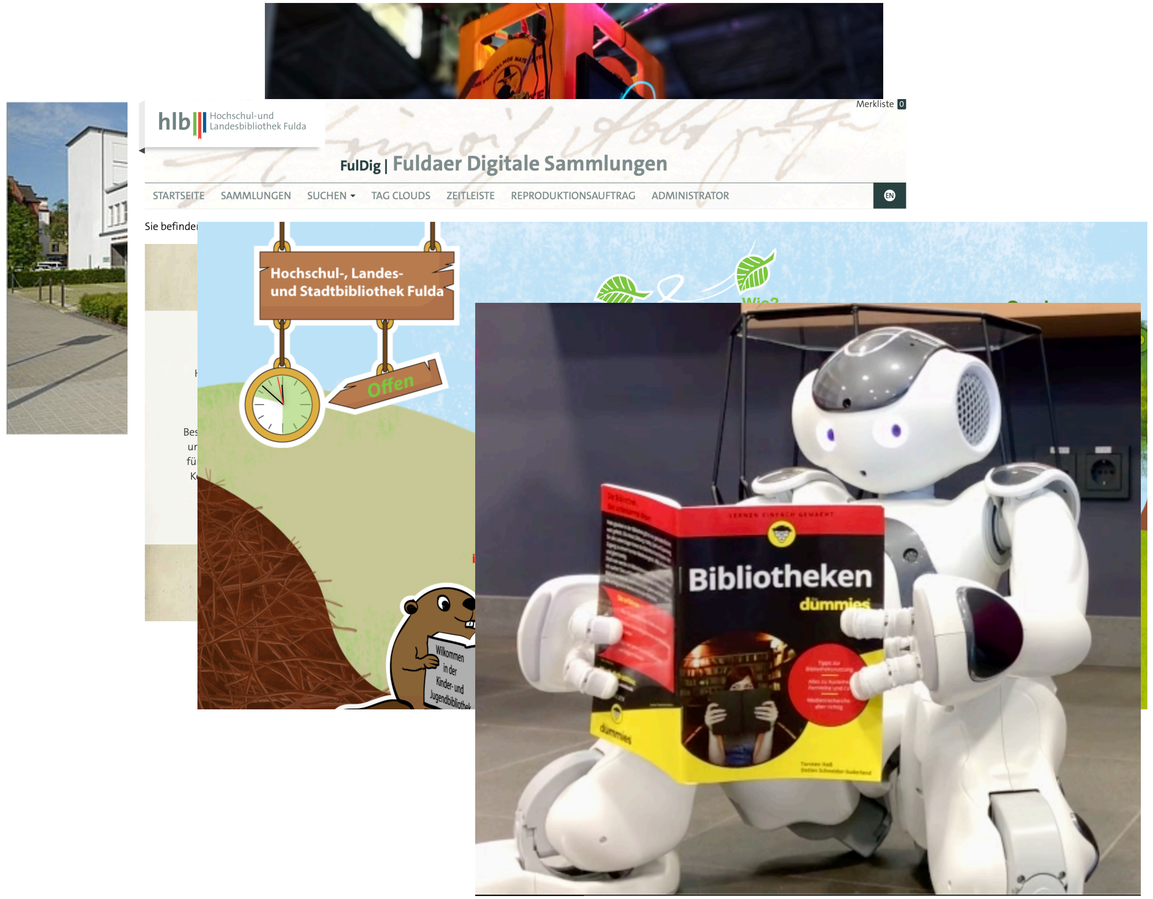
CCC Fulda
Chaos macht Schule

hlb

Leiter Digitale Dienste
IT-Sicherheitsbeautragter
Digitalisierung
Sonderprojekte
MINT/Makerspace

HUMBOLDT-UNIVERSITÄT ZU BERLIN

M.A. LIS



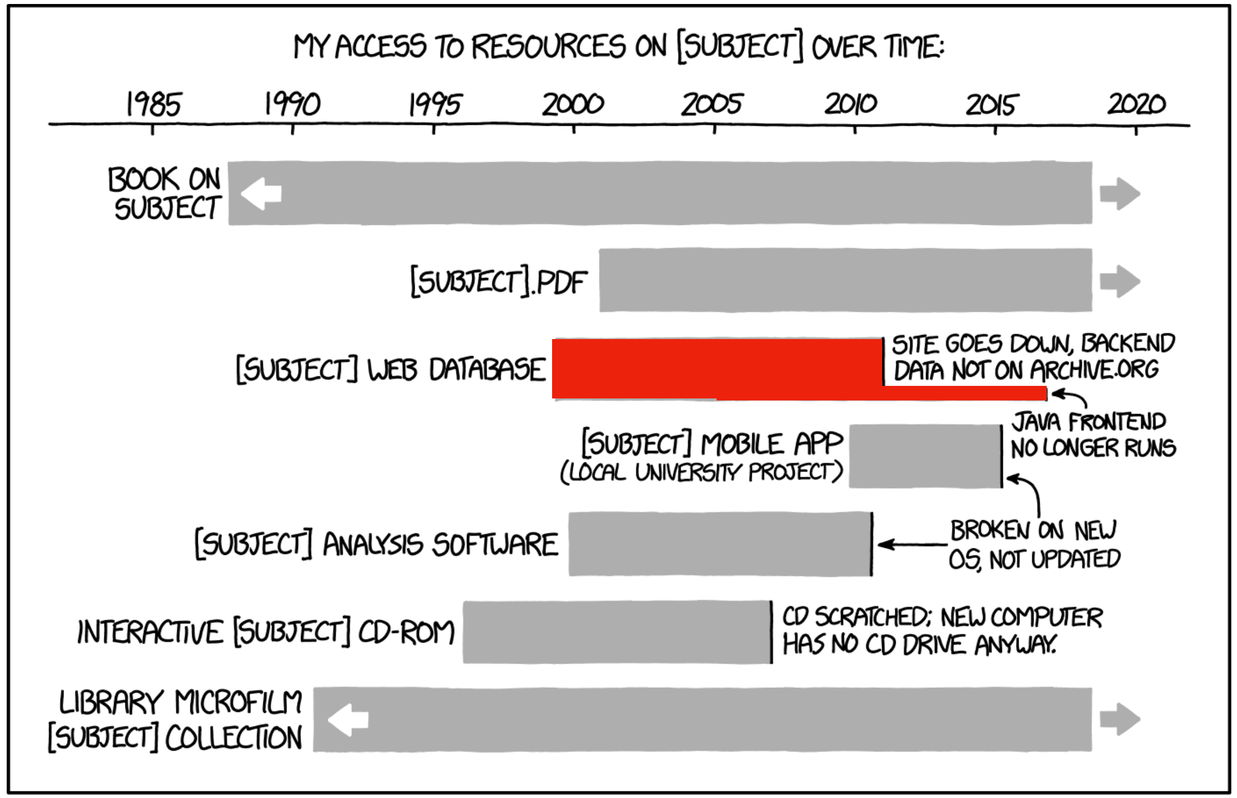
THE ADVENTURES OF

IBIANA STUDENT

AND THE RAIDERS OF

DEAD LINKS AND THE LOST RESSOURCES

PROBLEM?



IT'S UNSETTLING TO REALIZE HOW QUICKLY DIGITAL RESOURCES CAN DISAPPEAR WITHOUT ONGOING WORK TO MAINTAIN THEM.

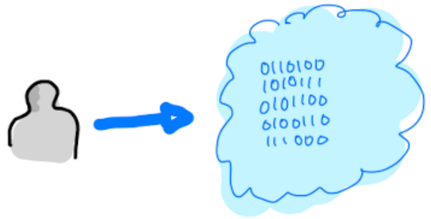
https://xkcd.com/1909/

Beispiel 404 Webseite

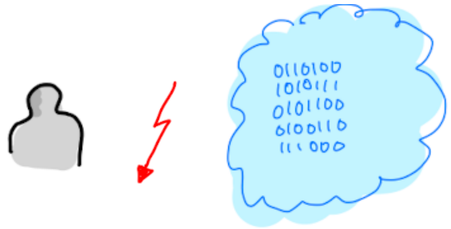
LINKS? RESSOURCEN?

URI = **dnb.de**

URL = **http://dnb.de**



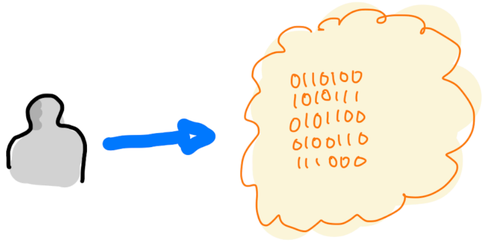
link rot



content drift

= reference rot

soft 404



„…Beobachtungen sind geeignet,
jegliches Vertrauen in die wissenschaftlichen Bibliotheken
zu **zerstören**, wenn es um Persistent Identifier geht."

Schändet nun auch die SUB Hamburg **Permalinks**?

Permalinks sind der ThULB Jena **wurscht**.

DNB **inkompetent**: Nach wie vor **defekte Links**.

Digitale historische Tageszeitungen
Allzu lückenhaft und ungepflegt (**defekte Links** z.B. in **Fulda**)

Blogtitel von https://archivalia.hypotheses.org

„…Beobachtungen sind geeignet,
jegliches Vertrauen in die wissenschaftlichen Bibliotheken
zu **zerstören**, wenn es um Persistent Identifier geht."

Schändet nun auch die SUB Hamburg **Permalinks**?

Permalinks sind der ThULB Jena **wurscht**.

DNB **inkompetent**: Nach wie vor **defekte Links**.

Digitale historische Tageszeitungen
Allzu lückenhaft und ungepflegt (**defekte Links** z.B. in **Fulda**)

Schatz kommst du ins Bett?

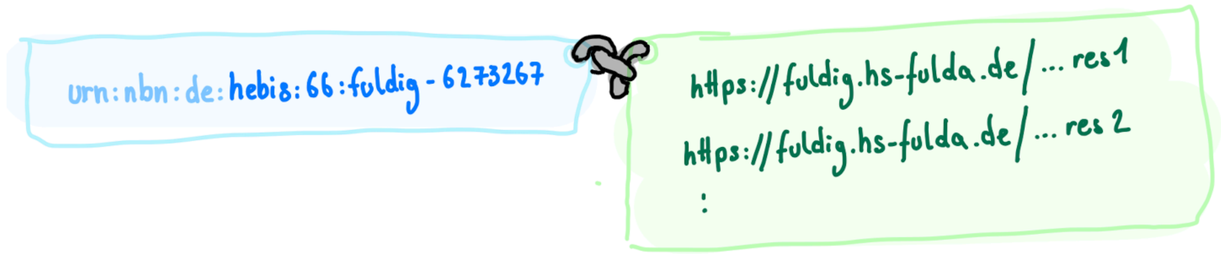
Was?

Ich kann nicht.
Das ist wichtig!

Jemand im Internet **irrt** sich!

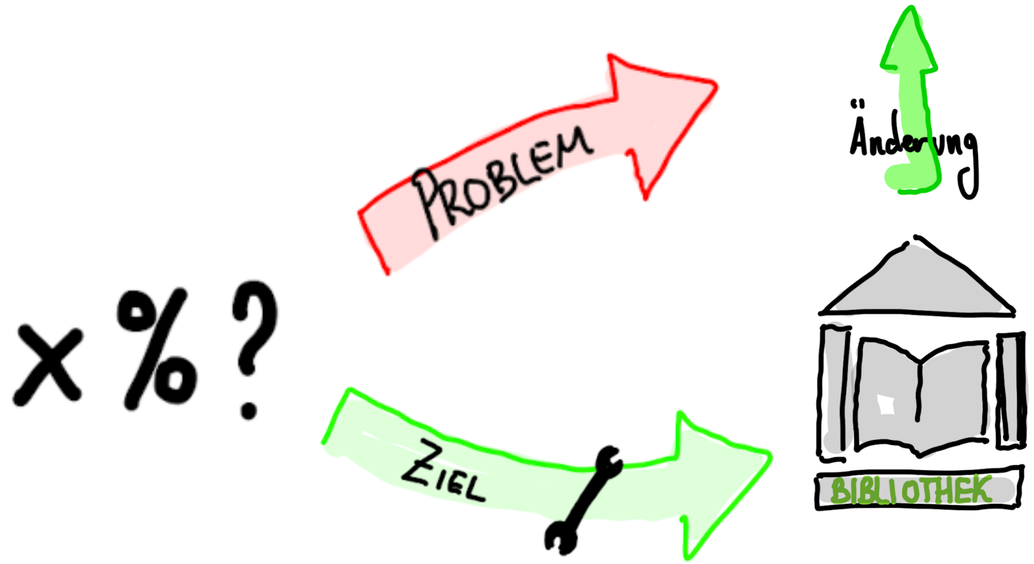
Blogtitel von https://archivalia.hypotheses.org

PERSISTENTE IDENTIFIER



DEUTSCHE NATIONAL BIBLIOTHEK

URN Service



10.000.000 zufällige URLs

testet **URL** zu urn

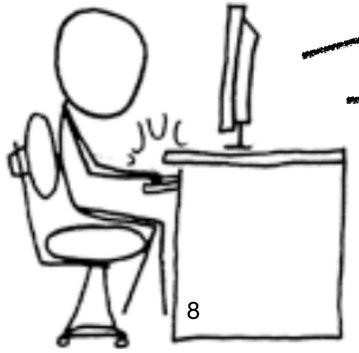
DNB?

Ja?

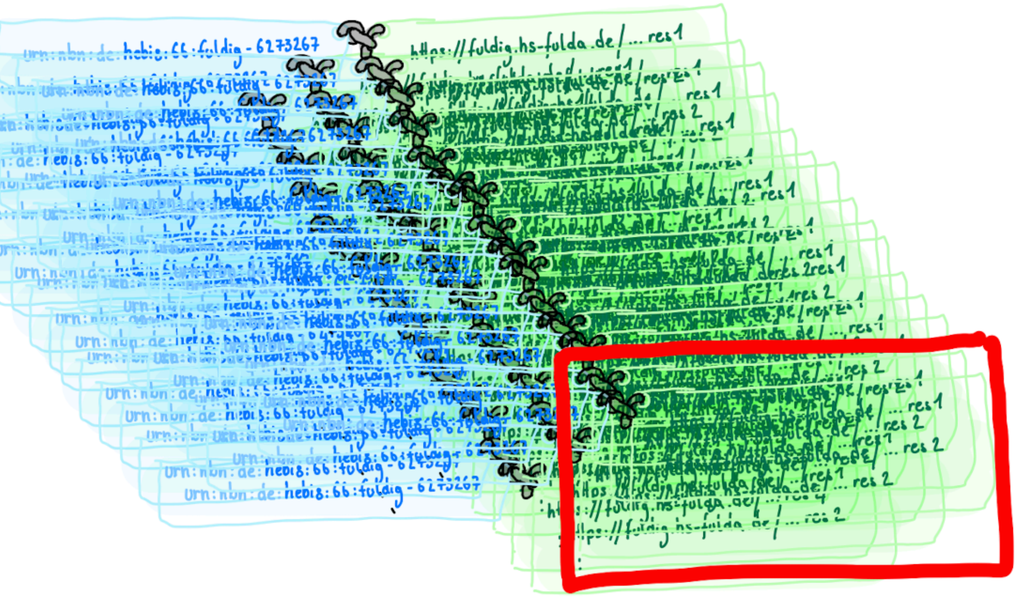
URNs?

klar!

hier **~36.000.000 URNs**



QUANTIFIZIERUNG DER ERREICHBARKEIT



Zufällige URNs

Kontrolle

Soft 404 Prüfung

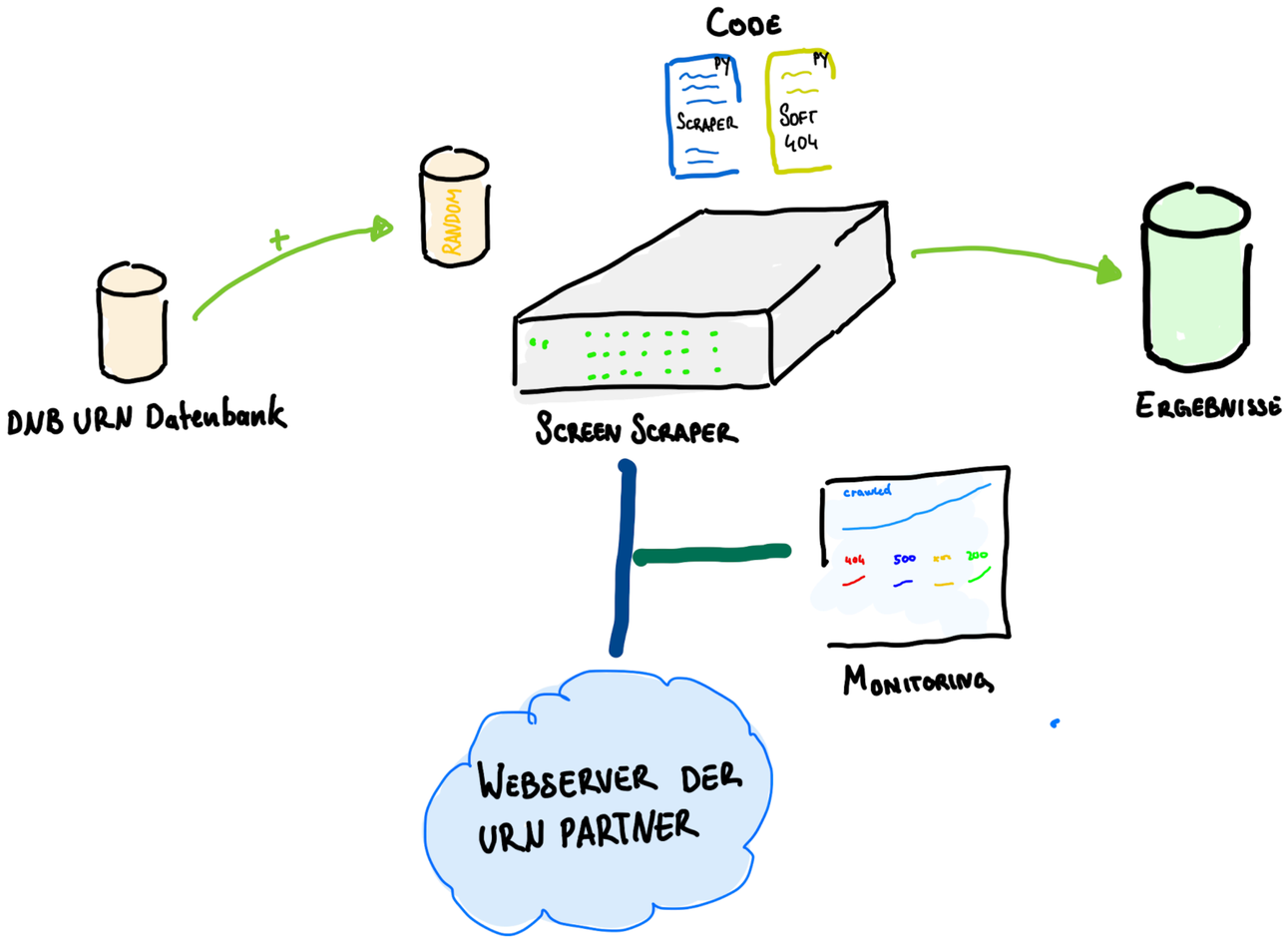
Text und Datamining

v.B. ethischer und rechtlicher Aspekte

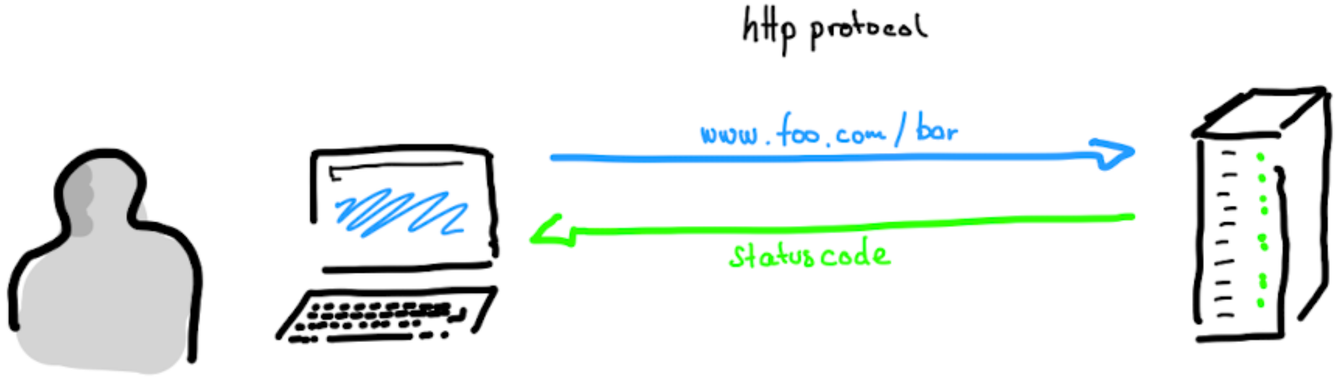
Erkenntnis

x % ?

HARD- & SOFTWARE



HTTP STATUSCODES



200 *ok* Die Anfrage wurde erfolgreich verarbeitet und die angefragte Ressource wird übertragen (vgl. rfc7231, S.51).

400 *Bad Request* Der Server kann die Anfrage nicht bearbeiten, da die Anfrage fehlerhaft war (vgl. rfc7231, S.58).
401 *Unauthorized* Die Anfrage wurde ohne gültige Authenifizierungsdaten gestellt (vgl. rfc7235, S.6).
403 *Forbidden* Die Anfrage wurde abgelehnt. Weitere Informationen sind in der Antwort enthalten (vgl. rfc7231, S.59).
404 *Not Found* Die angeforderte Ressource ist nicht verfügbar beziehungsweise konnte nicht gefunden werden.(vgl. rfc7231, S.59).
409 *Conflict* Die Anfrage konnte nicht einfeutig einer Ressource zugewiesen werden (vgl. rfc7231, S.60).
410 *Gone* Die angeforderte Ressource wurde permanent entfernt (vgl. rfc7231, S.60).

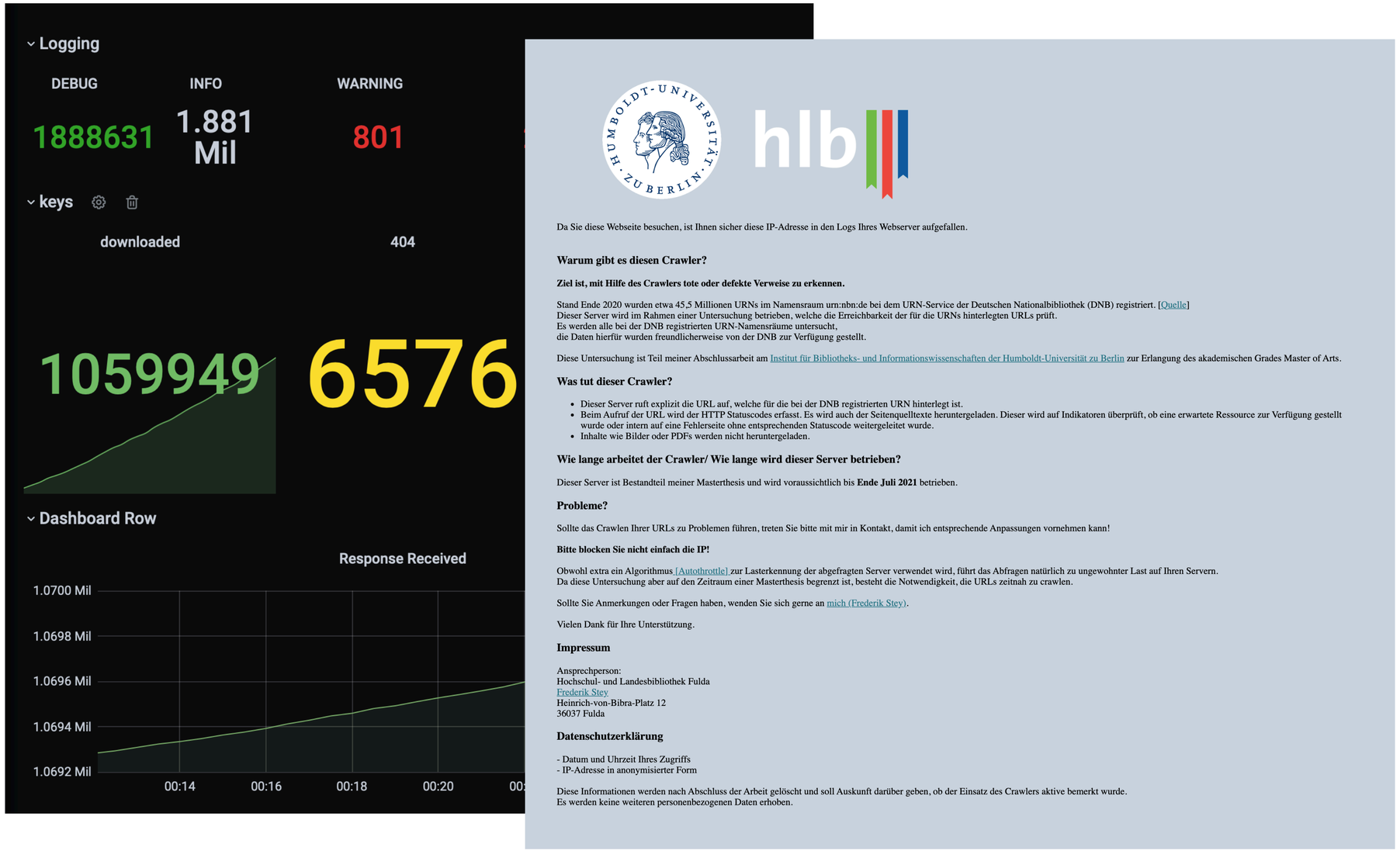
500 *Internal Server Error* Dies ist ein „Sammel-Statuscode“ für unerwartete Serverfehler (vgl. rfc7231, S.63).
501 Not Implemented Die Funktionalität, um die Anfrage zu bearbeiten, wird von diesem Server nicht bereitgestellt. Ursache ist zum Beispiel eine unbekannte oder nicht unterstützte HTTP-Methode.
502 *Bad Gateway* Der angefragte Server dient als Gateway/Proxy und hat eine ungültige Antwort von einem Server erhalten (vgl. rfc7231, S.63).
503 *Service Unavailable* Der Server kann temporär keine Anfragen, z.B. aufgrund Überlastung oder Wartungsarbeiten, verarbeiten (vgl. rfc7231, S.63).
504 *Gateway Timeout* Der angefragte Server dient als Gateway/Proxy und hat für eine Anfrage an einen Server ein Timeout erhalten. (vgl. rfc7231, S.63).

ETHISCHE & RECHTLICHE ASPEKTE

CDoS

Text &
Datamining

MONITORING



Da Sie diese Webseite besuchen, ist Ihnen sicher diese IP-Adresse in den Logs Ihres Webserver aufgefallen.

Warum gibt es diesen Crawler?

Ziel ist, mit Hilfe des Crawlers tote oder defekte Verweise zu erkennen.

Stand Ende 2020 wurden etwa 45,5 Millionen URNs im Namensraum urn:nbn:de bei dem URN-Service der Deutschen Nationalbibliothek (DNB) registriert. [Quelle]
Dieser Server wird im Rahmen einer Untersuchung betrieben, welche die Erreichbarkeit der für die URNs hinterlegten URLs prüft.
Es werden alle bei der DNB registrierten URN-Namensräume untersucht,
die Daten hierfür wurden freundlicherweise von der DNB zur Verfügung gestellt.

Diese Untersuchung ist Teil meiner Abschlussarbeit am Institut für Bibliotheks- und Informationswissenschaften der Humboldt-Universität zu Berlin zur Erlangung des akademischen Grades Master of Arts.

Was tut dieser Crawler?

- Dieser Server ruft explizit die URL auf, welche für die bei der DNB registrierten URN hinterlegt ist.
- Beim Aufruf der URL wird der HTTP Statuscodes erfasst. Es wird auch der Seitenquelltexte heruntergeladen. Dieser wird auf Indikatoren überprüft, ob eine erwartete Ressource zur Verfügung gestellt wurde oder intern auf eine Fehlerseite ohne entsprechenden Statuscode weitergeleitet wurde.
- Inhalte wie Bilder oder PDFs werden nicht heruntergeladen.

Wie lange arbeitet der Crawler/ Wie lange wird dieser Server betrieben?

Dieser Server ist Bestandteil meiner Masterthesis und wird voraussichtlich bis **Ende Juli 2021** betrieben.

Probleme?

Sollte das Crawlen Ihrer URLs zu Problemen führen, treten Sie bitte mit mir in Kontakt, damit ich entsprechende Anpassungen vornehmen kann!

Bitte blocken Sie nicht einfach die IP!

Obwohl extra ein Algorithmus [Autothrottle] zur Lasterkennung der abgefragten Server verwendet wird, führt das Abfragen natürlich zu ungewohnter Last auf Ihren Servern.
Da diese Untersuchung aber auf den Zeitraum einer Masterthesis begrenzt ist, besteht die Notwendigkeit, die URLs zeitnah zu crawlen.

Sollte Sie Anmerkungen oder Fragen haben, wenden Sie sich gerne an mich (Frederik Stey).

Vielen Dank für Ihre Unterstützung.

Impressum

Ansprechperson:
Hochschul- und Landesbibliothek Fulda
Frederik Stey
Heinrich-von-Bibra-Platz 12
36037 Fulda

Datenschutzerklärung

- Datum und Uhrzeit Ihres Zugriffs
- IP-Adresse in anonymisierter Form

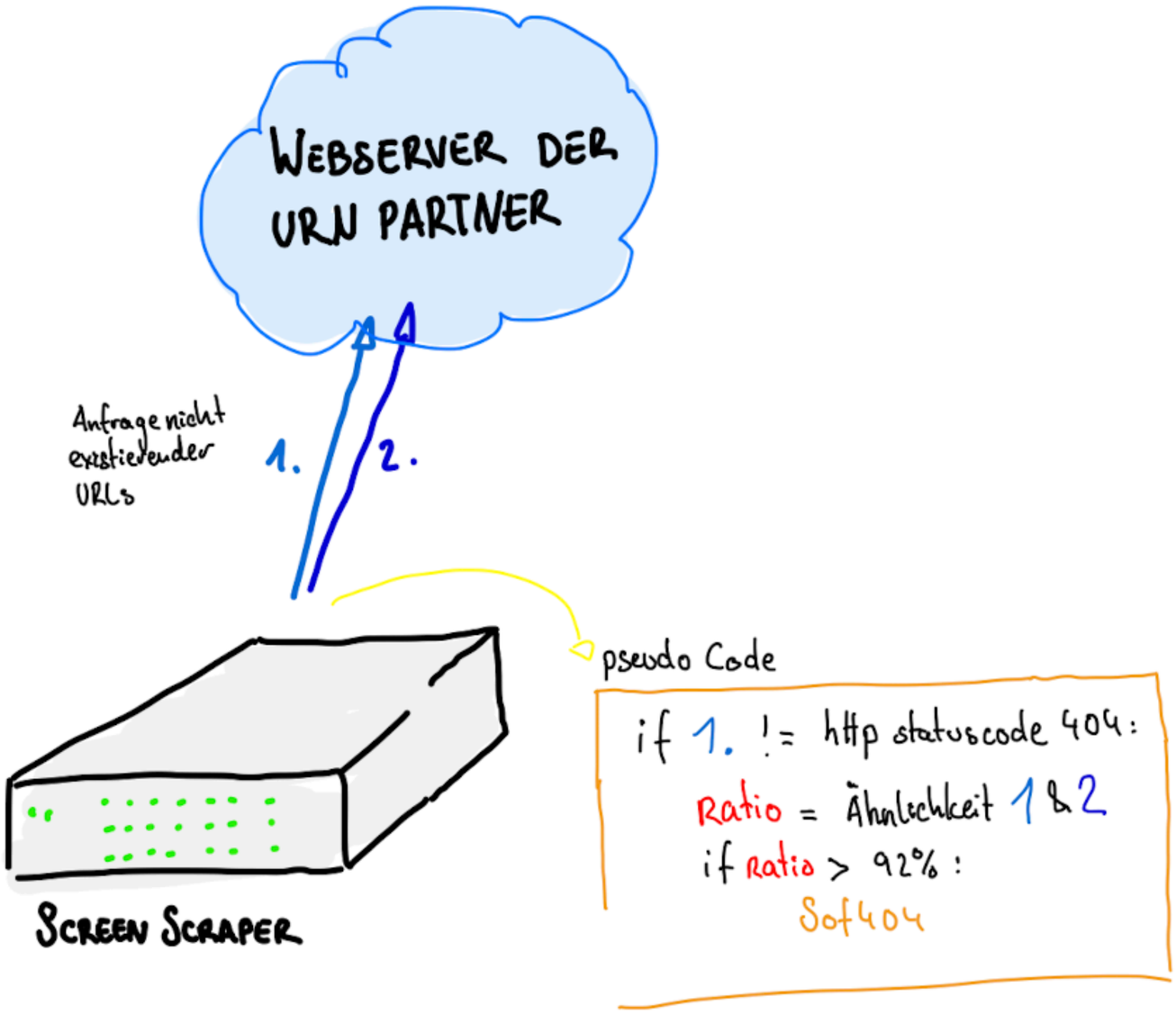
Diese Informationen werden nach Abschluss der Arbeit gelöscht und soll Auskunft darüber geben, ob der Einsatz des Crawlers aktive bemerkt wurde.
Es werden keine weiteren personenbezogenen Daten erhoben.

URN Link Rot Checker (HU Berlin / HLB Fulda / DNB - frederik.stey@hlb.hs-fulda.de +http://193.174.30.7

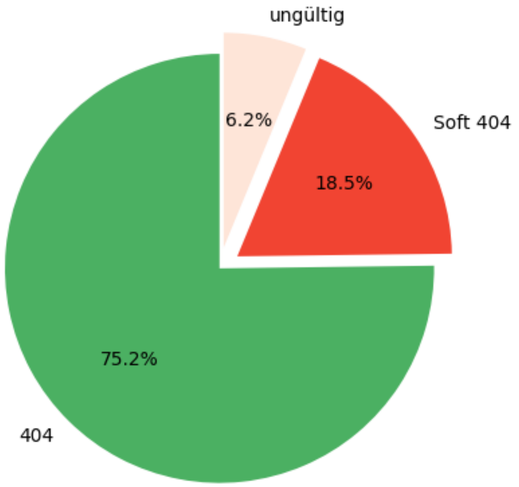
Soft 404


constructions-online.de
Domain erwerben
Sie können die Domain constructions-online.de für 195 EUR vom Inhaber kaufen.
Diese Webseite wurde vom Domain Inhaber dynamisch generiert, der das Sedo Domain Parking Programm nutzt. Die auf dieser Seite automatisiert bereitgestellten Werbeanzeigen kommen von dritter Seite und stehen mit Domain-Inhaber oder Sedo in keiner Beziehung.
Privacy Policy
Elements
Console
Preserve log
Filter
100 ms
200 ms
300 ms
1500 ms
1600
Name
resolver?urn=urnchecker
200
document
Other
4.7 kB
70 ms

SOFT 404

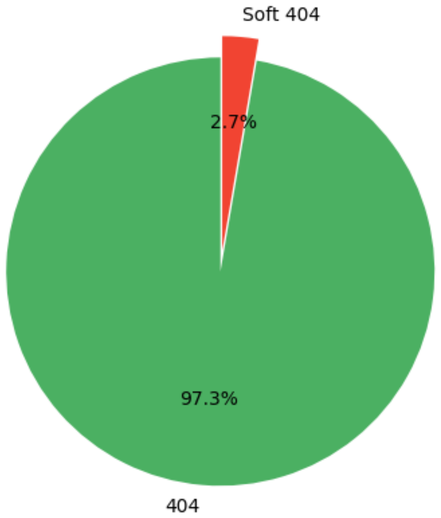

WEBSERVER DER URN PARTNER
Anfrage nicht existierender URLs
1.
2.
SCREEN SCRAPER
pseudo Code
if 1. != http statuscode 404:
Ratio = Ähnlichkeit 1 & 2
if Ratio > 92% :
Soft404

Resultate: Soft 404

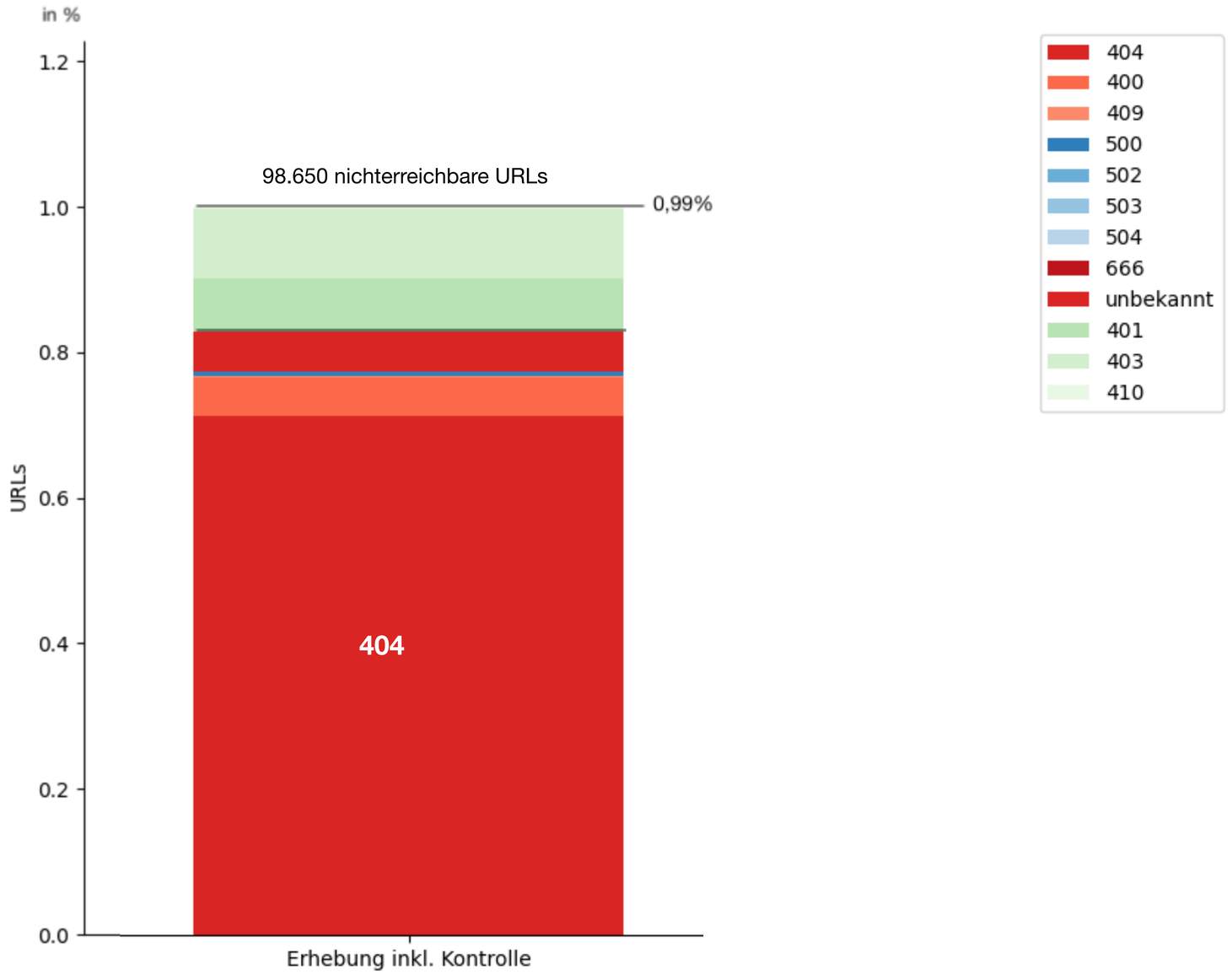


496 Domains auf Soft 404 getestet

nur bei 4 von 14 überprüfte Domains Soft 404 Fehler
40.471 Seiten geprüft

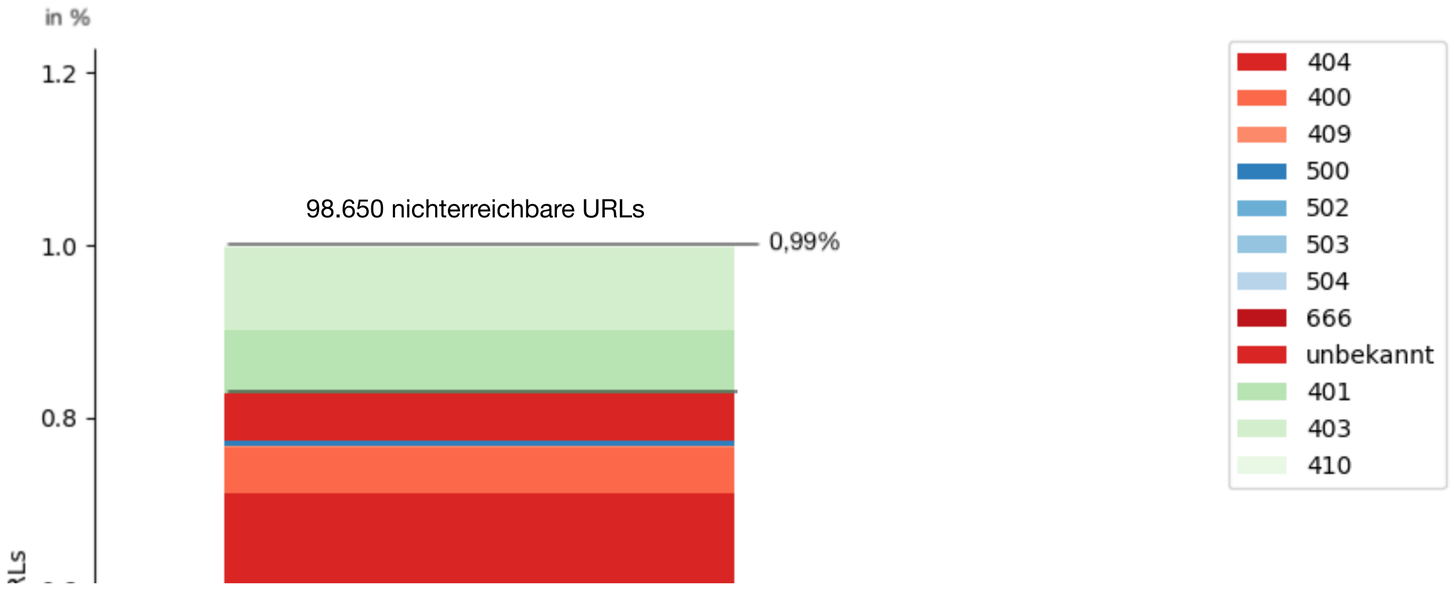


AUSWERTUNG



n = 10.000.000 URLs

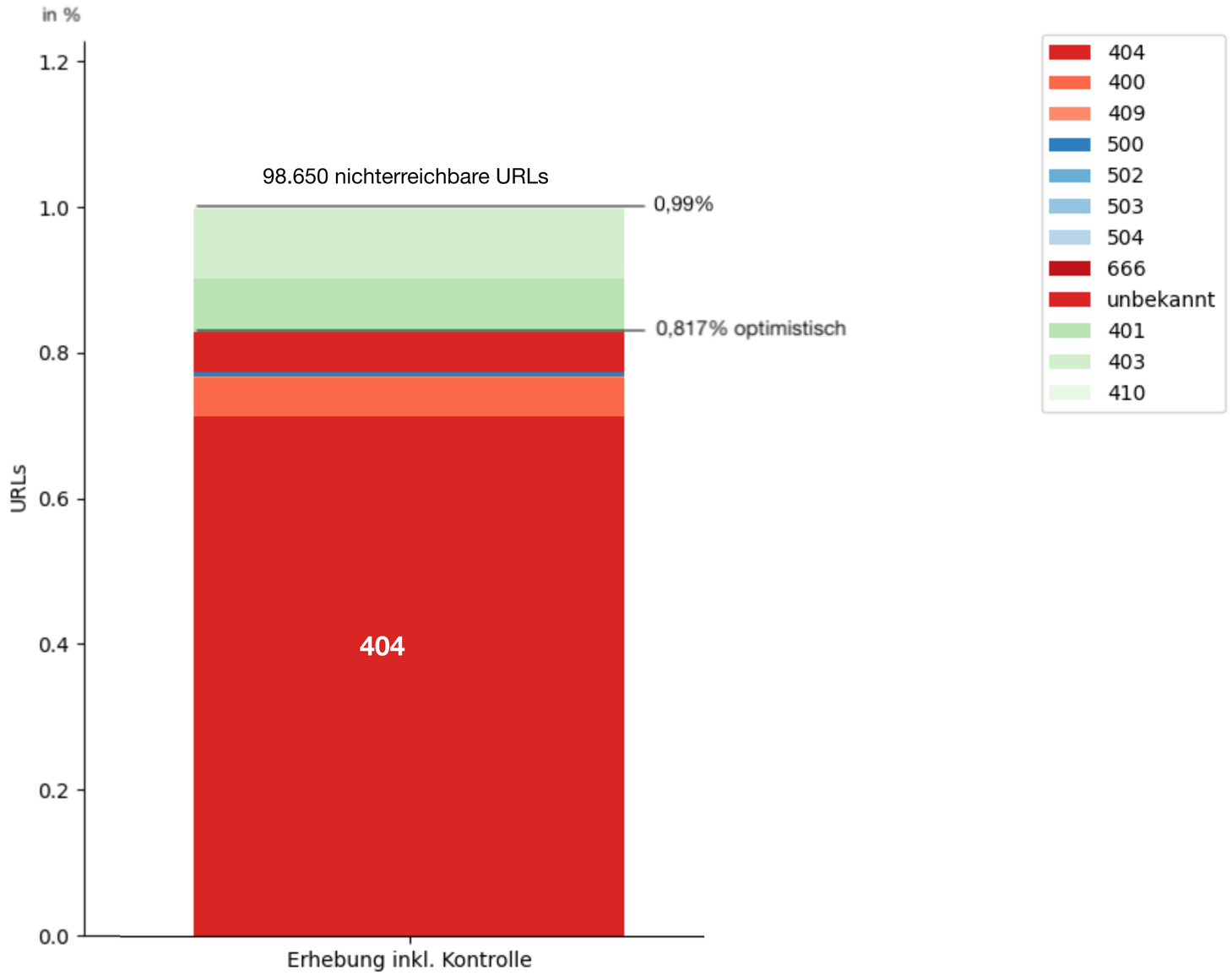
AUSWERTUNG



400 ***Bad Request*** Der Server kann die Anfrage nicht bearbeiten, da die Anfrage fehlerhaft war
401 ***Unauthorized*** Die Anfrage wurde ohne gültige Authenifizierungsdaten gestellt
403 ***Forbidden*** Die Anfrage wurde abgelehnt. Weitere Informationen sind in der Antwort enthalten
404 ***Not Found*** Die angeforderte Ressource ist nicht verfügbar beziehungsweise konnte nicht gefunden werden.(vgl. rfc7231, S.59).
409 ***Conflict*** Die Anfrage konnte nicht einfeutig einer Ressource zugewiesen werden.
410 ***Gone*** Die angeforderte Ressource wurde permanent entfernt.

n = 10.000.000 URLs

AUSWERTUNG

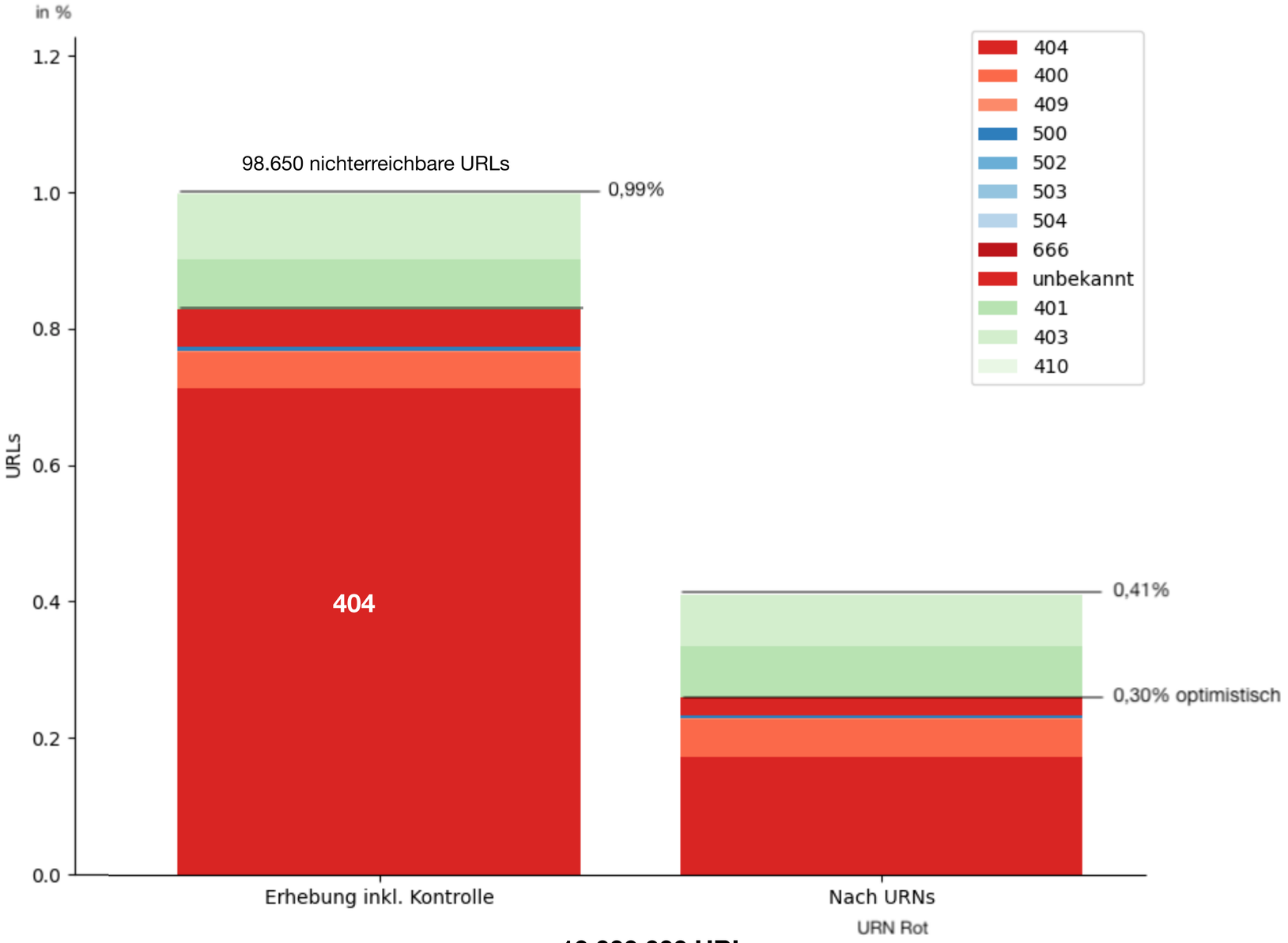


n = 10.000.000 URLs

URN Rot

Eine URN zählt nur als nicht erreichbar,
wenn alle dazugehörigen URLs nicht erreichbar sind.

AUSWERTUNG



n = 10.000.000 URLs

CODE & NACHNUTZBARKEIT

- Python 3
- Scrapy

https://scrapy.org

INSTITUTIONS URNs

DEUTSCHE NATIONAL BIBLIOTHEK

URN Service

```
urn:nbn:de:hebis:66-opus-405;http://fuldok.hs-fulda.de/opus4/files/21/DA_Bayer.pdf;1000021605
urn:nbn:de:hebis:66-opus-405;http://fuldok.hs-fulda.de/opus4/frontdoor/index/index/docId/21;1000021605
urn:nbn:de:hebis:66-opus-405;http://d-nb.info/1060786575/34;1000021605
urn:nbn:de:hebis:66-opus-412;http://d-nb.info/111273662X/34;1000021606
urn:nbn:de:hebis:66-opus-412;http://fuldok.hs-fulda.de/opus4/files/22/david_regionalmarken.pdf;1000021606
urn:nbn:de:hebis:66-opus-412;http://fuldok.hs-fulda.de/opus4/frontdoor/index/index/docId/22;1000021606
urn:nbn:de:hebis:66-opus-412;https://fuldok.hs-fulda.de/opus4/frontdoor/index/index/docId/22;1000021606
urn:nbn:de:hebis:66-opus-412;https://fuldok.hs-fulda.de/opus4/files/22/david_regionalmarken.pdf;1000021606
urn:nbn:de:hebis:66-opus-425;http://fuldok.hs-fulda.de/opus4/frontdoor/index/index/docId/23;1000021607
urn:nbn:de:hebis:66-opus-425;http://fuldok.hs-fulda.de/opus4/files/23/fabion_brutvogel_heidelstein.pdf;1000021607
urn:nbn:de:hebis:66-opus-425;http://d-nb.info/1060670704/34;1000021607
urn:nbn:de:hebis:66-opus-436;http://fuldok.hs-fulda.de/opus4/files/24/sauer_gvo.pdf;1000021608
urn:nbn:de:hebis:66-opus-436;http://d-nb.info/1060786583/34;1000021608
urn:nbn:de:hebis:66-opus-436;http://fuldok.hs-fulda.de/opus4/frontdoor/index/index/docId/24;1000021608
urn:nbn:de:hebis:66-opus-447;http://fuldok.hs-fulda.de/opus4/files/25/Schlussbericht_2004.pdf;1000021609
urn:nbn:de:hebis:66-opus-447;http://d-nb.info/1060670720/34;1000021609
urn:nbn:de:hebis:66-opus-447;http://fuldok.hs-fulda.de/opus4/frontdoor/index/index/docId/25;1000021609
urn:nbn:de:hebis:66-opus-457;http://fuldok.hs-fulda.de/opus4/files/26/Schlussbericht_2005.pdf;1000021610
urn:nbn:de:hebis:66-opus-457;http://d-nb.info/1060670755/34;1000021610
urn:nbn:de:hebis:66-opus-457;http://fuldok.hs-fulda.de/opus4/frontdoor/index/index/docId/26;1000021610
urn:nbn:de:hebis:66-opus-481;http://fuldok.hs-fulda.de/opus4/files/3/REK_rhoen_2007.pdf;1000021952
urn:nbn:de:hebis:66-opus-481;http://d-nb.info/1060670518/34;1000021952
urn:nbn:de:hebis:66-opus-481;http://fuldok.hs-fulda.de/opus4/frontdoor/index/index/docId/3;1000021952
urn:nbn:de:hebis:66-opus-463;http://fuldok.hs-fulda.de/opus4/files/27/Segetalflora_Dreienberg_Schlussbericht.pdf;1000021954
urn:nbn:de:hebis:66-opus-463;http://d-nb.info/1060670798/34;1000021954
urn:nbn:de:hebis:66-opus-463;http://fuldok.hs-fulda.de/opus4/frontdoor/index/index/docId/27;1000021954
urn:nbn:de:hebis:66-opus-478;http://fuldok.hs-fulda.de/opus4/files/28/weber_holz_als_energietraeger.pdf;1000021955
urn:nbn:de:hebis:66-opus-478;http://d-nb.info/1060786591/34;1000021955
urn:nbn:de:hebis:66-opus-478;http://fuldok.hs-fulda.de/opus4/frontdoor/index/index/docId/28;1000021955
urn:nbn:de:hebis:66-opus-496;http://fuldok.hs-fulda.de/opus4/files/29/Steinkopf_Schlussbericht.pdf;1000021956
urn:nbn:de:hebis:66-opus-496;http://d-nb.info/1060670801/34;1000021956
urn:nbn:de:hebis:66-opus-496;http://fuldok.hs-fulda.de/opus4/frontdoor/index/index/docId/29;1000021956
urn:nbn:de:hebis:66-opus-504;http://fuldok.hs-fulda.de/opus4/files/30/drinnenberg.pdf;1000024385
urn:nbn:de:hebis:66-opus-504;http://d-nb.info/1060786605/34;1000024385
urn:nbn:de:hebis:66-opus-504;http://fuldok.hs-fulda.de/opus4/frontdoor/index/index/docId/30;1000024385
urn:nbn:de:hebis:66-opus-546;http://fuldok.hs-fulda.de/opus4/files/32/Ideenskizze_Hess_Rhoen.pdf;1050039154
urn:nbn:de:hebis:66-opus-546;http://d-nb.info/1060670828/34;1050039154
urn:nbn:de:hebis:66-opus-546;http://fuldok.hs-fulda.de/opus4/frontdoor/index/index/docId/32;1050039154
urn:nbn:de:hebis:66-opus-555;http://fuldok.hs-fulda.de/opus4/files/33/Masterarbeit_Knorpp_2009_Salzrestriktion.pdf;1200048262
urn:nbn:de:hebis:66-opus-555;http://d-nb.info/1060786613/34;1200048262
urn:nbn:de:hebis:66-opus-555;http://fuldok.hs-fulda.de/opus4/frontdoor/index/index/docId/33;1200048262
urn:nbn:de:hebis:66-opus-560;http://fuldok.hs-fulda.de/opus4/files/34/2004_1_01.pdf;1200048263
urn:nbn:de:hebis:66-opus-560;http://d-nb.info/1060670836/34;1200048263
urn:nbn:de:hebis:66-opus-560;http://fuldok.hs-fulda.de/opus4/frontdoor/index/index/docId/34;1200048263
```

URN ;URL ;URN_ID

urn:nbn:de:hebis:66-opus-1009;http://fuldok.hs-fulda.de/opus4/files/85/2008_5_09.pdf;1200078916

SCRAPY METHODEN

start_requests()

In dieser Funktion werden die abzufragenden URLs zusammengetragen und einem sogenannten Generator übergeben. Dies ermöglicht dem Framework durch die URLs zu iterieren und entsprechend der getätigten Einstellungen abzuarbeiten.

parse()

Diese Funktion wird von dem Framework aufgerufen, wenn eine Antwort des Servers, mit dem Statuscode *200 ok* beantwortet wird. Hier kann ausprogrammiert werden, was mit den Daten passieren soll, oder wie strukturierte Daten aus der Antwort extrahiert werden.

error_parse()

Tritt ein Fehler bei der Abfrage auf, ruft das Framework diese Methode auf. Antwortet der Server mit einem HTTP Fehler Code (*4XX* oder *5XX*) oder bricht auf Grund eines anderen Fehlers ab, können diese in dieser Funktion behandelt werden. Da es sich um temporäre Fehler handeln kann, wiederholt Scrapy bei bestimmten Fehlercodes oder bei Zeitüberschreitungen nach Abarbeitung aller URLs die fehlgeschlagenen Anfragen erneut.

(vgl. *Scrapy Dokumentation*, Retry Middleware).

SCRAPY SETTINGS

https://docs.scrapy.org/en/latest/topics/settings.html

```
# Enable and configure the AutoThrottle extension (disabled by default)
# See https://docs.scrapy.org/en/latest/topics/autothrottle.html
AUTOTHROTTLE_ENABLED = True
# The initial download delay
AUTOTHROTTLE_START_DELAY = 5
# The maximum download delay to be set in case of high latencies
AUTOTHROTTLE_MAX_DELAY = 60
# The average number of requests Scrapy should be sending in parallel to
# each remote server
AUTOTHROTTLE_TARGET_CONCURRENCY = 1.0
# Enable showing throttling stats for every response received:
AUTOTHROTTLE_DEBUG = True
```

CODE

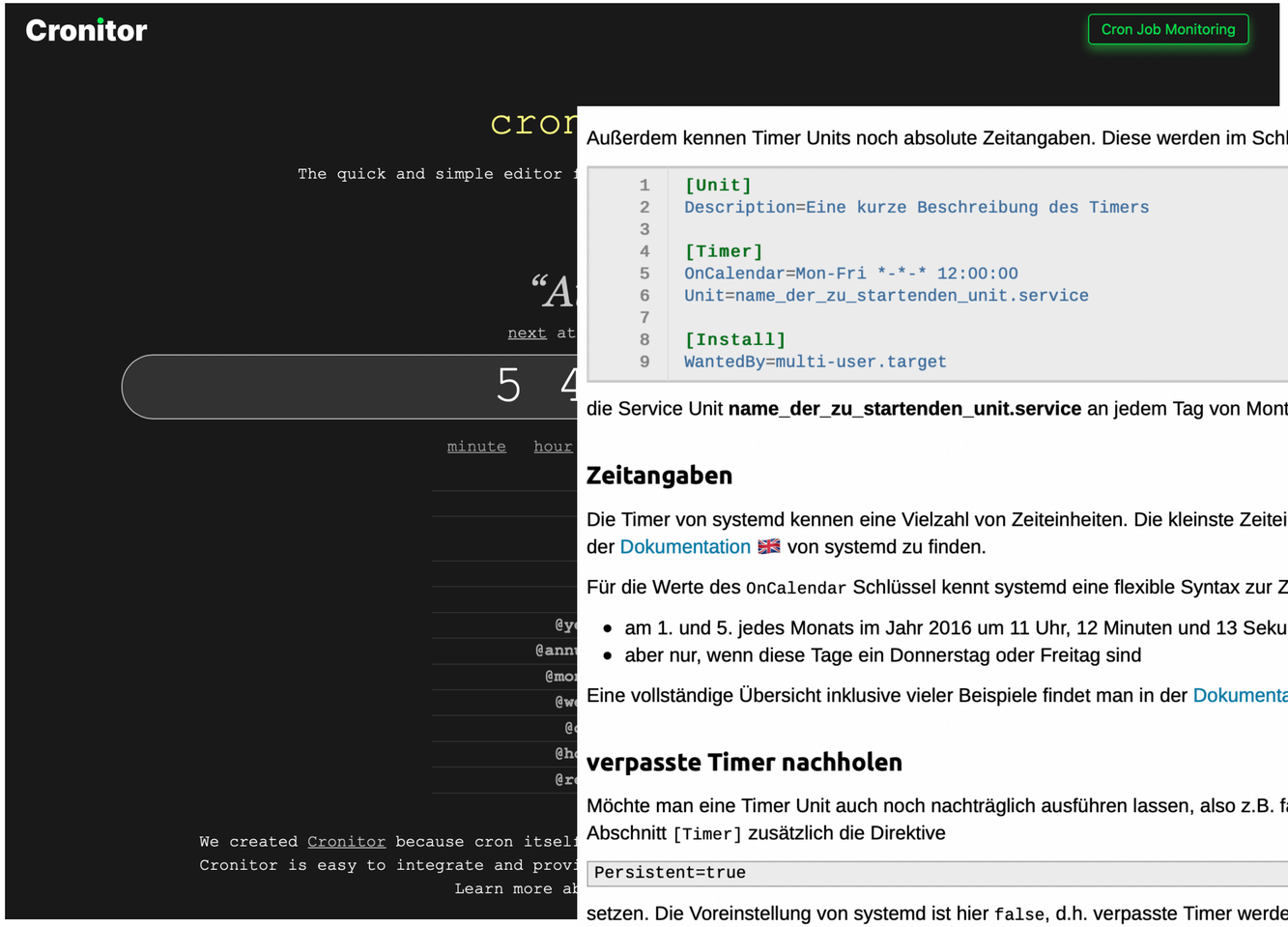
Der Code liegt in überarbeiteter Version unter https://github.com/fstey/urnChecker

```python
import scrapy
from scrapy.spidermiddlewares.httperror import HttpError

class BlogSpider(scrapy.Spider):
    name = 'linkchecker'

    def start_requests(self):
        filename = "input_fuldok.txt"
        # URN / URL Liste der DNB
        # URN;URL;URN_ID
        # urn:nbn:de:hebis:66-opus-1009;http://d-nb.info/1060671646/34;1200078916

        with open(filename, "r") as fd:
            lines = fd.read().splitlines()
            for line in lines:
                if line.startswith('urn:'): # for preventing header
                    urn, url, urn_id = line.split(';')

                    yield scrapy.Request(url=url, cb_kwargs=dict(urn=urn),
                        callback=self.parse, errback=self.error_parse, dont_filter=True)

    # 200 Status
    def parse(self, response, urn):
        url=response.url
        # check for soft404
        #print(urn, url, '_____________________> ',response.status)

    def error_parse(self, failure):
        urn = failure.request.cb_kwargs['urn']
        url = failure.request.url
        if failure.check(HttpError):
            response = failure.value.response
            print(urn, url,"HttpError", response.status)
        elif failure.check(DNSLookupError):
            print(urn+';'+url+';DNSLookupError')

        elif failure.check(TimeoutError, TCPTimedOutError):
            print(urn+';'+url+';DNSLookupError')

```

SOFT404

```
audit = difflib.SequenceMatcher(None,nonExistingWebsiteCode, WebsiteCode,autojunk=False)
```

CRON & systemd.timer



https://crontab.guru

Außerdem kennen Timer Units noch absolute Zeitangaben. Diese werden im Schlüssel `OnCalendar` angeben. So würde z.B. die folgende Timer Unit:

```
[Unit]
Description=Eine kurze Beschreibung des Timers

[Timer]
OnCalendar=Mon-Fri *-*-* 12:00:00
Unit=name_der_zu_startenden_unit.service

[Install]
WantedBy=multi-user.target
```

die Service Unit **name_der_zu_startenden_unit.service** an jedem Tag von Montag bis Freitag um 12 Uhr ausführen.

Zeitangaben

Die Timer von systemd kennen eine Vielzahl von Zeiteinheiten. Die kleinste Zeiteinheit sind Mikrosekunden, die größte Jahre. Eine komplette Übersicht inklusive der Einheiten ist in der Dokumentation 🇬🇧 von systemd zu finden.

Für die Werte des `OnCalendar` Schlüssel kennt systemd eine flexible Syntax zur Zeitangabe. So steht z.B. der Wert `Thu,Fri 2016-*-1,5 11:12:13` für:

- am 1. und 5. jedes Monats im Jahr 2016 um 11 Uhr, 12 Minuten und 13 Sekunden ausführen
- aber nur, wenn diese Tage ein Donnerstag oder Freitag sind

Eine vollständige Übersicht inklusive vieler Beispiele findet man in der Dokumentation 🇬🇧 von systemd.

verpasste Timer nachholen

Möchte man eine Timer Unit auch noch nachträglich ausführen lassen, also z.B. falls der Rechner zum Zeitpunkt der Fälligkeit des Timers ausgeschaltet war, muss man im Abschnitt `[Timer]` zusätzlich die Direktive

```
Persistent=true
```

setzen. Die Voreinstellung von systemd ist hier `false`, d.h. verpasste Timer werden nicht nachgeholt. Die Direktive funktioniert nur in Kombination mit der `OnCalendar` Direktive, nicht mit anderen Direktiven für Zeitpunkte.

Timer Units aktivieren

Nach dem Erstellen müssen Timer Units noch aktiviert und gestartet werden[1]:

```
sudo systemctl enable name_des_timers.timer
sudo systemctl start name_des_timers.timer
```


https://wiki.ubuntuusers.de/systemd/Timer_Units/

FAZIT

Wie groß ist also das Problem der Nichterreichbarkeit von Ressourcen im URN Namensraum der DNB wirklich?

Wie Umfangreich ist der Aufwand für eine Institution die Funktionalität der hinterlegten URLs sicherzustellen?

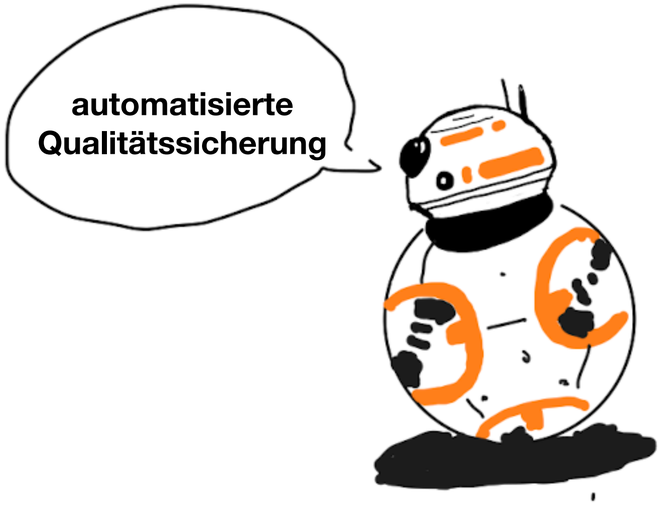
Korrektheit der Statuscodes ¯_(ツ)_/¯

Weiterleitungen ?

Soft-404 (╯°□°)╯ ⁀ ┻━┻

ok!

einfach



Vielen herzlichen Dank!

Graf, Klaus (2012a). *Defekte Links bei deutschen Hochschulschriftenservern*. URL: https://web.archive.org/web/20210709123604/https://archivalia.hypotheses.org/8844 (besucht am 09. 07. 2021).
– (2012b). *[InetBib] Defekte Links*. URL: https://web.archive.org/web/20210709121515/https://www.inetbib.de/listenarchiv/msg48263.html (besucht am 09. 07. 2021).
– (2012c). *Nibelungenlied-Links anno 2004*. URL: https://web.archive.org/web/20210709123348/https://archivalia.hypotheses.org/8847?s=netbib-wiki (besucht am 09. 07. 2021).
– (2014). *DNB inkompetent: Nach wie vor defekte Links*. URL: https://web.archive.org/web/20210709123342/https://archivalia.hypotheses.org/4575 (besucht am 09. 07. 2021).
– (2016a). *Permalinks sind der ThULB Jena wurscht*. URL: https://web.archive.org/web/20210128053109/https://archivalia.hypotheses.org/61126 (besucht am 09. 07. 2021).
– (2016b). *Schändet nun auch die SUB Hamburg Permalinks?* URL: https://web.archive.org/web/20210125121229/https://archivalia.hypotheses.org/61794 (besucht am 09. 07. 2021).
– (2019). *Digitale historische Tageszeitungen*. URL: https://web.archive.org/web/20210709124113/https://archivalia.hypotheses.org/100954 (besucht am 09. 07. 2021).