deutsches literatur archiv marbach

Netzliteratur in Archiven: Von der technischen Analyse zur Emulation

Workshop, 24.06.2014

deutsches literatur archiv marbach

Übersicht

- Projekt und -stand
  - Projektvorstellung
  - Probleme
  - Lösungsansätze
    - Methoden
    - Metadaten
    - Entwickelte Programme

- Das BagIt-Format im Projekt Netzliteratur

Projekt

- Aufbau eines Quellencorpus für die seit den 1990er Jahren entstehende Literaturgattung „Netzliteratur“
- Projektlaufzeit: 3 Jahre (01.2013 – 12.2015)
- Sammlung und Auswahl
- Rechteeinholung und Katalogisierung
- Sichtung, Analyse, lokale Testumgebung, SWBContent
- Archivierung
- Verfügbarmachung

Projekt

- Aufbau eines Quellencorpus für die seit den 1990er Jahren entstehende Literaturgattung „Netzliteratur“
- Projektlaufzeit: 3 Jahre (01.2013 – 12.2015)
- Sammlung und Auswahl
- Rechteeinholung und Katalogisierung
- Sichtung, Analyse, lokale Testumgebung, SWBContent
- Archivierung
- Authentische Verfügbarmachung

Werkbeschaffung

- Crawl
- Daten aus anderen Quellen, z.B. Kontakt mit Autoren
- Screencast

Probleme: Crawl

- Serverseitige Funktionen
  - http://www.dadadata.de/
- Deep Web-Komponenten
  - http://www.assoziations-blaster.de
- externe Datenquellen
  - http://searchlutz.netzliteratur.net/
- JavaScript
- Systemanforderungen
  - http://www.berkenheger.netzliteratur.net/ouargla/websprudel/

Lösungsansätze: Crawl

- Serverseitige Funktionen || Deep Web-Komponenten
  - Quellcode beschaffen, Seite tatsächlich betreiben
  - Momentaufnahme
  - Screencast

- externe Datenquellen
  - Screencast
  - Proxies

Lösungsansätze: Crawl

- JavaScript
  - Verbesserung der Crawler
- Systemanforderungen
  - Verbesserung der Crawler Crawler

Ergebnisse der Werkbeschaffung

- warc-Datei
- Videodatei
- Quellcode

Probleme: Wiedergabe warc

- Serverseite: Wayback Machine und Authentizität
- Clientseite: Browser, Plugins, Hardware, ...

deutsches literatur archiv marbach

Lösungsansatz: Wiedergabe warc

Auf Grundlage von warc-Dateien eine authentische Abspielumgebung emulieren

Lösungsansatz: Metadaten

- METS, MODS und PREMIS
- Beschreibung der Abspielumgebung: Client- und Serverseite

deutsches literatur archiv marbach

Entwickelte Programme

- MIME-Type-Parser für warc-Dateien
- warc-“Entpacker“

BagIt allgemein

hierarchische Verzeichnisstruktur

- Payload-Verzeichnis „data“
- Metadatendateien „bagit.txt“ (beschreibt die Bag)
- sowie „manifest-<alg>.txt“

BagIt allgemein

```
bag/
|
|-- data
|   \-- nyancat.jpg
|
|-- manifest-md5.txt
|     51afb385ha019f34b671a3f0a615fae1 data/nyancat.jpg
|
\-- bagit.txt
      BagIt-version: 0.97
      Tag-File-Character-Encoding: UTF-8
```

BagIt im Projekt Netzliteratur

hierarchische Verzeichnisstruktur

- Payload-Verzeichnis „data“ und
- Metadatendateien „bagit.txt“
- sowie „manifest-<alg>.txt“

- metadata.xml
- structMD.xml
- screenshot_00.jpg
- screenshot_00.tif

BagIt im Projekt Netzliteratur

```
bag/
|
|-- data
|   \-- nyancat.jpg
|
|-- manifest-md5.txt
|     51afb385ha019f34b671a3f0a615fae1 data/nyancat.jpg
|
|-- bagit.txt
|     BagIt-version: 0.97
|     Tag-File-Character-Encoding: UTF-8
|
|-- bag-info.txt
|
|-- metadata.xml
|
|-- structMD.xml
|
|-- screenshot_00.jpg
\-- screenshot_00.tif
```