Photo by jmtosses CC BY-NC 2.0 https://www.flickr.com/photos/jmt/3838802631

Information Gathering and Retrieval

Dr. Anne Kayem

Hasso-Plattner-Institute
University of Potsdam

Gatherer

- Captures as many documents as possible and keeps this database as up-to-date as possible

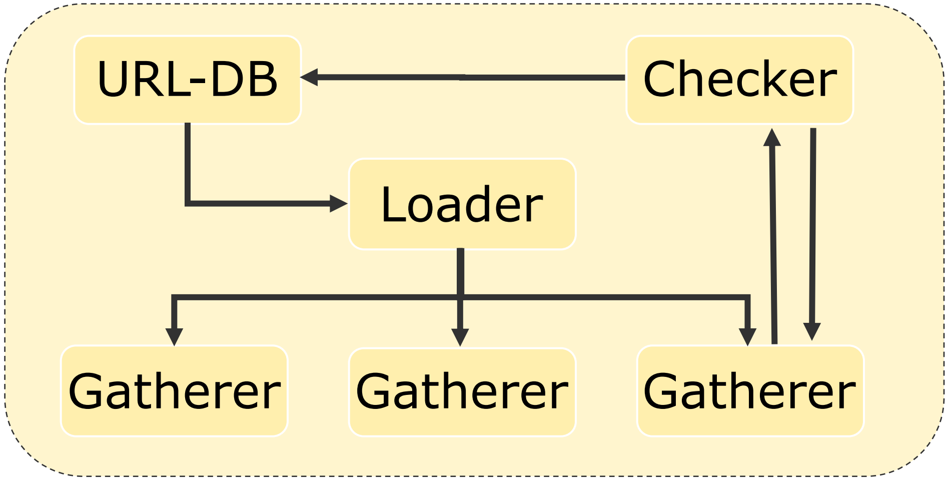
- **Operation**:
  - □ Uses HTTP GET-Request: Web

```
GET https://hpi.de/study/overview.html HTTP/1.1
User-Agent: Googlebot/2.1
```

- **Problem**:
  - □ Dynamic Resources, Islands, Dark Web
  - □ Web crawlers leave traces in the log file of the web server ...

Checker

- Normalises URLs
- Decides which documents from the Gatherer to send to the information retrieval system
- **Example:** Document types, Syntactic Correctness, Availability, …
- Eliminates duplicates

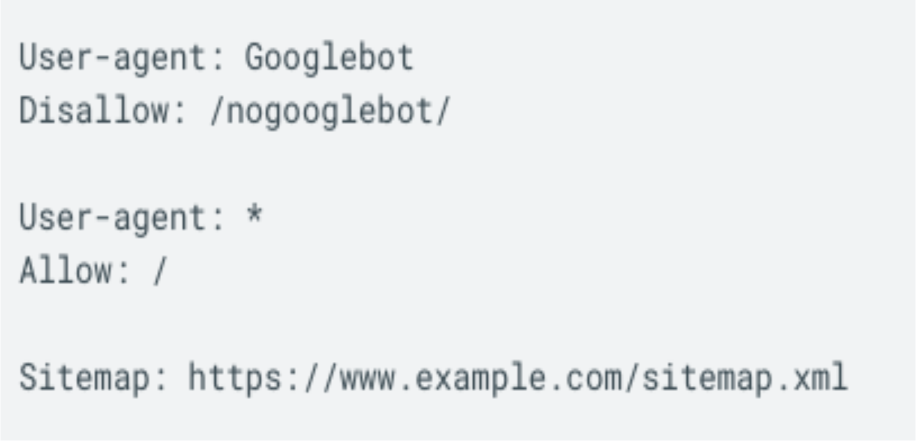


Further Tasks

- Checker

- Decides, for which links further searches should be done:

  - Robots.txt
  - Sitemap
  - Defective Links
  - SPAM Avoidance
  - Redirects
  - …



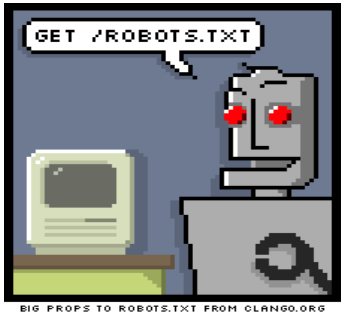
```
User-agent: Googlebot
Disallow: /nogooglebot/

User-agent: *
Allow: /

Sitemap: https://www.example.com/sitemap.xml
```

Controlling a Web Crawler

- HTML-Authors can control Web crawlers by using (incorporating) special Meta-Tags on their webpages

```
<META NAME="ROBOTS" CONTENT="NOINDEX, NOFOLLOW">
```

- Web-Server can control the Webcrawler using `/robots.txt`



```
User-agent: Googlebot
Disallow: /nogooglebot/

User-agent: *
Disallow: /
```

WebRobots Exclusion Protocol

- Robot-Netiquette recommends compliance with the Robot Exclusion Standard

- But! Not all web crawlers adhere to it

Sitemap Protocol

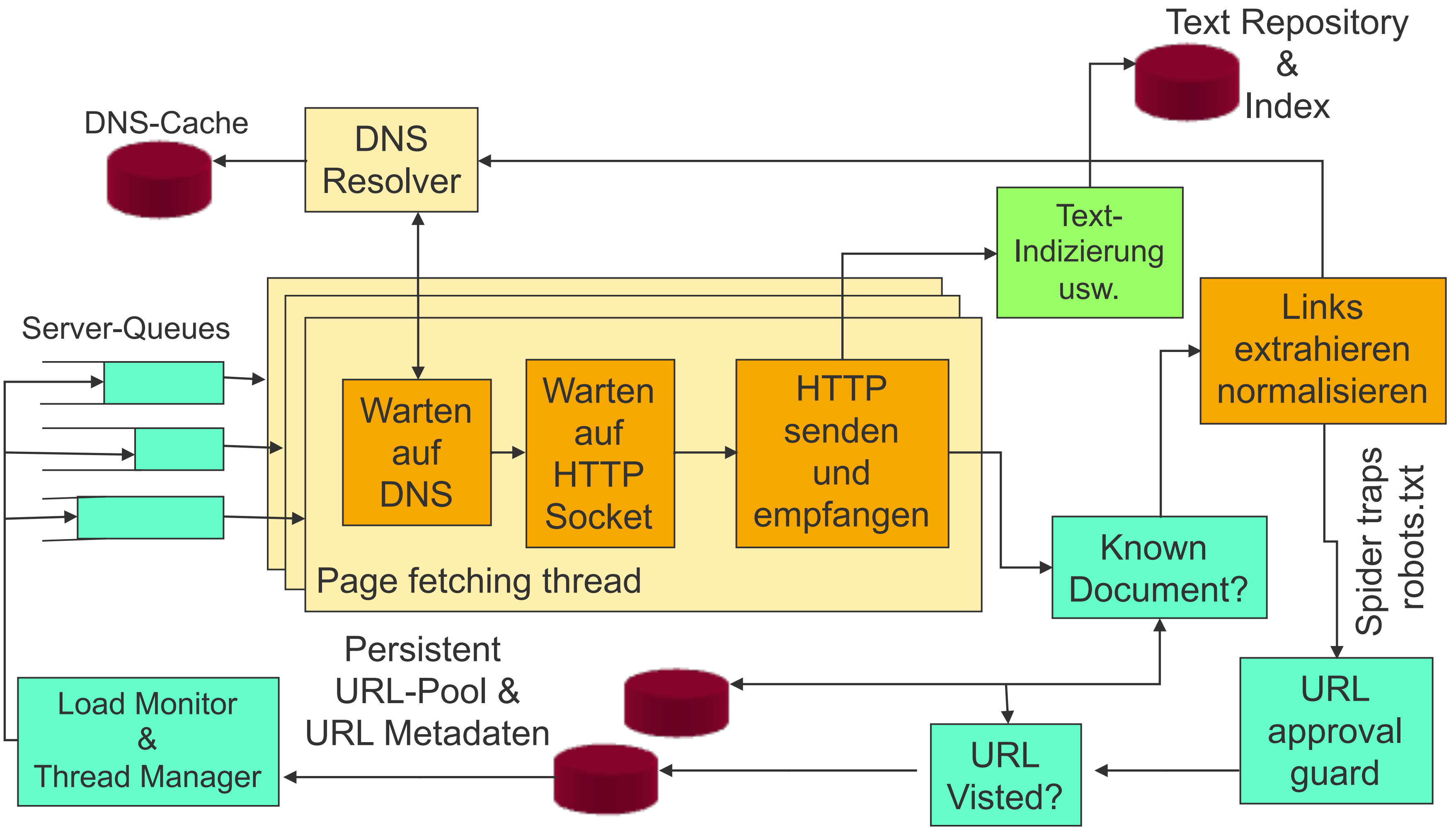
- Makes it easier for checkers to find important URLs
- Maps the linking structure to be searched

```
<?xml version="1.0" encoding="UTF-8"?>
<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9"
         xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
         xsi:schemaLocation="http://www.sitemaps.org/schemas/sitemap/0.9
                   http://www.sitemaps.org/schemas/sitemap/0.9/sitemap.xsd">
    <url>
        <loc>http://hpi.de/study/overview.html</loc>
        <changefreq>weekly</changefreq>
        <priority>1.00</priority>
    </url>
    <url>
         <loc>http://hpi.de/en/studies/overview.html</loc>
         <changefreq>weekly</changefreq>
         <priority>0.80</priority>
    </url>
 ...
</urlset>
```

Implementing a Web Crawler

- Requesting and transferring a WWW document is a time-consuming process
    - URL IP (DNS)
    - Establish TCP connection
    - Transfer data
    - Disconnect TCP connection
    - Detect duplicates
    - Extract hyperlinks from document
    - Extracting URLs from JavaScript files
- Parallelization of individual tasks

Web Crawler Architecture


Text Repository
&
Index
DNS-Cache
DNS
Resolver
Text-
Indizierung
usw.
Server-Queues
Links
extrahieren
normalisieren
Warten
auf
DNS
Warten
auf
HTTP
Socket
HTTP
senden
und
empfangen
Spider traps
robots.txt
Known
Document?
Page fetching thread
Persistent
URL-Pool &
URL Metadaten
URL
approval
guard
Load Monitor
&
Thread Manager
URL
Visted?