1912 100 JAHRE 2012 DEUTSCHE NATIONAL BIBLIOTHEK

Tobias Steinke

Herausforderungen bei der Langzeitverfügbarkeit von Webarchiven

1912 100 JAHRE 2012 DEUTSCHE NATIONAL BIBLIOTHEK

Inhaltsverzeichnis

1. **Digitale Langzeitarchivierung**
2. **Problematik bei Webarchiven**
3. **Metadaten zur Langzeitarchivierung**
4. **Standards und Tools**
5. **Archivierungsformat WARC**
6. **IIPC Preservation Working Group**

Digitale Langzeitarchivierung

- Problem 1: Erhalt des Datenstroms
  - Lösung: Redundante Datenhaltung, Umkopieren
- Problem 2: Langzeitverfügbarkeit
  - Änderung der Hard- und Softwareumgebungen
  - Dateiformate werden obsolet
  - Interpretierbarkeit des Datenstroms erhalten
  - Lösungen: Migration und Emulation

Lösungsstrategien zur Langzeitverfügbarkeit

- Migration
  - Dateiformatkonvertierung für aktuell nutzbare Umgebung
  - Aufwand skaliert mit Anzahl von betroffenen Objekten
  - Risiko von Verlust bei Inhalten
- Emulation
  - Nachstellen von früheren Systemumgebungen mit Software
  - Aufwand für Softwareentwicklung pro Umgebung
  - Risiko von geändertem Systemverhalten

Langzeitarchivierung bei Webarchiven

- Webstandards: HTML, CSS, Javascript, JPEG, GIF, etc.
- Browser als systemunabhängige Zugriffsumgebung
- Plugins für Inhaltsobjekte (Flash, PDF, 3D, Video, etc.)
- Hohe Verknüpfung durch Links führt zu Abhängigkeiten
- Webstandards ändern sich (HTML 4, XHTML, PNG, etc.)
- Seiten sind für die Darstellung mit bestimmten Browsern optimiert (Dynamic HTML, Videocodecs, etc.)

Langzeitverfügbarkeitsstrategien bei Webarchiven

- Migration
  - Hohe Verknüpfung: Links zu Konvertierungen berücksichtigen
  - Überschaubare Dateiformate für Großteil der Inhalte
- Emulation
  - Keine kompletten Systeme, nur Browserumgebungen
  - Standardumgebungen für zeitliche Abschnitte
  - Plugins können komplexere Systemanforderungen haben

Metadaten zur Langzeitarchivierung

- Gezielte Migration oder geeignete Emulation starten
- Technische Metadaten: Genaues Dateiformat (z. B. PDF 1.4), passendes Darstellungsprogramm (z. B. Internet Explorer 4.0), genutzte Hardware (etwa bestimmter Scanner), Farbtiefe, Codec, Komprimierung, etc.
- Metadaten zur Protokollierung von Änderungen
- Signifikante Eigenschaften: Welche Merkmale müssen erhalten bleiben?

Standards und Tools

- PREMIS: Langzeitarchivierungsmetdaten für digitale Objekte
- METS: Containerformat für Metadaten und Objektstruktur
- OAIS: Referenzmodell für digitale Langzeitarchive
- JHOVE, DROID, FITS: Formatidentifikation, Validierung und Metadatengenerierung

Archivierungsformat WARC

- ISO-Standard: Web ARChive (WARC)
- Nachfolger von ARC (entwickelt von Internet Archive)
- Binärformat für Crawlerresultate
- Records für Dateien und Verwaltungsinformationen
  - Internet Application Layer Protocols (HTTP, DNS, FTP, etc.)
  - Metadaten
  - Resultate von Transformationen (Migrationen)

IIPC Preservation Working Group (PWG)

- International Internet Preservation Consortium
- Arbeitsgruppen für Harvesting, Access, Preservation
- PWG: Verschiedene Arbeitspakete zu Browserabhängigkeiten, Charakterisierungs-Tools, Informationspaketen, etc.
- WARC Tools zur Unterstützung von Archiven
- JHOVE Modul für WARC (BnF)
- Formatidentifikation:
  http://netpreserve.org/about/Poster_ipres2010_webarchivefileformats_oury.pdf