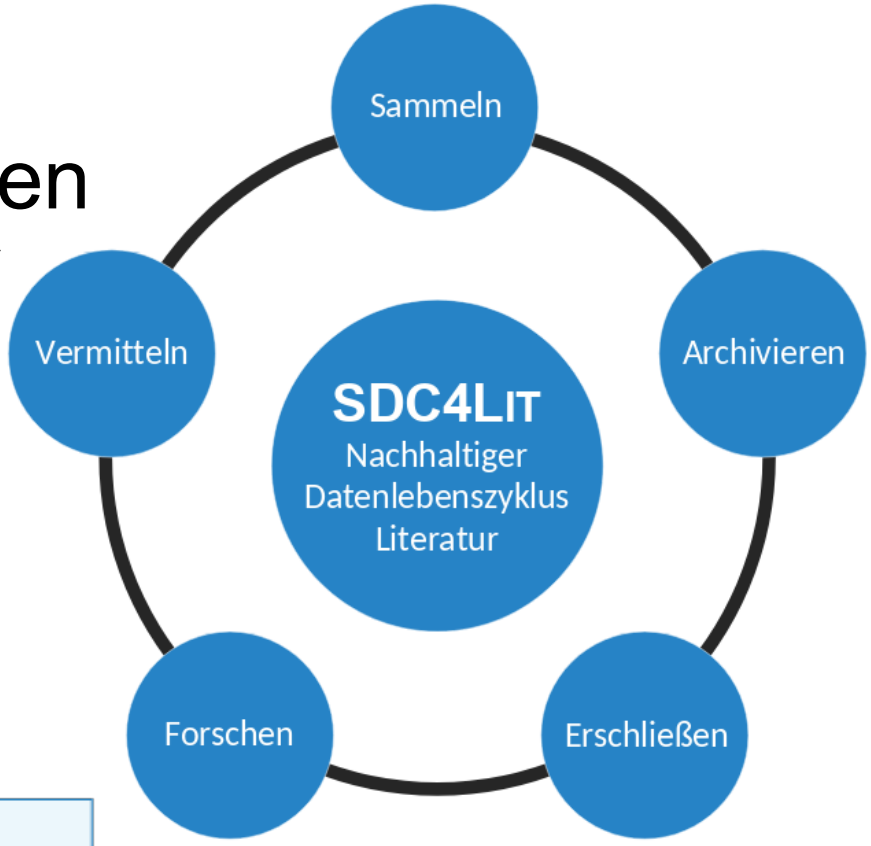
Digitale Objekte in der Webarchivierung aus textanalytischer Perspektive

Claus-Michael Schlesinger, Mona Ulrich, André Blessing
Science Data Center for Literature
Deutsches Literaturarchiv Marbach
Universität Stuttgart

Literatur im Netz archivieren und erforschen

Deutsches Literaturarchiv Marbach (DLA)
Höchstleistungsrechenzentrum Universität Stuttgart (HLRS)
Institut für Maschinelle Sprachverarbeitung Universität Stuttgart (IMS)
Institut für Literaturwissenschaft Universität Stuttgart (ILW)



Das Internet ist kein Speichermedium

Literatur im Netz 2010/2022: Wieviele Seiten sind noch online?

- Prüfkorpus: 134 Webseiten, davon 71 literarische Werke
- Seiten und Werke online 2022 (live web)
  - Webseiten: 79 von 134 (59 %)
  - Werke: 47 von 71 (66 %)

Webseiten und Werke
in Datenbanken und Archiven

Catalogue	Sites	Works
swb	22	22
iza	3	3
the_next	9	8
ia	107	59
kallias	33	28
cell	36	35
adel	45	42
elmcip	55	33

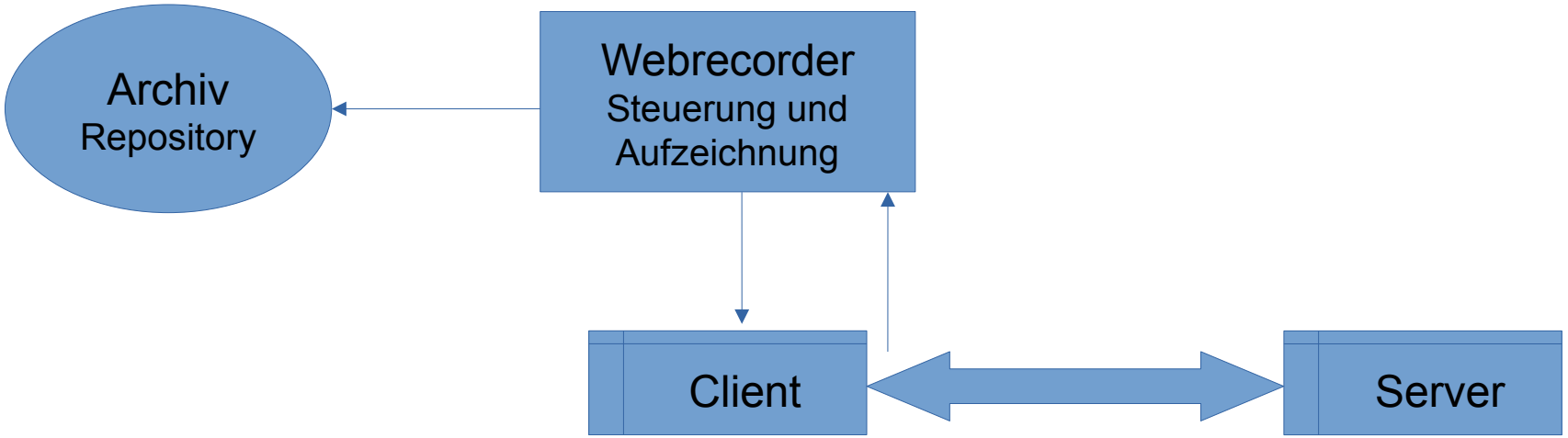
Ulrich/Schlesinger/Jorczick: Counting the Loss of E-Lit Works, DOI: 10.5281/zenodo.6587607

Literatur im Netz archivieren

- Live Web
  - Seiten können nicht stabil referenziert werden
  - Seiten verschwinden

- Archiv
  - stabile Referenz
  - Erschließung
  - Erhaltung

- Bestand DLA
  - 330 Blogs
  - 60 Netzliteraturwerke
  - 90 literarische Zeitschriften
  - WARC-Format
  - regelmäßige vollständige Crawls



URL: http://www.aaleskorte.de/

Titel: Die Aaleskorte der Ölig

Verfasser: Frank Klötgen, Dirk Günther

Jahr: 1998

WARC-Format

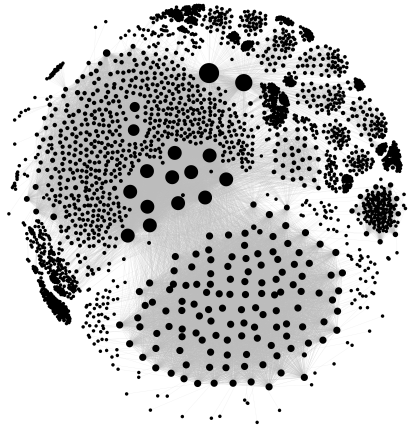
- ISO 28500:2017
- enthält alle durch den Client angefragten Ressourcen und die Kommunikation (requests und responses) zwischen Client und Server

```
WARC/1.0
WARC-Type: response
WARC-Target-URI: http://www.aaleskorte.de/anfang.htm
WARC-Date: 2015-07-02T07:25:47Z
WARC-Payload-Digest: sha1:IJLVDQYXWECXQ5AXT6GXEQ25NCSNDCJP
WARC-IP-Address: 82.165.104.8
WARC-Record-ID: <urn:uuid:796b9674-0a37-4f2a-8207-d48a5c4b9824>
Content-Type: application/http; msgtype=response
Content-Length: 1744

HTTP/1.1 200 OK
Date: Thu, 02 Jul 2015 07:25:47 GMT
Server: Apache
Last-Modified: Mon, 27 Oct 2003 17:52:49 GMT
ETag: "f9bb4026-5de-3caaed488fe40"
Accept-Ranges: bytes
Content-Length: 1502
Connection: close
Content-Type: text/html

<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 3.2 Final//EN">

<HTML>
<HEAD>
    <TITLE>Aaleskorte der Ölig</TITLE>
</HEAD>

<BODY bgcolor="black" text="#cccccc" link="red" alink="red"
vlink="red">
<table width="500"><TR><td>
<a href="index.htm">Stop!</A><P>
Zunächst müssen Sie Ihr Drehbuch für die "Aaleskorte der Ölig"
zusammenstellen. Keine Angst, diese Mühe kostet Sie ganze zwanzig
Mausklicks - dann startet der Film.<P>
<HR width="500" align="left">
Auf den folgenden Seiten werden Sie sehen:<BR>
die Portraits der Protagonisten, die für die jeweilige Szene anwählbar
sind, sowie eine kurze Regieanweisung.<P>
</TD></TR></TABLE>

<table width="500"><TR><td width="100"><img src="bilder/clipaal.jpg"
```

WARC-Objekte analysieren

- **Linkextraktion:** Graphdaten und Netzwerkanalyse
- **Textextraktion:** textbasierte Methoden (Suchindex)

Datenbasis	Derivate	Analyse
WARC	Links	Netzwerke
	Texte	Suche Textstatistik?
	...	



Visualisierung der internen Linkstruktur von "The Unknown" (Hypertext Novel), unknownhypertext.com, in: Viehhauser/Hein/

Archives Unleashed

https://archivesunleashed.org/

LINK GATE Alpha

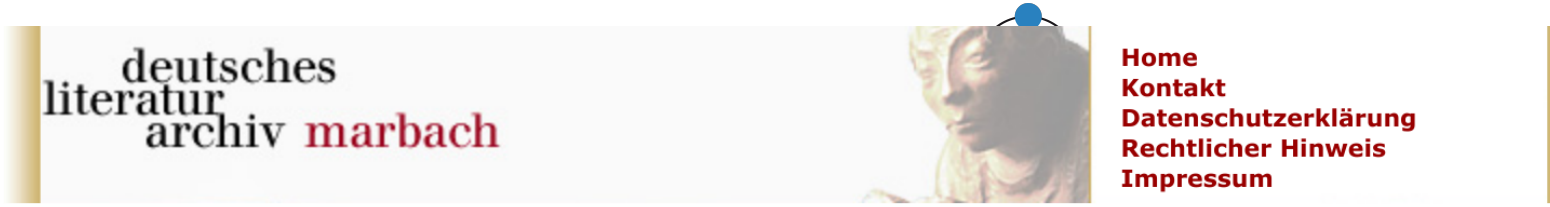
Web Archive Graph Visualization

https://linkgate.bibalex.org/
https://netpreserveblog.wordpress.com/2020/04/23/linkgate-update/

INTERNET ARCHIVE WayBackMachine

https://web.archive.org

WARC-Objekte analysieren: Literarische Blogs



Joachim Lottmann: Auf der Borderline nachts um halb eins, Crawl vom 12.4.2007, URN: urn:nbn:de:bsz:mar1-dd001-49f6aea4-3d5a-44d4-b29f-38e487e166ae1, Screenshot (http://literatur-im-netz.dla-marbach.de/jspview/downloads/frei/49f6aea4-3d5a-44d4-b29f-38e487e166ae/0/index.html)

WARC-Objekte analysieren: Objektstruktur Crawls

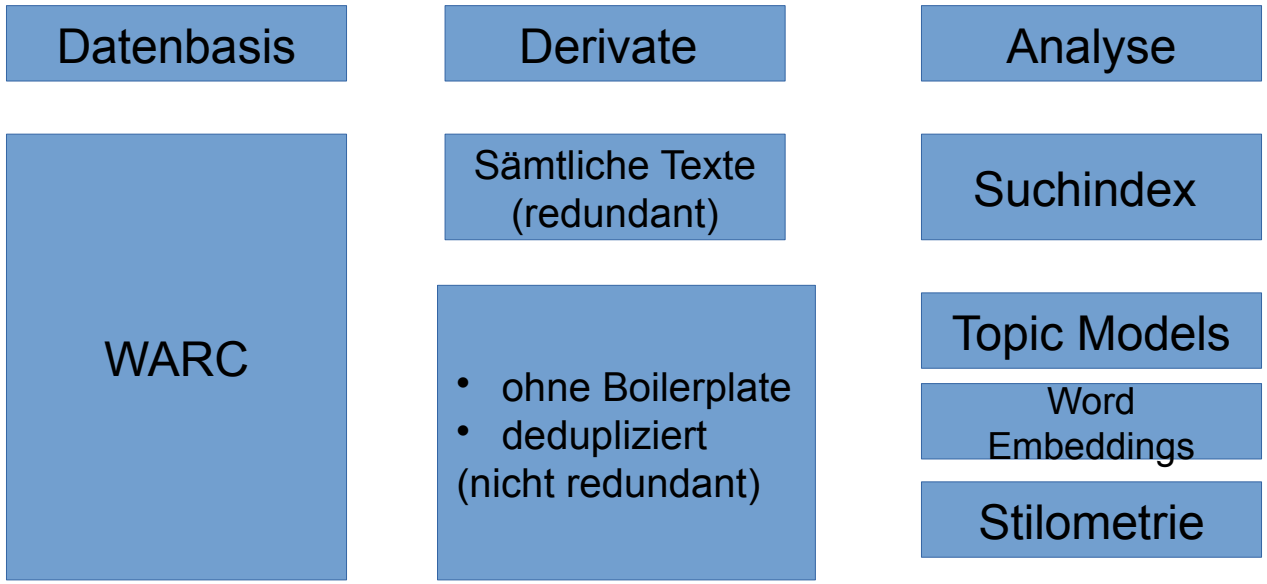
- Der Server liefert die Texte in verschiedenen Ansichten aus, z.B. Übersicht, Einzelbeitrag, Kategorienansicht usw.
- Ein vollständiger Crawl speichert alle Ansichten, die der Server ausliefert.

- => Einzeltexte sind im Crawl mehrfach redundant enthalten
- => Redundanz ist abhängig von serverseitigen Einstellungen und kann deshalb nicht kontrolliert werden

- Analyse: Die Redundanz der Daten ist ein Problem
- Replay: Die Redundanz der Daten ist ein Feature!

WARC-Objekte analysieren: Preprocessing und Derivate

Textstatistische Methoden wie Word Embeddings, Topic Modeling oder Most-Frequent-Words-Ansätze erfordern nicht-redundante Daten.



Schwierigkeiten bei der Einzelobjektanalyse:

- WARC-Crawls enthalten oft Daten außerhalb einer Seite
- Entfernung Boilerplate nicht exakt
- Deduplizierung nur eingeschränkt automatisierbar

Strukturierte Daten via API

Viele Content Management Systeme und Plattformen bieten Schnittstellen für den Download von strukturierten Daten an.

twitter.com

tumblr.com


```
{'contributors': None,
 'truncated': True,
 'text': 'katzenfutter\nkatzenfutter und rettich\n\nrettich\
nrettich und brokkoli\n\nkatzenfutter\nkatzenfutter und
brokkoli\n\nkatzen… https://t.co/cJk5nOvRsv',
 'is_quote_status': False,
 'in_reply_to_status_id': None,
 'id': 1447516333000859653,
 'favorite_count': 2,
 'source': '<a href="https://twitter.com/gomringador"
rel="nofollow">gomringador</a>',
 'retweeted': False,
 'coordinates': None,
 'entities': {'symbols': [],
  'user_mentions': [],
  'hashtags': [],
  'urls': [{'url': 'https://t.co/cJk5nOvRsv',
    'indices': [117, 140],
    'expanded_url':
'https://twitter.com/i/web/status/1447516333000859653',
```


Twitterbot @gomringador von Kathrin Passig, Ausschnitt aus dem Datensatz für einen Tweet, Tweet-ID: 1447516333000859653, Download 11.10.2021.

Ergänzende Archivierung via API

Struktur eines Datenpakets für @gomringador

```
.
├── screenshots
│   ├── 2019-10-15-000130_762x684_scrot.png
│   ├── Screenshot 2022-05-17 at 20-30-33 gomringador ( gomringador) Twitter.png
│   ├── Screenshot 2022-05-17 at 20-31-22 gomringador on Twitter.png
│   ├── Screenshot 2022-05-17 at 20-31-46 gomringador on Twitter.png
│   └── Screenshot 2022-05-17 at 20-32-11 gomringador ( gomringador) Twitter.png
├── sourceCodes
│   ├── source.v1.zip
│   └── source.v2.zip
├── structuredData
│   └── structuredData.zip
└── webArchive
    └── gomringador_20220519.wacz
```

Strukturierte Daten herunterladen

"Dschungel. Anderswelt" von Alban Nikolai Herbst

- 14995 Einträge (Mo 27. Jun 19:43:12 CEST 2022)
- 97 Kategorien
- 9875 Tags
- CMS: Wordpress, self-hosted
- API: Wordpress REST-API v2
- API Documentation: https://developer.wordpress.org/rest-api/
- API URL: https://dschungel-anderswelt.de/wp-json/wp/v2/
- Hosting Provider: Hetzner
- **Download without authentication: yes**

Strukturierte Daten herunterladen

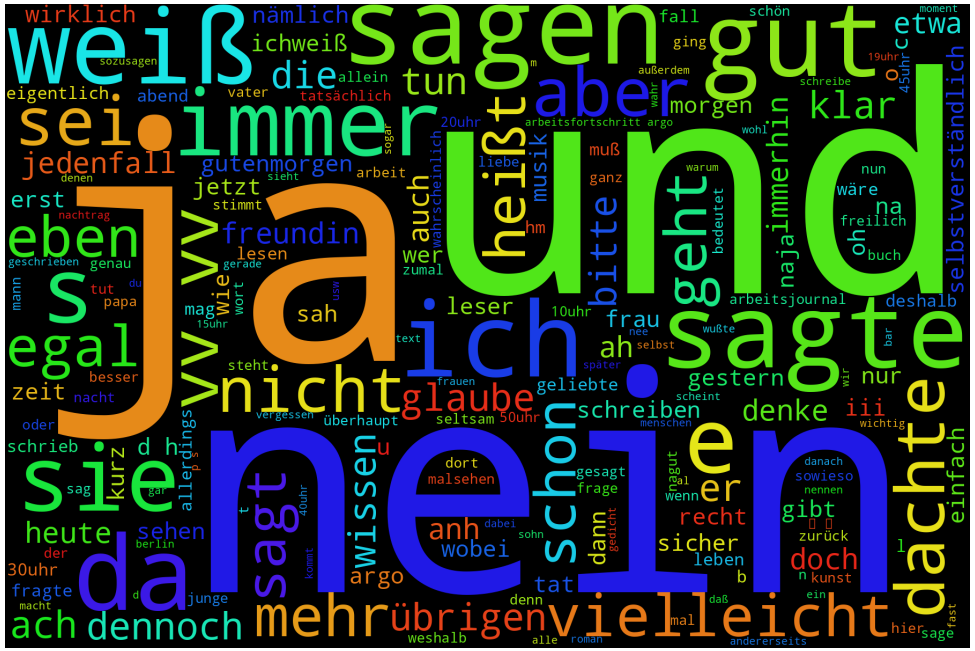
"Techniktagebuch" von Kathrin Passig u.a.

- ca. 7000 Einträge (Stand 11/2021)
- API: Tumblr
- **Download without authentication: yes**

	Webcrawl	API-Crawl
Größe	9.7 GB	<1 MB
Dauer	~25 h	30 Sekunden

Literatur im Netz analysieren

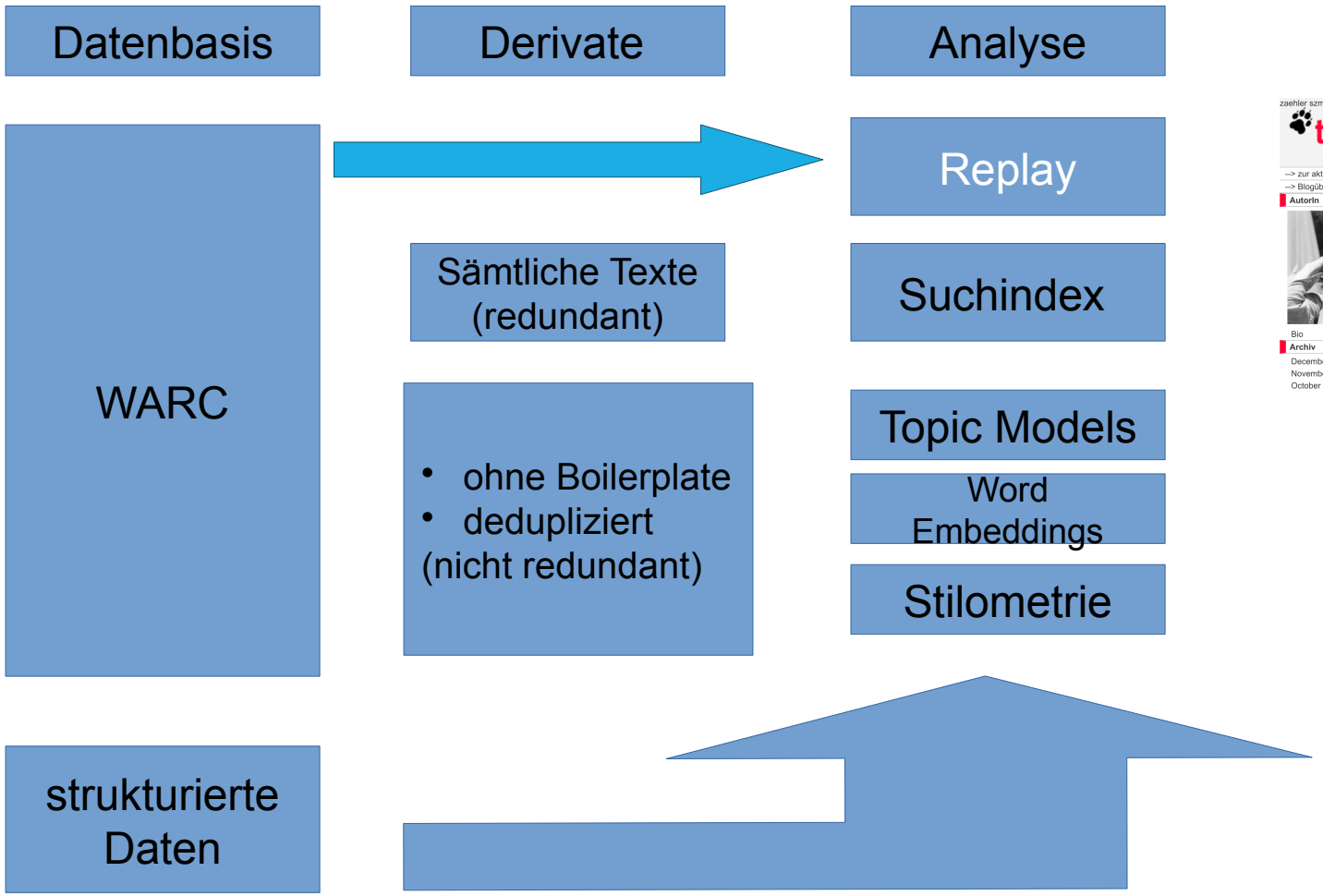
Datenbasis	Derivate	Analyse
WARC	Sämtliche Texte (redundant)	Suchindex
	• ohne Boilerplate • dedupliziert (nicht redundant)	Topic Models
		Word Embeddings
		Stilometrie
strukturierte Daten		

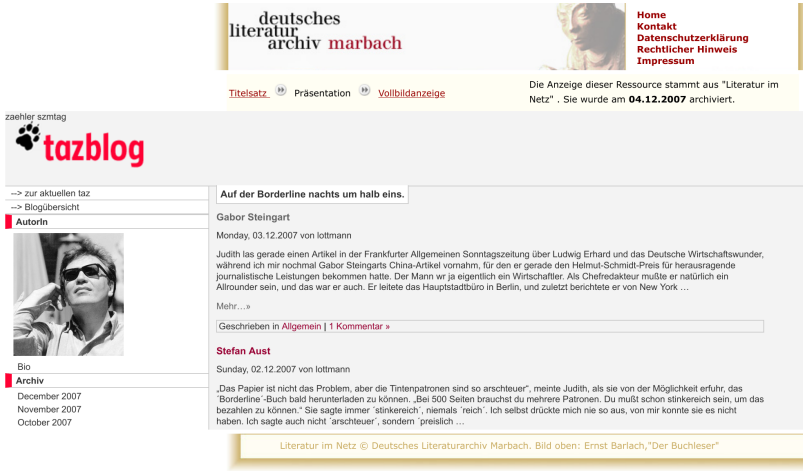


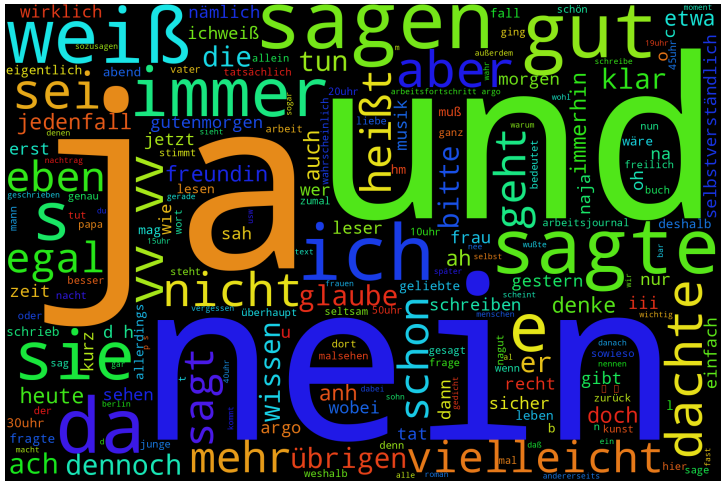
Wordclouds!

Wordcloud "Dschungel. Anderswelt" von Alban Nikolai Herbst, Datenbasis: Vollständiger API-Crawl vom 28.6.2022.

Literatur im Netz analysieren







Danke für Ihre Aufmerksamkeit!

Science Data Center for Literature (SDC4Lit) <https://sdc4lit.de>

Claus-Michael Schlesinger <cms@ilw.uni-stuttgart.de>

Mona Ulrich <mona.ulrich@dla-marbach.de>

Andre Blessing <andre.blessing@ims.uni-stuttgart.de>

DOI: 10.5281/zenodo.7107005