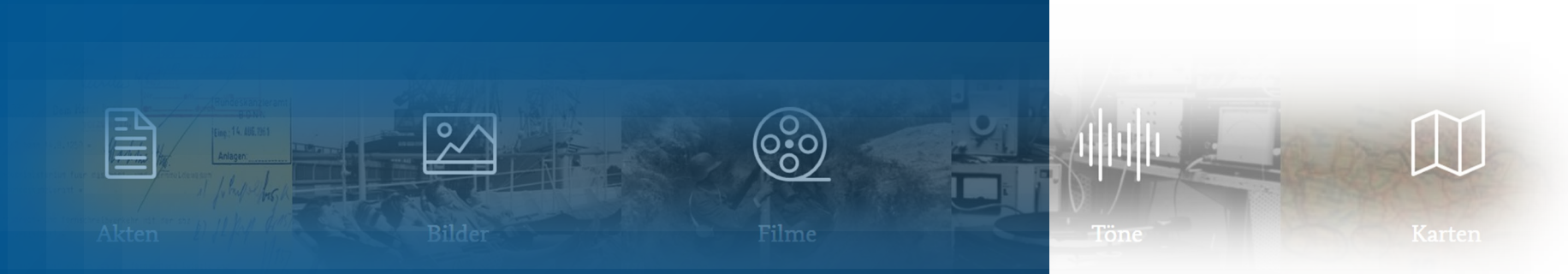
Perspektiven für eine semiautomatische Qualitätssicherung bei der Archivierung von Webseiten

Vortrag auf der
27. Tagung des Arbeitskreises **Archivierung von Unterlagen aus digitalen Systemen**
5.-6. März 2024, Zürich

Felix Lange, Referat B 4, Bundesarchiv, Deutschland



Agenda

- Einleitung
  - Webarchivierung im Bundesarchiv
  - QS von Webarchiven: State of the Art
- Hauptteil: QS von Webarchiven im Bundesarchiv
  - Zielstellung
  - Operationalisierung
  - Ergebnisse und Herausforderungen
- Zusammenfassung und Ausblick

Webarchivierung im Bundesarchiv

Grundlegendes

- Zeitraum: Webarchivierungen seit Ende 2022, Infrastruktur noch im Aufbau
- Erfassungsrahmen Stellen: Selektive Auswahl von Webseiten einzelner Stellen oder Projekte zu bestimmten - i.A. einmaligen - Sicherungszeitpunkten (bspw. bei Auflösung der Stelle)
- Erfassungsrahmen typologisch: Nicht nur klassische Webseiten (Mediensammlungen), sondern auch Webanwendungen
- Verzeichnung im Kontext der Archivierung (genuin) digitaler Unterlagen der abgebenden Stelle (Provenienzprinzip) - Kein separates Webarchiv im Bundesarchiv

Erhaltungsziele

- **Inhalt**: Erhaltung aller Seiten einer Site, inkl. des gesamten textuellen und audiovisuellen Inhalts
  - *Aber:* u.U. Begrenzung der gesicherten Medienarten (Ausschluss von Filmen u.a.)
- **Struktur:** Abbildung der Binnenreferenzen und des HTML
- **Funktionalität** Abbildung der urspr. Funktionalität und weitestgehend des dynamischen Verhaltens einer Site
  - *Aber:* Reines Webscraping (d.h. „Abkratzen“ einer Oberfläche), keine vollständige Sicherung oder Rekonstruktion serverbasierter Suchanwendungen

Relevanz der QS bei der Webarchivierung durch Archive

Besonderheit der Webarchivierung aus Sicht von Archiven: Auf Datenebene formiert das Archiv das Archivale (d.h. das SIP), nicht die Provenienzstelle. Entsprechend trägt das Archiv eine gesteigerte Verantwortung für die Authentizität des Archivales.

Qualitätssicherung von Webarchiven: State of the Art

Bisherige Praxis in Archiven (s. IIPC Web Arch. Conf. 2023)

- Studie von Reyes [2014, USA]: Vorranging händische QS mit eigenem Personal und halbstrukturierten Auswertungsbögen
- **Library of Congress** [USA, 2023]: 52 Prüfende (inhouse), Bewertungsmetrik auf Grund „halb-subjektiver“ Qualitätsurteile
- **National Archives** [Großbritannien, 2023]: Schritte in Richtung automatisierter QS mit Hilfe von Skripten zum Abgleich von Original und Archiv sowie zwischen Archiven verschiedener Softwarepakete

Theoretische Ansätze I

Definition von Gütekriterien

Das ARC-Framework der Library of Congress

- **Archivability –** Analyse der zu archivierenden Site, Definition realistischer Archivierungsziele (bspw. Backend Fachverfahren)
- **Relevance –** Vor und nach dem Abruf: Bestimmung des Erfassungsrahmens für eine konkrete Site
- **Correspondence –** Visual, interactional, Completeness

Theoretische Ansätze I

Das ARC-Framework der Library of Congress

- **Operationalisierung**: Bewertung durch semistrukturierte Fragebögen
- **Kritik**: Theorie „User-centered“. Archivierung nicht für Internetnutzer, sondern als Grundlage künftiger Auswertungen. Vollständigkeits- und Authentizitätsanspruch. „User“ nutzen ein Angebot entsprechend intendiertem Nutzungszweck.
  - ➢ Archive müssen Qualitätsmaße möglichst objektiv definieren und am Ideal des exakten Abbilds bzw. signifikanter Eigenschaften desselben ausrichten. Die „User Experience“ ist ein nachgelagertes Problem, es ist nicht der Maßstab der Archivierung per se.

Theoretische Ansätze II (technische Verfahren)

- **Zeitversatz:** Benn Sa u.a. (2011): Messung der Veränderung von Sites über den Zeitverlauf
- **Visuelle Übereinstimmung** (Kiesel et al. 2018): Analyse auf Ebene einzelner Bildpunkte

➢ QS für Webarchive ist ein aktives Forschungsthema in der Informationswissenschaft.

Grundsätzliche Problemstellungen

- **Aufgabe**: QS-Software muss Original aus dem Internet abrufen (1) und mit Webarchiv vergleichen (2)
  - (1) ist Aufgabe eines Crawlers!
    Gefahr: **QS nur so gut wie der Crawler der QS-Software**
  - Für (2) Fehlen eines objektiven Qualitätsmaßes; wie werden Abweichungen bemessen und bewertet?
- **Lösung**:
  - Verschiedene berechenbare Qualitätsmaße für verschiedene Erhaltungsziele
  - Verschränkung menschlicher und maschineller Prüfverfahren

QS: Operationalisierung

Ansätze

- Zu (1) Inhalt: Listenvergleich von Ressourcen aus dem Browser mit denen im Abruf-Protokoll (crawl.log).
- Zu (2) Struktur: Analyse der HTML-Objekt-Struktur der im Frontend ausgelieferten Seite durch Skript
- Zu (3) Funktionalität: Simulation des Frontends durch Skript, dass dynamische Elemente der Seite (Links etc.) ansteuert. Durch generische Softwarewerkzeuge zum Website-Testing

QS-Operationalisierung: Ressourcen

Dateivergleich durch Auslesen der Logs

```
sha1:6PWZ63BKBVBJJPRP23RQIDDRUL6XZEWV - -
2023-12-21T21:13:01.928Z   200     845877 https://www.coronawarn.app/assets/js/cwaa.js - - text/javascript #001 20
2023-12-21T21:13:05.196Z   200     234351 https://www.coronawarn.app/assets/js/app.js - - text/javascript #001 202
2023-12-21T21:13:08.333Z   200      49089 https://www.coronawarn.app/de/ R https://www.coronawarn.app/ text/html #
```

crawl.log

```
{
  "startedDateTime": "2024-02-27T07:42:51.383+01:00",
  "request": {
    "bodySize": 0,
    "method": "GET",
    "url": "https://www.coronawarn.app/assets/js/app.js",
    "httpVersion": "HTTP/2",
    "headers": [
      {
        "name": "Host",
        "value": "www.coronawarn.app"
      },
      {
        "name": "User-Agent",
        "value": "Mozilla/5.0 (Windows NT 10.0; Win64; x64;
      },
      . . .
    "response": {
      "status": 200,
      "statusText": "OK",
      "httpVersion": "HTTP/2",
      "headers": [
        {
```

Browser HAR

Operationalisierung : DOM-Struktur

Abgleich der Strukturen im Document Object Model

- Umsetzung:
  - Python-Skript und Bibliothek Selenium zum Laden der Daten
  - Spezial-Bibliotheken zur Berechnung numerischer Ähnlichkeitswerte und Ausgabe der Differenzen (lxml.html.diff, html_similarity)

Konkrete Unterschiede (Ausgabe in Datei)

<ins>..</ins>
~~..~~

Numerische Ähnlichkeitswerte

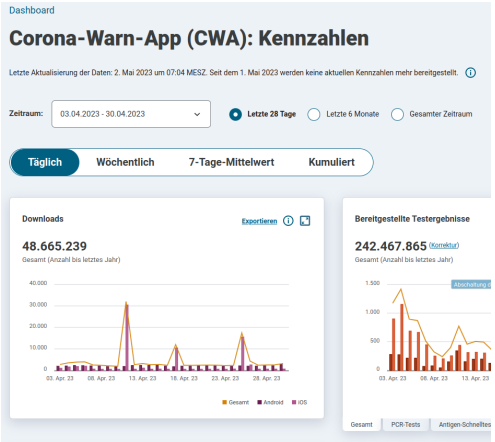
```
Kumulierter Ähnlichkeitswert: 0.9409589302769819
Ähnlichkeitswerte für Seite: http://localhost:8080/gkr_sept/20230925180603/https:[redacted]/gkr/ueber-uns/
Strukturelle Ähnlichkeit: 0.946058091286307
Stilistische Ähnlichkeit: 0.9436619718309859
```

QS-Operationalisierung: Anzeige und dynamisches Verhalten

Simulation des Nutzerverhaltens durch Website-Testsoftware

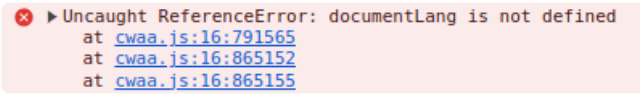
- Umsetzung:
  - Bibliothek Selenium
  - Ansteuern interaktiver Elemente, Prüfung des Ergebnisses

Original



Archiv

Fehler (Archiv)



QS-Operationalisierung : Ergebnis und Herausforderungen

- Ergebnis: Prüfprotokoll mit Fehler- und Erfolgsmeldungen sowie numerischen Gütewerten und gefundenen Fehlern
- Bleibende Herausforderungen:
  - Schwierigkeit, Abweichungen vom Original nachvollziehbar zu gewichten
  - Vermeidung und ggf. Dokumentation von Fehlern bei der Skriptausführung

Zusammenfassung und Perspektiven

Thesen zum Abschluss

- Automatisierung der QS schafft Skalierbarkeit, Zuverlässigkeit und Objektivität.
- Vollständige Automatisierung ist ein unrealistisches Ziel.
- Größte Herausforderung ist dynamisches Verhalten. Dies betrifft allgemeine Webtechnologien (clientseitige Skripte – Javascript). Nur eine **gemeinsame Anstrengung** kann die Produktion geeigneter Testumgebungen ermöglichen.

Präsentation:

Titel:	Perspektiven für eine semiautomatische Qualitätssicherung bei der Archivierung von Webseiten
vorgetragen von:	Lange, Felix
vorgetragen am:	05.03.2024

Kontaktdaten:

Ansprechpartner/-in:	Lange, Felix
Telefon:	+49 30 187770 8951
Email:	b4@bundesarchiv.de
Anschrift:	Bundesarchiv Referat B 4 Potsdamer Straße 1 Koblenz

Literatur

- Kiesel, Johannes, Florian Kneist, Milad Alshomary, Benno Stein, Matthias Hagen, und Martin Potthast (2018). „Reproducible Web Corpora: Interactive Archiving with Automatic Quality Assessment“. *Journal of Data and Information Quality* 10, Nr. 4 (31. Dezember 2018): 1–25. https://doi.org/10.1145/3239574.
- Feissali, Kourosh, und Jake Bickford (2023). „Open Auto QA at UK Government Web Archive“. Gehalten auf der IIPC General Assembly and Web Archiving Conference, Hilversum (NL), 10. Mai2023. https://digital.library.unt.edu/ark:/67531/metadc2143893/m2/1/high_res_d/IIPC_WAC2023-JAKE_BICKFORD_KOUROSH_FEISSALI-The_.pdf.
- Reyes Ayala, Brenda (2023). „Correspondence as the Primary Measure of Information Quality for Web Archives: A Human-Centered Grounded Theory Study“. *International Journal on Digital Libraries* 23, Nr. 1 (1. März 2022): 19–31. https://doi.org/10.1007/s00799-021-00314-x.
- Dies., Mark Edward Phillips, und Lauren Ko (2023). „Current Quality Assurance Practices in Web Archiving“. UNT Scholarly Works. University of North Texas, 19. August 2014. https://digital.library.unt.edu/ark:/67531/metadc333026/m2/.

Literatur

- Saad, Myriam Ben, und Stéphane Gançarski (2011). „Improving the Quality of Web Archives through the Importance of Changes“. In *Database and Expert Systems Applications*, herausgegeben von Abdelkader Hameurlain, Stephen W. Liddle, Klaus-Dieter Schewe, und Xiaofang Zhou, 6860:394–409. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011. https://doi.org/10.1007/978-3-642-23088-2_29 .