Tobias Steinke

Webarchivierung als internationale Aufgabe

DEUTSCHE NATIONAL BIBLIOTHEK

Inhaltsverzeichnis

Das Web als internationales Medium

- Web auf Basis des Internet seit 1989: International von Anfang an
- Kein nationales Web
  - Ursprung einer Webseite für Nutzer nicht ersichtlich
  - Nationale Top-Level-Domain (.de, .fr, .uk) nur Anhaltspunkt
  - Nicht-nationale Top-Level-Domains (.com, .net, .org)
- Webarchivierung
  - Zuerst auf internationaler Ebene vom Internet Archive
  - Erhalt von nationalen Kulturgütern als Aufgabe: Nationalbibliotheken für Webarchivierung verantwortlich

Webarchive weltweit (1)

- Internet Archive (archive.org)
  - US-amerikanische Non-Profit-Organisation (Fair Use)
  - Sammlung von „allen“ Webseiten weltweit seit 1996
  - Eigene Open-Source-Tools (Heritrix, Wayback Machine)
  - Über 450 Milliarden Webseiten, 20 Petabyte
  - Weltweiter Zugang per URL, keine Volltextsuche
  - Webarchivierungsdienstleistung (Archive-It)
- PANDORA (Australien)
  - Webarchiv von Nationalbibliothek und State Libraries seit 1996
  - Selektive Sammlung mit Rechteinholung für Zugang
  - HTTrack und eigenes Archiv- und Verwaltungssystem PANDAS

Webarchive weltweit (2)

- Nordic Web Archive
  - Projekt der Nationalbibliotheken von Dänemark, Norwegen, Schweden, Finnland und Island (2000 – 2002)
  - Entstehung nationaler Webarchive: Sammlung nationaler Top-Level-Domains
- UK Web Archive
  - Webarchiv der British Library, NL of Scotland, National Archives, Wellcome Library, JISC (seit 2003)
  - Selektive Crawls mit Rechteeinholung für öffentliche Bereitstellung
  - Seit 2013 gesetzliche Sammelauftrag für .uk-Crawl: Zugang nur in Lesesälen

Internationale Zusammenarbeit: IIPC

- International Internet Preservation Consortium: Technische und inhaltliche Zusammenarbeit (seit 2003)
- Derzeit 50 Organisationen weltweit (Bibliotheken, Archive, Forschungseinrichtungen, Firmen)
- Jährliche Konferenz, Arbeitsgruppen (Harvesting, Access, Preservation), Projektförderung, Trainingsevents, gemeinsame Crawls
- Tool-Weiterentwicklung: Heritrix, Open Wayback

Standards

- WARC (2009)
  - ISO-Standard für Webarchiv-Container
  - Aktuell ISO-Revision, IIPC-Gruppe zur Überarbeitung
  - Unterstützung in Tools und Services
  - Einsatz auch in digitaler Archivierung anderer Publikationen
- Statistiken und Qualitätskriterien (2013)
  - ISO-Bericht „Statistics and Quality Indicators for Web Archiving“
  - Einführung und Begriffe der Webarchivierung
  - Vergleichbare statistische Indikatoren und Qualitätsfaktoren

Memento

- Framework zur übergreifenden Nutzung von Webarchiven
- Entwickelt und getragen von Los Alamos National Lab.
- Protokoll zur Bekanntgabe der Webarchivinhalte
- Zugriff über timetravel.mementoweb.org
- Inhalte von Internet Archive, UK Web Archive, u. a.
- Time Travel Reconstruct: Angezeigte Webseite als Zusammensetzung aus verschiedenen Webarchiven

Ausblick: Weltweites Webarchiv?

- Web mit internationalem, vernetztem Charakter braucht weltweit vernetzte Webarchive
- Unterschiedliche rechtliche Gegebenheiten
- Technische Herausforderungen und kontinuierliche Änderungen im Web: Scheinbare Redundanzen in Webarchiven ergänzen sich
- Selektive Archivierung: Zusammenarbeit bei Auswahl und Sammlungserstellung