Von Spidern und Baggern

Tools im Projekt Netzliteratur

Steffen Fritz

01.12.2015

deutsches literatur archiv marbach

Werke

Anzahl: 50
Zeitraum: 1995 - 2012
Typen: Hypertext, Flash, Präsentationen, ...

deutsches literatur archiv marbach

Workflow (vereinfacht)

1. Rechteeinholung
2. Katalogisierung
3. Wiki-Eintrag
4. Vorabanalyse/Sichtung Liveweb
5. Harvesting und Dokumentation
   - Online und crawlbar: Crawl
   - Online oder emulierbar: Screencast
   - Online oder emulierbar: Screenshot
   - Quellcode verfügbar: Quellcode
6. Wiki-Eintrag aktualisieren
7. Bagerstellung
8. Datenübetragung an BSZ
9. Verfügbarmachung BSZ

deutsches literatur archiv marbach

Workflow (vereinfacht)

1. Rechteeinholung
2. Katalogisierung
3. Wiki-Eintrag
4. **Vorabanalyse/Sichtung Liveweb**
5. **Archivierung und Dokumentation**
   - Online und crawlbar: Crawl
   - Online oder emulierbar: Screencast
   - Online oder emulierbar: Screenshot
   - Quellcode verfügbar: Quellcode
6. Wiki-Eintrag aktualisieren
7. **Bagerstellung**
8. Datenübertragung an BSZ
9. Verfügbarmachung BSZ

deutsches literatur archiv marbach

Ausgangslage

50 Werke

deutsches literatur archiv marbach

Ausgangslage

50 Werke mit mehr als 3400 verlinkten Dokumenten

deutsches literatur archiv marbach

Ausgangslage

50 Werke mit mehr als 3400 verlinkten Dokumenten

$\Rightarrow$ Automatisierung

deutsches literatur archiv marbach

Automatisierung Vorabanalyse

Grundlage:

- `http://archiveready.com`
  - Webdienst
  - Analysiert einzelne Dokumente
  - API verfügbar
  - Tests:
    - Accessibility
    - Standards Compliance
    - Cohesion
    - Metadata

deutsches literatur archiv marbach

Automatisierung Vorabanalyse

Grundlage:

- `http://archiveready.com`
  - Webdienst
  - Analysiert einzelne Dokumente
  - API verfügbar
  - Tests:
    - Accessibility
    - Standards Compliance
    - Cohesion
    - Metadata

Ermittlung URL Dokumente:

- wget

deutsches literatur archiv marbach

Spider

```
#!/usr/bin/env zsh

wget -nv -r --spider -i $1 2 > &1 |
egrep " URL: " |
awk '{ print $3 }' |
sed "s@URL:@@g" >> corpus_link_liste.txt
```

deutsches literatur archiv marbach

Eingabegenerierung für ArchiveReady

```
#!/usr/bin/env zsh

cat corpus_link_liste.txt|sort|uniq >
corpus_link_liste_sorted_uniq.txt
```

deutsches literatur archiv marbach

corpus_link_liste_sorted_uniq.txt

```
(...)
http://oliver-gassner.de/textratouren/n99/index.html
http://oliver-gassner.de/textratouren/n99/ly.html
http://oliver-gassner.de/textratouren/n99/qu.html
http://oliver-gassner.de/textratouren/n99/re.html
(...)
```

deutsches literatur archiv marbach

arrr.py

Funktionsweise:

- URL aus Linkliste einlesen
- an ArchiveReady übergeben
- JSON-Antwort parsen und aufbereiten
- Bericht zu jedem Werk erstellen

deutsches literatur archiv marbach

Beispielbericht Vorabanalyse

Beispielbericht:
`https://wwik-prod.dla-marbach.de/line/images/4/48/Noise99_beurteilung_spiegelungsfaehigkeit.pdf`

deutsches literatur archiv marbach

Methode

- Wahl der Harvesting- und Dokumentationsmethode entsprechend Vorabanalyse
  - Screenshot (obl)
  - Screencast (fak)
  - Crawl
  - Sourcecode

deutsches literatur archiv marbach

Screenshot und Screencast

- Screenshot: Photoshop. Formate: jpg, tiff
- Screencast: Camtasia. mp4-Container (AVC/H.264, AAC)

deutsches literatur archiv marbach

Screenshot und Screencast

- Screenshot: Photoshop. Formate: jpg, tiff
- Screencast: Camtasia. mp4-Container (AVC/H.264, AAC)

- Liveweb
- Emulationsumgebung
  - Mac vMac
  - bwFLA

deutsches literatur archiv marbach

Crawl

- BSZ: SWBcontent (u.a. Heritrix und Wayback Machine)
- Download des Crawls (Heritrix Jobverzeichnis)

deutsches literatur archiv marbach

Sourcecode

Von den Autoren per Mail, ftp oder auf Datenträgern

deutsches literatur archiv marbach

zu archivierende Objekte

Screenshot: Bilddateien
Screencast: Videodateien
Crawls: Heritrix-Jobverzeichnis.tar.gz
Sourcecode: DATEINAME.tar.gz

deutsches literatur archiv marbach

Allgemein: BagIt

- Verzeichnisstruktur
- dient der zuverlässigen Speicherung und Übertragung digitaler Inhalte

deutsches literatur archiv marbach

Allgemein: BagIt

```
bag/
|
|-- data
|    \-- stimpack.dat
|
|-- manifest-md5.txt
|      51afb385ha019f34b67a615fae1  data/stimpack.dat
|
\-- bagit.txt
       BagIt-version: 0.97
       Tag-File-Character-Encoding: UTF-8
```

deutsches literatur archiv marbach

DLA: BagIt

```
bag/
|
|-- data
|    \-- screenshot_00.jpg
|    \-- screenshot_00.tif
|    \-- (...)
|
|-- manifest-sha1.txt
|      51afb385ha019f34b67ae1  data/screenshot_00.jpg
|      (...)
|
|-- metadata.xml
|
\-- bagit.txt
\-- bag-info.txt
```

deutsches literatur archiv marbach

metadata.xml

Beschreibung der Werke und Objekte mit METS, MODS und PREMIS[1]

deutsches literatur archiv marbach

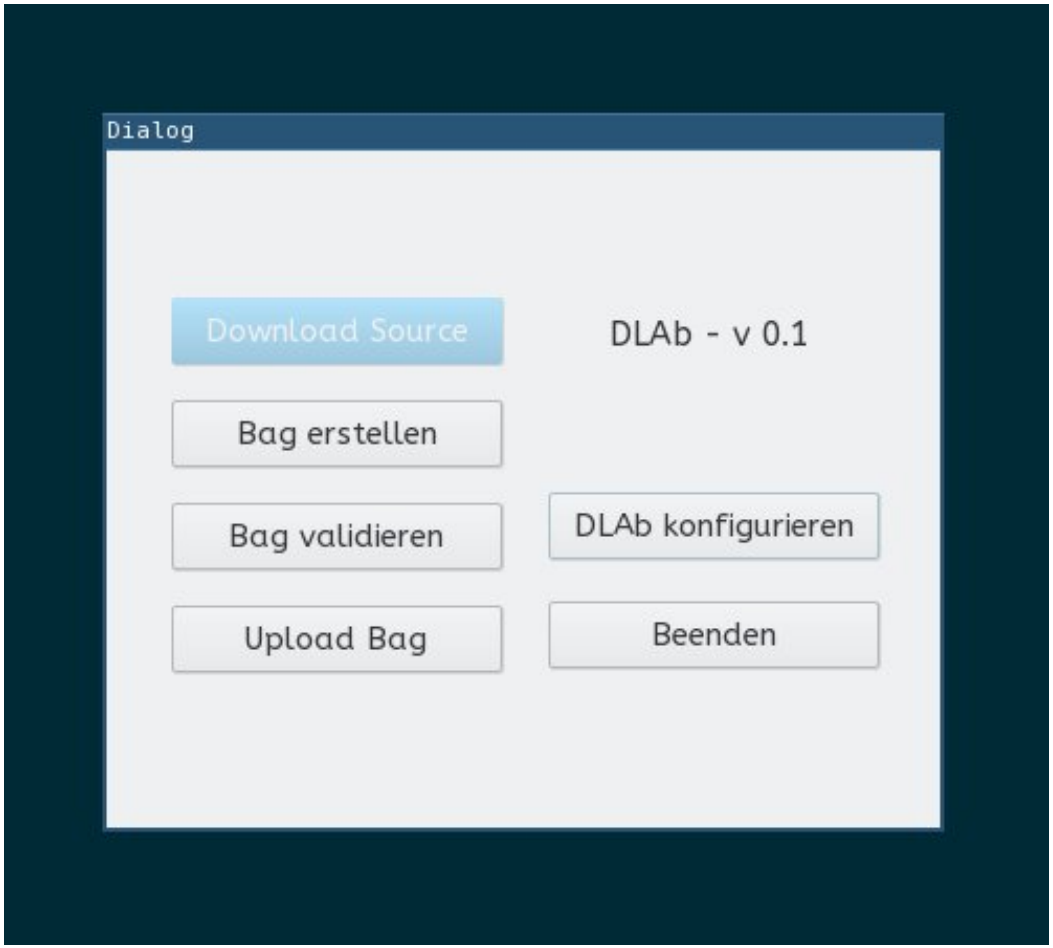
[1] 02.12.2015, 10 Uhr 15: Netzliteratur und Metadaten

DLAb

- Automatisierte Analyse der Payload (fido)
- Abruf bibliographische Daten (SRU-Schnittstelle SWB)
- Generierung der metadata.xml
- Erstellung Bag

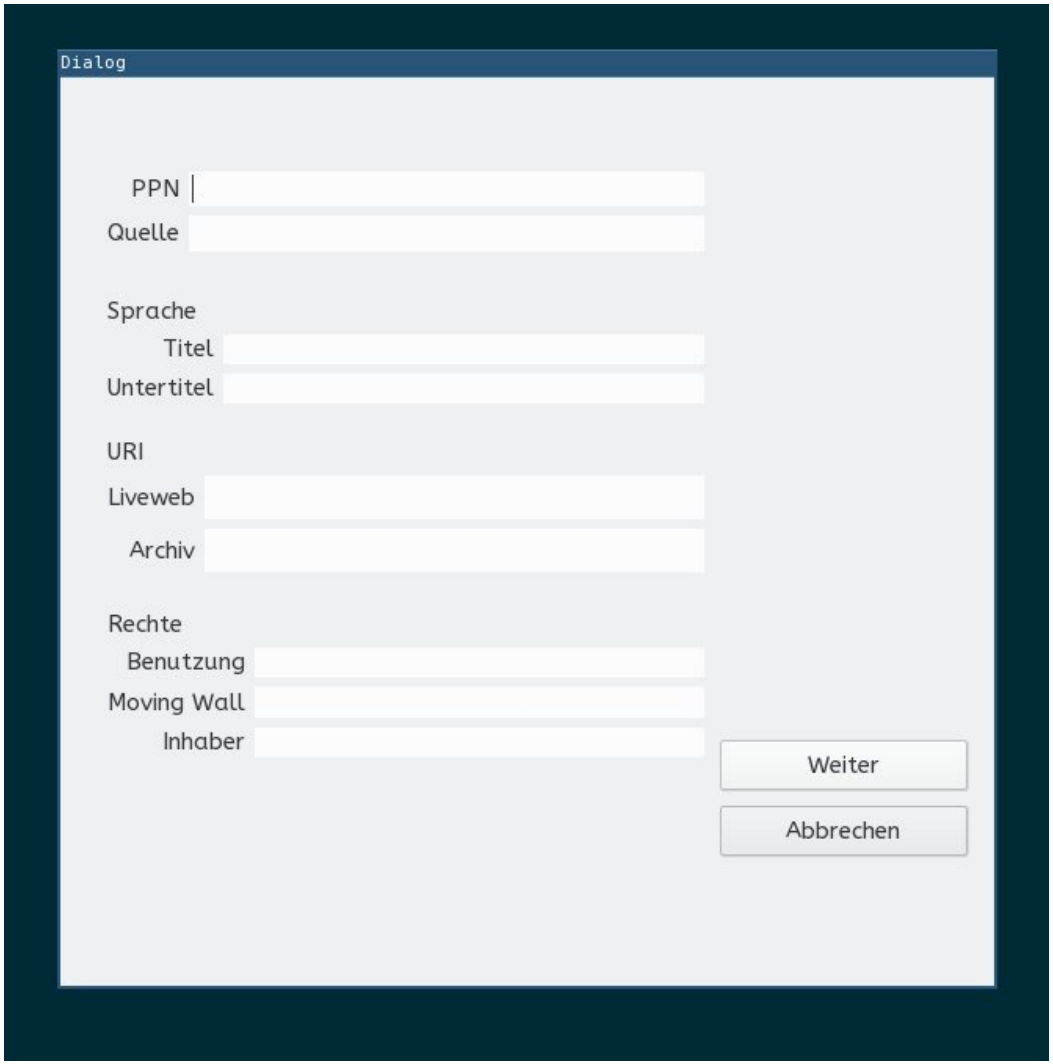
deutsches literatur archiv marbach

DLAb: Menü



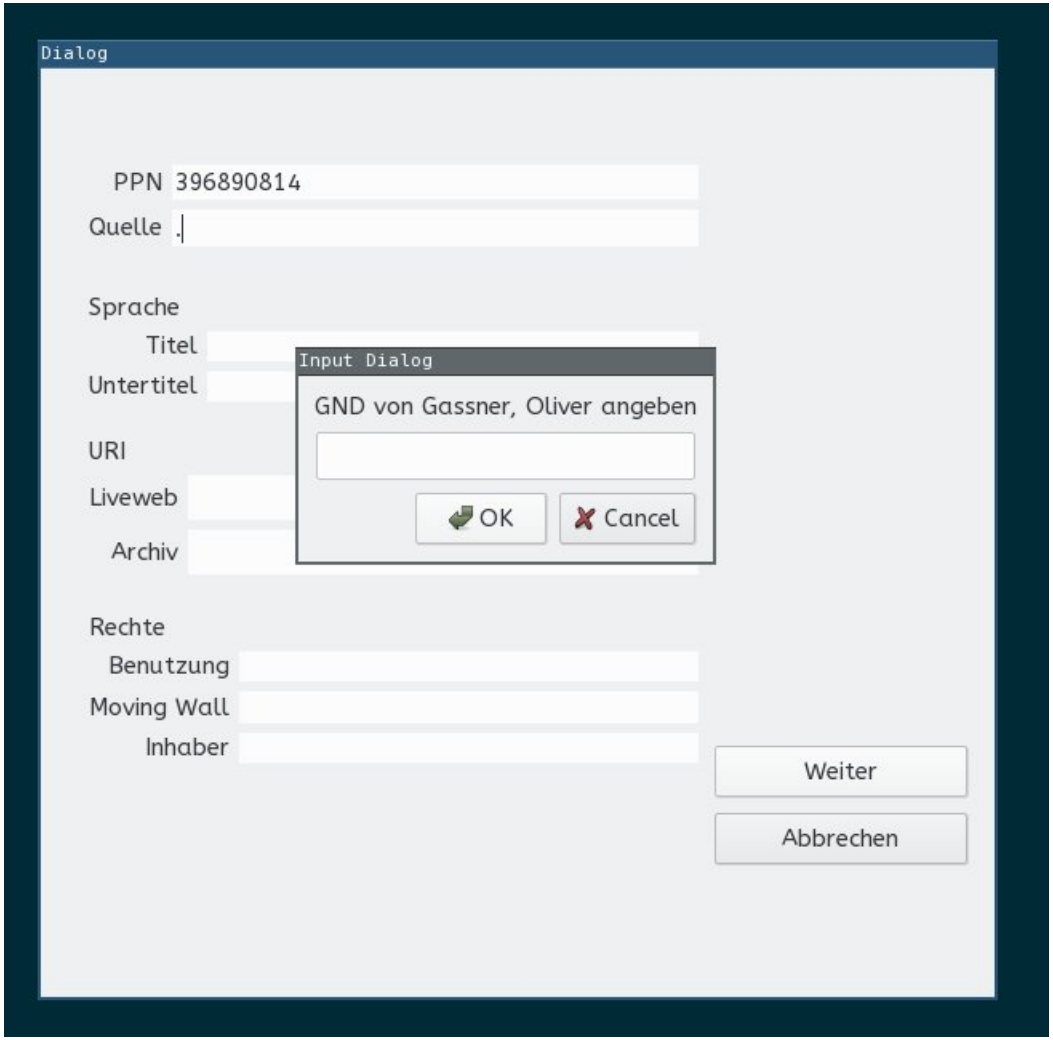
deutsches literatur archiv marbach

DLAb: Bibliographische Erfassung



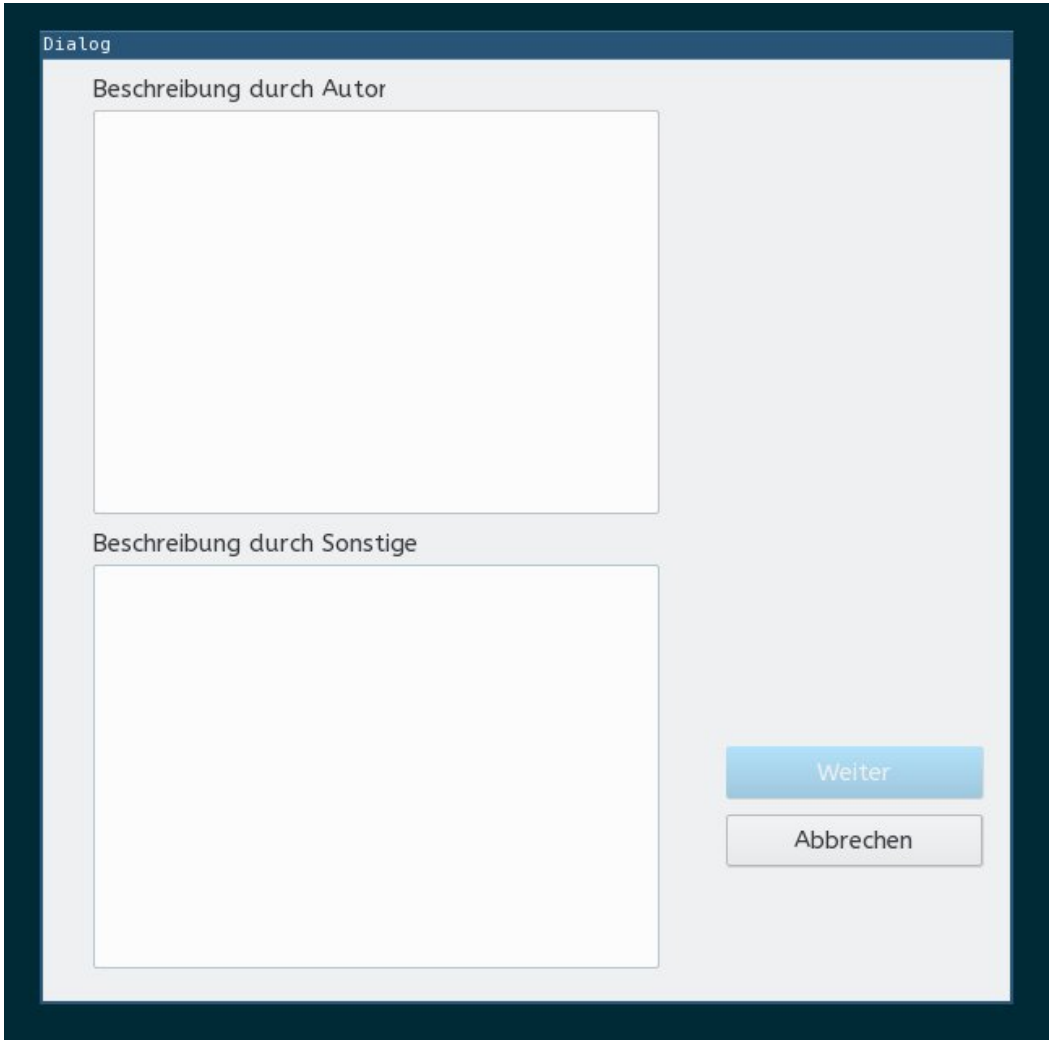
deutsches literatur archiv marbach

DLAb: GND



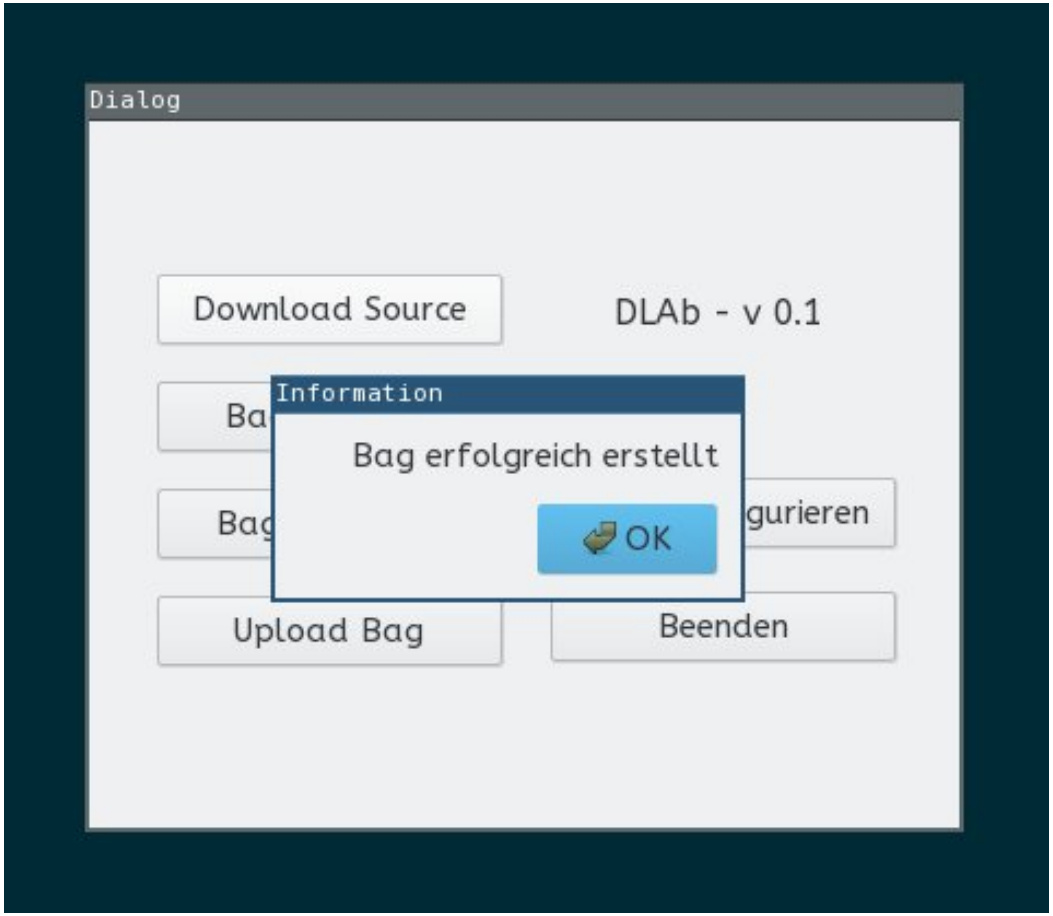
deutsches literatur archiv marbach

DLAb: Abstracts



deutsches literatur archiv marbach

DLAb: Bag erfolgreich erstellt



deutsches literatur archiv marbach

DLAb: Ergebnis

- BSZ96890814.bag
- BSZ96890814.bag.tar.gz
- BSZ96890814.bag.tar-gz-sha512.txt

deutsches literatur archiv marbach

BSZ: SWBcontent

- Upload Bag über Webinterface
- Darstellung: `http://literatur-im-netz.dla-marbach.de/bsz397988397.html`

deutsches literatur archiv marbach

html2warc

Generierung von WARC-Dateien aus Offlinequellen

deutsches literatur archiv marbach

Weiterführende Informationen

1. Workflow: `https://wwik-prod.dla-marbach.de/line/index.php/Projektpapiere`
2. ArchiveReady: `http://archiveready.com`
3. CLEAR: `https://doi.org/10.1007/s00799-015-0144-4`
4. wget: `https://www.gnu.org/software/wget/`
5. arrr.py: `https://github.com/netzliteratur/arrr`
6. bwFLA: `http://bw-fla.uni-freiburg.de/`
7. SWBcontent: `https://www.bsz-bw.de/mare/lza/swbcontent.html`
8. DLAb: `https://github.com/netzliteratur/dlab`
9. fido: `https://github.com/openpreserve/fido`
10. html2warc: `https://github.com/netzliteratur/html2warc`

Fragen?

fritz@dla-marbach.de

deutsches literatur archiv marbach