SCHÖNE NEUE NACHRICHTENWELT ARCHIVIERUNG VON NACHRICHTENSEITEN AN DER DEUTSCHEN NATIONALBIBLIOTHEK

STAATLICHE AKADEMIE DER BILDENDEN KÜNSTE STUTTGART

Stephanie Kuch
s.kuch@abk-stuttgart.de
Nov. 2013

Masterstudiengang Konservierung Neuer Medien und Digitale Information

EINLEITUNG

Wollen wir heute wissen, welche Nachrichten vor vielen Jahren die Menschen bewegten, so gehen wir in ein Archiv oder eine Bibliothek. Dort finden wir gedruckte Zeitungen und Flugblätter, die den Menschen vor mehreren hundert Jahren von mehr oder weniger aktuellen, von trivialen und wissenschaftlichen Nachrichten berichteten. Durch handschriftliche Überlieferungen auf Papier und Pergament, besitzen wir auch Kenntnisse aus der Zeit vor dem Buchdruck und haben eine Vorstellung, wie das Leben der Menschen damals aussah. Nachrichten verbreiteten sich zu Zeiten als es weder Telekommunikation noch ein weitverzweigtes Pressesystem gab noch langsam.

Unsere heutige Nachrichtenübermittlung sieht dagegen völlig anders aus. Nachrichten und Gerüchte verbreiten sich mit Hilfe des Internets in minuten- oder sogar sekundenschnelle weltweit über alle Kontinente hinweg. Doch selbst in dieser Zeit hat sich eine Form der Zeitung in der Welt des Internet etabliert - die Nachrichtenseiten. Die Thesis mit dem Titel „Schöne, Neue Nachrichten-Welt – Archivierung von Nachrichtenseiten an der Deutschen Nationalbibliothek“ soll die Schwierigkeiten und Herausforderungen der Archivierung von Nachrichtenseiten aufzeigen, sich mit ihnen auseinandersetzen und mögliche Lösungswege aufzeigen.

EIGENSCHAFTEN VON NACHRICHTENSEITEN

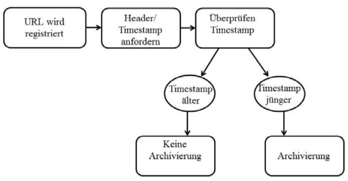
- häufige Aktualisierung (stündlich bis minütlich)
- große Vielfalt an verwendeten Formaten
- Inhalte werden durch paid content von der Archivierung ausgeschlossen bzw. deren Zugänglichkeit erschwert
  - 4 große Nachrichtenseiten in Deutschland verwenden paywalls
- große Datenmengen

ANFORDERUNGEN AN ARCHIVIERUNG

- häufige Archivierungszyklen, aus denen ein großer Bedarf an Ressourcen resultiert
- eine Standardisierung der Formate im Archiv ist schwer zu erreichen
- Lösung für den Umgang mit paid content muss gefunden werden
  - Inhalte hinter paywall werden ignoriert?
  - Zugang zu Inhalten nur über Kooperationen mit Nachrichtenseitenbetreiber
- Umgang mit den großen Datenmengen

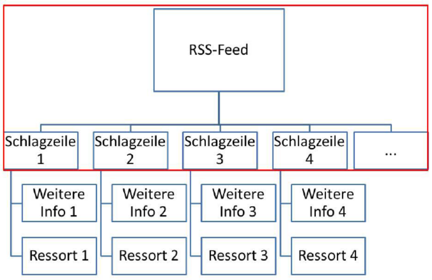
INKREMENTELL

- Inkrementell: bereits archivierte Inhalte werden von einer weiteren Archivierung ausgeschlossen
- Ziel: Einsparung von Ressourcen und Zeit
- Technik bisher noch nicht ausgereift
- bereits nach 24h ist die gecrawlte Datenmenge fast identisch
- es wird bereits nach Möglichkeiten gesucht, dieses Methode mittels eines Hash zu optimieren
- Hash: Prüfsumme einer Datei, an Hand derer festgestellt werden kann, ob sie verändert wurde
- auch eine Überprüfung des Timestamp wäre denkbar



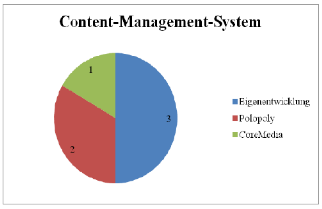
RSS-FEED

- Really Simple Syndication:
Ein RSS-Feed bietet den Nutzern einer Website die Möglichkeit, sich über neue Informationen auf einer Webseite zu informieren. Die Inhalte bzw. die Änderungen werden dabei in einer standardisierten, maschinenlesbaren Form dargestellt (meist xml-Format)
- denkbar wäre eine Archivierung des RSS-Feeds sowie aller verlinkten Artikel
- durchschnittliche Datenmenge bei Testspiegelungen: ca. 3 MB



DATEN-TRANSFER

- eine weitere Alternative stellt die direkte Übernahme der Daten vom Erzeuger, d.h. der Online Redaktion dar
- Problem: CMS verwenden interne Formate
- Lösung 1: CMS emulieren
  - sehr großer Aufwand, aber Layout bleibt erhalten
  - Ressourcen-intensiv
  - nur für kleine Menge an CMS geeignet



- Lösung 2: Übernahme in Standardformat
- mögliches Format NITF:
  - News Industry Text Format
  - entwickelt vom International Press Telecommunications Council
  - weit verbreitet
  - Open Source
  - Konsequenz: Layout-Informationen gehen verloren, nur Text bleibt erhalten

FAZIT

Da die Emulation der CMS auf Grund des hohen Anteils der Eigenentwicklungen einen zu großen Aufwand bedeuten würde und die inkrementelle Archivierung noch weiter entwickelt werden muss um praktikabel zu sein, wird die Archivierung des RSS-Feed mit den den dort verlinkten Artikeln momentan als die geeignetste Archivierungsmethode erachtet. Wird jedoch das Layout der jeweiligen Website nicht für signifikant gehalten, so wäre auch der Daten-Transfer praktikabel.