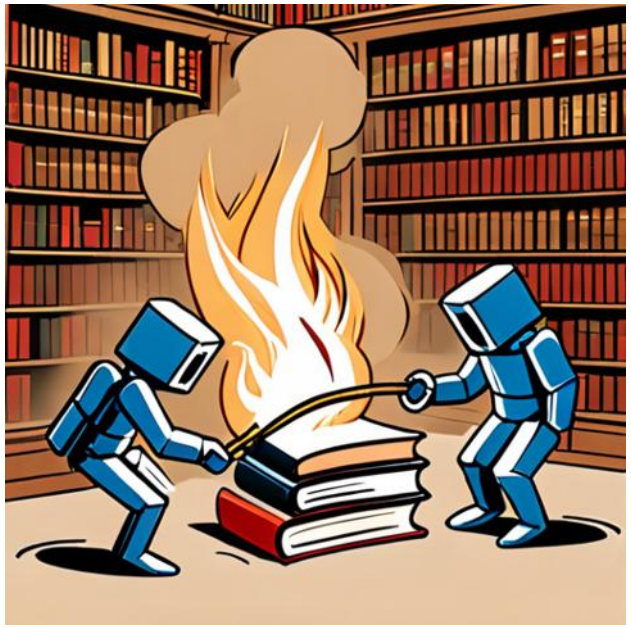
Maßnahmen zum Schutz der Mannheimer Digitalen Sammlungen vor aggressiven Bots



Kitodo Praxistreffen 2025 in Köln

Stefan Weil
Universitätsbibliothek Mannheim

Digitale Sammlungen – Schutzmaßnahmen für Kitodo-Server

- Wie viele Online-Angebote werden auch unsere Digitalen Sammlungen massiv von Webbots – insbesondere von Bots der bekannten KI-Anbieter – „heimgesucht“.

- Erste Maßnahmen waren Filter, die über den User-Agent bekannte „Bad Bots“ identifizieren und sperren. Vorbild dabei: Einstellungen von Wikimedia.

- Nach und nach wurden auch „gute Bots“ in die Filterliste aufgenommen.

- Geografische Filter ergänzt, sperren aktuell Argentinien, Brasilien und Ecuador.

- IP-basierte Filter ergänzt, sperren Alibaba.

  => **Trotzdem weiterhin Probleme.**

Problemanalyse und Verbesserungen

- Interaktive Analyse von Störungen (viele Apache-Prozesse, keine HTTP-Verbindung möglich, verstopfte MariaDB mit vielen sleeping connections) mit KI-Unterstützung für die Bewertung von Logdaten und Systemzustand
- Handlungsempfehlungen der KI ausprobiert
- Sitemap und robots.txt optimiert, werden von den namhaften Bots beachtet
- Fehlgeschlagene Einstellungen (Rate Limiting, Priorisierung) wieder entfernt
- Erfolgreiche Einstellungen auf allen Servern realisiert
  - Umstellung von Apache mpm-prefork + mod-php auf mpm-event + php-fpm
  - TYPO3 RealURL Schreibzugriffe in Datenbank blockiert
  - (Connection pooling für die Datenbankzugriffe, Wirksamkeit noch unklar)

Zustand nach Verbesserungen, weitere Maßnahmen

- Deutlich geringerer Ressourcenverbrauch (CPU-Last, Hauptspeicher).
- Gute Bots von Open AI, Anthropic, Google, Apple daher wieder zugelassen, denn diese sollen bei uns lernen.
- Zyklische Überwachung des Betriebszustandes und Neustart der relevanten Dienste (Apache, PHP-FPM) bei kritischem Zustand.
- Digitalisierungsserver liefert seitdem öfter mehr als eine Million Dateien am Tag aus, mit Spitzenwert bei 1,6 Millionen.
- Keine Störungen seit 9. November; davor wurden im November mehr als 400 Neustarts durch Überwachung protokolliert.

Denial-of-Service durch “friendly fire”

UB UNIVERSITÄTSBIBLIOTHEK MANNHEIM

- Insbesondere am 4. November und den folgenden Tagen gab es trotz aller Maßnahmen massive Betriebsstörungen.
- Suche nach häufigsten Abrufern findet Spitzenreiter:
  ```
  cat /var/log/apache2/access.log | sed 's/ .*//' | sort | uniq -c | sort -n | tail
  ```
- Namhafte wissenschaftliche Einrichtung ruft stundenlang mit zwei Servern Zeitungsdaten ab, oft die gleiche URL mehrfach (bis zu 49 x), mit sehr hohen Abrufraten (bis zu 30 Abrufe pro Sekunde).
- Maßnahme: Rücksprache mit dem Verursacher.

Zusammenfassung der Mannheimer Erkenntnisse

UB UNIVERSITÄTSBIBLIOTHEK MANNHEIM

- Überprüfung und Optimierung der eigenen Installation hilft viel.
  Da gibt es auch noch Verbesserungsmöglichkeiten im Kitodo-Code.
- Bei einigen (wenigen) besonders aggressiven Akteuren hilft nur Blockieren.
- Überwachung der Webdienste mit Neustart im Notfall ist unverzichtbar.
- Störungsursachen zeitnah überprüfen und darauf reagieren.
- Auf allgemeine Maßnahmen gegen Bots wie beispielsweise die Nutzung von Anubis verzichten wir, da wir selbst darunter leiden, wenn derartige Software woanders installiert ist.

Links

Referenzen

- Filtereinstellungen bei Wikimedia https://wikitech.wikimedia.org/wiki/Nova_Resource:Wikisource/Wikimedia_OCR
- Weil, S. (2024). Digitalisierungsaktivitäten und Kitodo an der UB Mannheim. Kitodo Praxistreffen 2024, Marburg

Abbildungsnachweis

- Die Abbildung auf der Titelseite wurde mit Stable Diffusion generiert.