Vom Heben verborgener Schätze – Literarische Blogs als Ressource

Schenk, Nicolas
nicolas.schenk@dla-marbach.de
Deutsches Literaturarchiv Marbach

Blessing, André
andre.blessing@ims.uni-stuttgart.de
Universität Stuttgart, Institut für Maschinelle Sprachverarbeitung

Hein, Pascal
pascal.hein@ilw.uni-stuttgart.de
Universität Stuttgart, Institut für Literaturwissenschaft

Hess, Jan
jan.hess@dla-marbach.de
Deutsches Literaturarchiv Marbach

Jung, Kerstin
kerstin.jung@ims.uni-stuttgart.de
Universität Stuttgart, Institut für Maschinelle Sprachverarbeitung

Schlesinger, Claus-Michael
claus-michael.schlesinger@ilw.uni-stuttgart.de
Universität Stuttgart, Institut für Literaturwissenschaft

30 Mio. Token, 140.000 Blogposts, über 200 aufbereitete Blogs

Bereits seit Ende der 90er Jahre gewannen Weblogs als Medium zur öffentlichen Darstellung unterschiedlicher Themen und Inhalte immer mehr an Popularität. Findige Literatur- und Kulturschaffende zögerten nicht lange, um das neue Medium auch für literarische Zwecke umzunutzen.[1] Blogs wie *Die Dschungel. Anderswelt*[2] von Alban Nikolai Herbst, *Abfall für alle*[3] von Rainald Goetz, Wolfgang Herrndorfs *Arbeit und Struktur*[4] oder das kollaborativ betriebene Blog *Die Riesenmaschine*[5] werden seither zwar regelmäßig als Gegenstand wissenschaftlicher Untersuchungen herangezogen (vgl. Fassio 2021, Giacomuzzi 2008, Knapp 2014, Knapp 2012), wie viele andere Formen von Literatur im Netz im Vergleich zu ihren genuin analogen Pendants jedoch immer noch durchaus stiefmütterlich behandelt. Hinzu kommt, dass sich die Verfasser:innen dieser Blog-Studien in erster Linie klassisch hermeneutisch-literaturwissenschaftlicher Methoden bedienen und nur in seltenen Fällen auf computergestützte Analysemethoden und -werkzeuge zurückgreifen (vgl. zuletzt: Fassio 2021), obwohl die Weblogs schon ihrem Begriff nach born-digital sind und damit eine ‚digitale', computergestützte Form der Analyse zunächst auf der Hand zu liegen scheint.[6]

In dem Beitrag zur Jahrestagung der DHd 2022 (Blessing et al 2022) haben die Autor:innen des vorliegenden Beitrags bereits am Beispiel des u. a. von der Autorin Kathrin Passig ins Leben gerufenen Techniktagebuch[7] verschiedene computergestützte Möglichkeiten sowie geeignete Werkzeuge für die Analyse literarischer Blogs vorgestellt. Der in der End-Anwendung später kaum sichtbare Aufwand an textuellen Preprocessing-Schritten, der bereits an diesem vermeintlich einfachen Fallbeispiel offenbar wurde, gepaart mit den Reaktionen und dem großen Interesse aus der wissenschaftlichen Community an der grundlegenden Methodik zum Umgang mit WARC-Dateien, lassen bereits die Gründe erahnen, weshalb in der Blog-Forschung digitale Methoden bislang vergleichsweise selten zum Einsatz kamen. Im hier nun vorliegenden Beitrag sollen – als Fortsetzung und Erweiterung des vorangehenden Beitrags auf der DHd 2022 – nicht nur exemplarisch ebendiese Herausforderungen, sondern vielmehr auch Lösungsmöglichkeiten im Umgang mit (archivierten) Blogs aufgezeigt werden. Im Fokus der Untersuchung steht dabei nicht mehr nur ein einzelnes Blog, sondern vielmehr ein insgesamt aus über 200 aufbereiteten Blogs mit ca. 140.000 Blogposts und 30 Millionen Token bestehender Fundus literarischer Blogs. Ebensolche literarische Weblogs sammelt die Bibliothek des Deutschen Literaturarchivs Marbach neben literarischen Zeitschriften und Objekten der Netzliteratur bereits seit 2008.[8] Bislang erfolgt die Bereitstellung dieser Sammlungsobjekte über die Plattform Literatur-im-Netz[9], 2023 werden diese zusammen mit der hier vorgestellten Ressource über das SDC4Lit-Repositorium bereitgestellt.[10]

Bausteine von Weblogs

Typische Bausteine in Blogs sind Blogposts als (meist datierte) Inhaltseinheiten verschiedenster Länge und unter Verwendung unterschiedlicher Modalitäten wie Text, Bild, Animation, Video, Referenz (z. B. Hyperlinks), etc. Weiterhin finden sich auf der Ebene der Blogposts oft Kommentarformulare und Kommentare sowie eine Etikettierung von Einträgen in Form von Tags, die der Gruppierung und Beschreibung der Einträge dienen und sich auf Themen, Stimmungen, Autoren, etc. des Blogposts beziehen können (vgl. zu den Bausteinen von Blogs: Ernst 2010, 286f.). Zusätzliche Elemente wie Übersichtsseiten, die als Einstiegsseiten der jeweiligen Blogs die einzelnen Blogposts (zusätzlich) in einer bestimmten Reihenfolge auflisten, oder Archivseiten, die den Nutzer:innen einen Zugang zu älteren Blogposts ermöglichen sollen, prägen die Struktur der Blogs und damit ggf. deren Analyse.

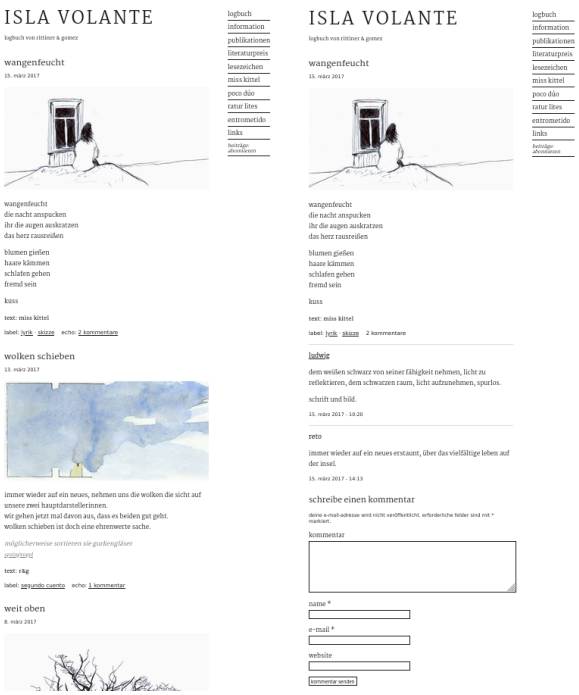

Abb.1: Links: Übersichtsseite (Home) mit mehreren Blogposts. Rechts: Seite des ersten Blogposts mit Tags, Kommentaren und Kommentarfunktion. Beispiele aus "Logbuch Isla volante : Bilder und Texte von der Insel". Spiegelung 2017.03.17_01, URN: urn:nbn:de:bsz:mar1-dd001-fe31dc38-7c43-4da2-ae3d-fd9619f88ea45.

Die Spiegelungen der Blogs, die am DLA durchgeführt werden, erfolgen zunächst mit einem Crawling-Vorgang, der der Hyperlinkstruktur im Blog folgt und die clientseitig (wie durch einen Browser) empfangenen Daten gemeinsam mit einigen Metadaten zum Crawlingprozess im Web ARChive- oder kurz: WARC-Format ablegt. Das aus dem ARC-Format des Internet Archive weiterentwickelte Archivformat hat sich inzwischen als internationaler Standard für die Archivierung von Webinhalten etabliert (IIPC, n.d.). Beim Crawling können Inhalte, die ggf. nicht Teil des Blogs sind, wie z. B. zufällig eingebundene Werbeanzeigen oder externe Inhalte, auf die von diesen Werbeanzeigen verwiesen wird, Teil des Archivobjekts werden. Aber auch bezüglich der tatsächlichen Blog-Elemente sind im Archivobjekt Inhalte (z. B. bestimmte Textpassagen) so oft abgelegt, wie der Crawler ihnen auf verschiedenen Seiten begegnet ist. Der Inhalt eines Blogposts kann an verschiedenen Stellen für den Crawler erreichbar sein: auf der Übersichtsseite, auf der eigenen Seite des Posts, auf der Seite jedes Schlagworts, mit dem der Post versehen ist, im Archiv, etc. Aber auch Textpassagen aus Strukturelementen wie Kopf- und Fußzeilen führen zu mehrfachen Vorkommen von Textpassagen oder Begriffen.

Werkzeuge, die Strukturelemente ausblenden und (Text-)Duplikate erkennen, sind daher bei der Vorverarbeitung der Daten für Analysen notwendig, arbeiten oft aber statistisch und müssen ggf. auf jedes zu untersuchende Blog neu angepasst werden, was im Spannungsfeld mit dem maschinell unterstützten Distant Reading steht. Im Folgenden wird daher das Vorgehen bei der Aufbereitung der Blogs beschrieben, das notwendig ist, um die Texte auch für quantitative Analysen verfügbar zu machen.

Aufbereitung der Blogs als Ressource für quantitative Analysen

Die am DLA archivierten Blogs sind nicht nur inhaltlich und stilistisch sehr heterogen, sondern auch in Bezug auf die technische Umsetzung. In den meisten Fällen werden bekannte Blog-Hoster wie wordpress (40%), twoday (15%), blogger, blogspot usw. verwendet. Diese Blog-Hoster wiederum setzen Content Management Systeme (CMS) ein, die die Blogpost-Erstellung, -Verwaltung sowie die Blogdarstellung sowohl für die Blogverfasser:innen, als auch für die Blogleser:innen vereinfachen.[11]

Für die Analyse des Inhalts eines Blogs sind nur die Blogposts relevant, da alle Übersichtsseiten (Home, Tags, Categories) eines Blogs automatisch daraus generiert werden. Darüber hinaus hat sich gezeigt, dass beim Erstellen von WARC-Crawls auch einiges an ‚Beifang', also Webseiten, die nicht zum Blog gehören, aber für das Abspielen teilweise nützlich sein könnten, mit-archiviert werden, was wiederum bei der Aufbereitung der Blogs beachtet werden muss. Die hier beschriebene Umsetzung zielt darauf ab, ein sauberes Textkorpus für jedes Blog zu extrahieren und alle irrelevanten und redundanten Inhalte zu ignorieren.

Bei der Aufbereitung werden Verfahren aus dem Information Retrieval eingesetzt (vgl. Manning, Raghavan und Schütze 2008, 443–459). Das Besondere am aktuellen Datensatz liegt darin, dass die Daten bereits im WARC-Format vorliegen. Cormack, Smucker und Clarke (2011) haben bereits gezeigt, wie bei sehr großen Web-Crawls [12], die als WARCs vorliegen, relevante Informationen extrahiert werden können. Unser Szenario unterscheidet sich dahingehend, dass hier eine klare Definition der gesuchten Information – in Form von Blogposts – existiert. Dies erfordert einen feingranularen Ansatz.

Die Umsetzung läuft in mehreren Schritten: In Schritt 1 werden alle HTML-Seiten des Blogs aus den WARCs extrahiert und erkannte Fremdinhalte entfernt. Schritt 2 erkennt das verwendete CMS-System. Schritt 3 basiert auf einem Regelsystem, das für jedes Blog die Posts ermittelt. Die URL-Pfade der Blogs sind ein Hauptmerkmal, welches in den Regelsystemen verwendet wird. In Schritt 4 werden zu jedem Post Text- und Metadaten mittels trafilatura[13] extrahiert. Alle Schritte sind als Jupyter-Notebooks implementiert und werden zusammen mit der Ressource veröffentlicht. Nach der Aufbereitung liegen die Daten im JSON- und txt-Format vor. Ein JSON-Datensatz enthält Text und Metadaten und kann z. B. direkt in Analyse-Tools importiert werden. Zum Aufbereitungsprozess gibt es Logfiles, die für die iterative Verbesserung des Prozesses ausgewertet werden.

Da in allen vier dieser Umsetzungsschritte Anpassungen notwendig sind, gehen Weiterentwicklung und Optimierung der Aufbereitung immer einher mit dem Einsatz von Analysewerkzeugen. Bei der Verwendung solcher Werkzeuge können systematische Auffälligkeiten sicht-

bar werden, wodurch beispielsweise noch vorhandene Redundanzen zum Vorschein kommen. Ein Werkzeug, mit dem große Textkorpora durchsucht und auf grammatikalische sowie syntaktische Strukturen hin untersucht werden können, ist CQPweb[14] (Hardie 2012, 380–409). Ein Auszug aus einer Beispielanfrage, in der – angelehnt an eine mögliche Forschungsfrage zu Formen der Erinnerung – nach dem Zeichenstring „erinnere mich" gesucht wird, bringt redundante Textstellen zutage (Abb. 2). Dieser Hinweis fließt in die Optimierung der Regeln in den Aufbereitungsskripten ein.

wirklich ein Gedanke , der mir noch überhaupt nicht kam und ich	erinnere mich	an keine böse Tat . Ich weiß wirklich nicht , was da
wirklich ein Gedanke , der mir noch überhaupt nicht kam und ich	erinnere mich	an keine böse Tat . Ich weiß wirklich nicht , was da
auf den Acker , sein Getreide war im das Heiligste . Ich	erinnere mich	an einen großen Sturm , der all das Gesäte vernichtete und das
kommentieren . Und Fischgrät wollte ich auch nicht erwähnen ...Ich	erinnere mich	nicht an Blackstar in dem Blog . Das muss ich doch direkt
kommentieren . Und Fischgrät wollte ich auch nicht erwähnen ...Ich	erinnere mich	nicht an Blackstar in dem Blog . Das muss ich doch direkt
in den Spiegel schauen . Ich fühle mich wie Restmüll . Ich	erinnere mich	an die Blicke dieser Männer . Meine Lebensfreude ist dahin . Ich
in den Spiegel schauen . Ich fühle mich wie Restmüll . Ich	erinnere mich	an die Blicke dieser Männer . Meine Lebensfreude ist dahin . Ich
Ihr Krankheits-Management " , ich denke Ihre NT hat Recht , ich	erinnere mich	, wie Sie den Wortmacher in seine Schranken wiesen und sagten :
Ihr Krankheits-Management " , ich denke Ihre NT hat Recht , ich	erinnere mich	, wie Sie den Wortmacher in seine Schranken wiesen und sagten :
Ich tu zum Teil nur noch so als ob , und	erinnere mich	an die Zeit vor Normal + XR , also an mein Durchlittenes
Ich tu zum Teil nur noch so als ob , und	erinnere mich	an die Zeit vor Normal + XR , also an mein Durchlittenes
aufgewühlt . Ich tu zum Teil nur so als ob , und	erinnere mich	an die Zeit vor Normal + XR , also an mein Durchlittenes
aufgewühlt . Ich tu zum Teil nur so als ob , und	erinnere mich	an die Zeit vor Normal + XR , also an mein Durchlittenes

Abb. 2: Auszug aus einer CQPweb-Testinstanz mit dem Suchbegriff „erinnere mich".

Das Durchsuchen und Abspielen von WARC-Dateien wird durch die Web-Applikation SolrWayback[15] ermöglicht. In der Ergebnisliste zur Suche nach dem Begriff „techniktagebuch" verweisen die ersten Treffer alle auf Twitter (Abb. 3). Der Inhalt des Tweets bleibt dabei gleich, jedoch steht das Webinterface in verschiedenen Sprachen zur Verfügung. Insgesamt werden bei der Suchanfrage knapp 50 verschiedene Sprachen für die Twitterseite gefunden.

Abb. 3: Auszug aus der Trefferliste einer SolrWayback-Testinstanz mit der Suchanfrage „techniktagebuch" und dem Beginn der CSV-Trefferliste.

Der Erfolg bei der Aufbereitung lässt sich anhand der Zahlen in Tabelle 1 ablesen. Hier wird auch die Heterogenität der Blogs sichtbar: Die Aufbereitungsschritte haben je nach Blog unterschiedlich starke Auswirkungen. Während der Unterschied der Anzahl von WARC-Records und HTML-Seiten deutlich macht, wie viele externe Inhalte und Nicht-HTML-Inhalte entfernt werden müssen, zeigt sich an der Differenz zwischen der Anzahl an HTML-Seiten und der Anzahl an Blogposts, wie hoch die Redundanz durch generierte Übersichtsseiten ausfällt.

Tab. 1: Übersicht der resultierenden Ressourcengrößen nach der schrittweisen Aufbereitung für vier Beispielblogs.

Blog	WARC-Records	HTML-Seiten	Blogposts	Token
Die.Dschungel.Anderswelt[16]	123.859	65.014	14.106	7,8 Mio
Techniktagebuch[17]	310.131	127.532	5.233	1,6 Mio
Lux autumnalis[18]	11.121	6.426	3.092	1,3 Mio
Henrikes Tagebuch[19]	499.878	168.856	861	0,2 Mio

Evaluation

Die Zahlen zur Ressourcengröße aus Tabelle 1 veranschaulichen, dass die Aufbereitung der Blogs nicht vollständig manuell validiert werden kann. Trotzdem ist eine Aussage über die Qualität der aufbereiteten Blog-Daten wichtig. Zu diesem Zweck wurde ein Annotations-Jupyter-Notebook entworfen, mit dessen Hilfe automatisch aus den jeweiligen CMS-Familien wie beispielsweise Wordpress oder Blogger repräsentative Datenmengen entnommen werden, die anschließend von mehreren Annotator:innen bewertet werden. Im ersten Schritt geht es um das Erkennen der Blog-Posts in Abgrenzung zu Übersichtsseiten oder weiteren Seiten wie beispielsweise dem Impressum. Wenn es sich bei der Seite um einen Post handelt, dann wird im zweiten Schritt die Qualität der Metadaten- und Textextraktion bewertet. Dabei wird u. a. darauf geachtet, ob Datum und Überschrift richtig extrahiert wurden und ob der extrahierte Text vollständig ist oder beispielsweise Teile aus der Seitennavigation enthalten sind.

Es hat sich gezeigt, dass die beschriebene manuelle Bewertung auch für Menschen nicht immer trivial ist. Daher ist der Annotationsprozess aktuell noch nicht abgeschlossen. Die Evaluationsergebnisse werden jedoch mit dem Release der Daten bereitgestellt.

Nutzungsszenarien für die aufbereiteten Blog-Daten

Nach den oben beschriebenen Aufbereitungsschritten sind die daraus resultierenden Daten (nahezu) frei von Dubletten und Fremdinhalten, sodass sie für verschiedene Formen der Textanalyse sowie für weitere Prozessierungsschritte verwendet werden können. Dazu gehören korpuslinguistische Untersuchungen, wie sie mit CQPweb durchgeführt werden können, das Erstellen von Sprachmodellen (z. B. embeddings, topic-models) oder automatische Keyword-Erkennung. Auch die Untersuchung der in den Blogposts enthaltenen Metadaten (Datum, Autor:in, Verschlagwortung) kann für die Forschung von Interesse sein. Beispielhaft erwähnt sei an dieser Stelle die Verschlagwortung der einzelnen Blogs durch die jeweiligen Blog-Autor:innen, die mit End-Anwendungen wie Keshif[20] erst auf Basis der aufbereiteten Blog-Daten überhaupt sinnvoll visualisiert und analysiert werden können (vgl. Abb. 4).

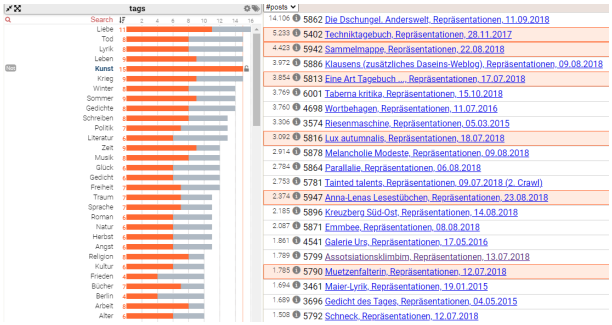

Abb. 4: Screenshot aus einer Keshif-Testinstanz. Rechts werden alle Blogs hervorgehoben, in denen das Schlagwort Kunst vergeben wurde, links befindet sich eine Liste aller Schlagworte.

In Blessing et al (2022) wurde am Fallbeispiel des Techniktagebuchs bereits exemplarisch aufgezeigt, welche komplexeren computergestützten Analysen durch die Verwendung der extrahierten Text- und Metadaten vorgenommen werden können. Unter anderem wurde bereits untersucht, welche Zusammenhänge zwischen Inhalt – repräsentiert durch automatische Keyword-Erkennung und die Verschlagwortung durch die Autor:innen – und Form bzw. Sprache erkennbar werden.

Schätze heben und nutzen

30 Millionen Zeichen, 140.000 Blogposts, über 200 Blogs: Schon wegen seiner Größe ist das hier vorgestellte, aufbereitete Korpus für viele Bereiche der Digital Humanities, beispielsweise die Computational Literary Studies, die Digital History oder für NLP-Untersuchungen, eine wichtige Quelle für Inhaltsanalysen oder das Trainieren von Sprachmodellen. Wie in diesem Beitrag gezeigt stellt vor allem die Struktur der Weblogs eine Herausforderung dar, enthalten die WARC-Dateien, in denen die Blogs zunächst vorliegen, doch sehr viel Redundantes, das für eine Vielzahl von Inhaltsanalysen nicht nur uninteressant, sondern sogar hinderlich ist. Mit den im Zuge der Veröffentlichung der SDC4Lit-Plattform 2023[21] zur Verfügung gestellten Blog-Daten in aufbereiteter Form wird eine robuste Ressource bereitgestellt, die neben den Rohdaten im WARC-Format auch das bereinigte Textkorpus in Form der inhaltlich relevanten Blogposts sowie die zugehörigen Metadaten zu jedem Post enthält. Dank der ebenfalls über SDC4Lit zur Verfügung gestellten WARC-Volltextsuche und WARC-Player wie SolrWayback können die Blogs bzw. die einzelnen Blogposts zudem – unabhängig von allen weiteren Analyseschritten – möglichst originalgetreu in ihrer ursprünglichen Repräsentationsform angesehen und erforscht werden, auch wenn die originalen Webseiten bereits nicht mehr vorhanden sind oder geändert wurden. Die Implementierung der Aufbereitung wird in Form von dokumentierten Jupyter-Notebooks bereitgestellt, dank der auch weitere, über das hier präsentierte Korpus hinausgehende (literarische) Blogs aufbereitet und damit für weitere DH-Bereiche zugänglich gemacht werden können, sodass die nunmehr bereits gehobenen Weblog-Schätze künftig nicht die einzigen bleiben.

Fußnoten

1. In Bezugnahme auf Ernst (2010, 294 –297); Giacomuzzi (2012, 183) und Jürgensen (2011, 407) liefert Fassio (2021, 97f.) eine ausführliche Diskussion der bisherigen Definitions- und Typisierungsversuche literarischer Weblogs.
2. https://dschungel-anderswelt.de/ , (zugegriffen: 14. Dezember 2022).
3. Das Blog erschien zunächst in einer Folge von 343 Blogposts auf rainaldgoetz.de (Quelle offline), anschließend auch in Buchform: Goetz, Rainald. 2015. "Abfall für alle. Roman eines Jahres." Frankfurt a. M.: Suhrkamp.
4. https://www.wolfgang-herrndorf.de/ , (zugegriffen: 14. Dezember 2022) ; auch als Print-Ausgabe: Wolfgang Herrndorf. 2015. "Arbeit und Struktur." Rowohlt: Reinbek.
5. http://riesenmaschine.de/ , (zugegriffen: 14. Dezember 2022).
6. Neben hermeneutischen Untersuchung en einzelner literarischer Blogs (Ainetter 2006, Knapp 2012, 2014) oder theoretischer bzw. struktureller Überlegungen (Ernst 2010) stellt ein Großteil der Studien zu (literarischen) Weblogs die Beziehung bzw. Abgrenzung verwandter Textsorten wie Flugblatt, Zeitung, Autobiographie oder – am häufigsten – dem Tagebuch in den Fokus (Augustin 2015, Flüh 2017, Jürgensen 2011, Michelbach 2019).
7. https://techniktagebuch.tumblr.com/ , (zugegriffen: 14. Dezember 2022).
8. Das Deutsche Literaturarchiv Marbach sammelt deutschsprachige Literatur von 1750 bis zur Gegenwart, sodass unser Korpus vorwiegend aus deutschsprachigen Blogs besteht. Allerdings finden sich in unserer Sammlung auch an manchen Stellen nicht-deutschsprachige Absätze oder gar ganze Blogeinträge. Die Technik zur Extraktion der Texte lässt sich zu weiten Teilen auch auf andere Sprachen übertragen, sodass die entwickelte Pipeline auch für andere Sprachen vollständig oder zumindest größtenteils nachnutzbar ist.
9. http://literatur-im-netz.dla-marbach.de , (zugegriffen: 14. Dezember 2022).
10. Im Rahmen des Projekts *SDC4Lit – Aufbau eines nachhaltigen Datenlebenszyklus für Literaturforschung und -vermittlung (* https://www.sdc4lit.de/ , zugegriffen: 14. Dezember 2022) entsteht seit 2019 die SDC4Lit-Plattform, über die die gesammelten Blogs zusammen mit den sonstigen Beständen von Literatur im Netz nicht nur in einem Repositorium verfügbar gemacht, sondern auch (explorative) Zugangs- und Analysemöglichkeiten aufgezeigt und bereitgestellt werden.
11. Bei einigen Blog-Hostern stehen APIs zur Verfügung, über die der Download einer inhaltsorientierten Version der Blogs möglich ist. Diese Repräsentationen sind bisher nicht Teil der Sammlung.
12. Der ClueWeb09-Datensatz umfasst 1.040.809.705 Webseiten. http://lemurproject.org/clueweb09/ (zugegriffen: 14. Dezember 2022).
13. https://github.com/adbar/trafilatura , (zugegriffen: 14. Dezember 2022), eine Python-Bibliothek zur Boilerplate-Entfernung und Metadatenerkennung (Barbaresi 2019).
14. https://cqpweb.lancs.ac.uk , (zugegriffen: 14. Dezember 2022).

15. https://github.com/netarchivesuite/solrwayback , (zugegriffen: 14. Dezember 2022).
16. https://dschungel-anderswelt.de , (zugegriffen: 14. Dezember 2022).
17. https://techniktagebuch.tumblr.com , (zugegriffen: 14. Dezember 2022).
18. http://www.luxautumnalis.de , (zugegriffen: 14. Dezember 2022).
19. http://henrikeheiland.blogspot.de/ , (zugegriffen: 14. Dezember 2022).
20. https://github.com/adilyalcin/Keshif , (zugegriffen: 14. Dezember 2022).
21. https://www.sdc4lit.de/, (zugegriffen: 14. Dezember 2022).

Bibliographie

Ainetter,Sylvia. 2006. "Blogs - literarische Aspekte eines neuen Mediums. Eine Analyse am Beispiel des Weblogs Miagolare". Wien: Lit Verlag.

Augustin, **Elisabeth**. 2015. "BlogLife. Zur Bewältigung von Lebensereignissen in Weblogs." Bielefeld: transcript.

Barbaresi, Adrien. 2019. "Generic Web Content Extraction with Open-Source Software". In *Proceedings of the 15th Conference on Natural Language Processing* (KONVENS 2019): 267–268.

Blessing, André, Jan Hess und Kerstin Jung. 2022. "'Ja, jetzt ist das langweilig. Aber in zwanzig Jahren!' - Bereitstellung, Zugang und Analyse literarischer Blogs am Beispiel des Techniktagebuchs." *DHd 2022 Kulturen des digitalen Gedächtnisses. 8. Tagung des Verbands "Digital Humanities im deutschsprachigen Raum" (DHd 2022), Potsdam*. Zenodo: https://doi.org/10.5281/zenodo.6322488; https://doi.org/10.5281/zenodo.6328029.

Cormack, Gordon V., Mark D. Smucker und Charles L. A. Clarke. 2011. "Efficient and effective spam filtering and re-ranking for large web datasets." In *Information retrieval 14*: 441–465.

Ernst, Thomas. 2010. "Weblogs. Ein globales Medienformat." In *Globalisierung und Gegenwartsliteratur. Konstellationen - Konzepte - Perspektiven*, hg. von Wilhelm Amann, Georg Mein und Rolf Parr, 281–302. Heidelberg: Synchron Wissenschaftsverlag der Autoren.

Fassio, Marcella. 2021. "Das literarische Weblog. Praktiken, Poetiken, Autorschaften". Bielefeld: Transcript.

Flüh, Thorsten. 2017. "Flugblatt – Zeitung – Blog. Materialität und Medialität als Literaturen." Wien: Passagen Verlag.

Giacomuzzi, Renate. 2008. "Die ‚Dschungel. Anderswelt' und A. N. Herbsts ‚Poetologie des literarischen Bloggens'." *Die Horen 53*: 137–149.

Giacomuzzi, Renate. 2012. "Deutschsprachige Literaturmagazine im Internet. Ein Handbuch." Innsbruck: Studien-Verlag.

Hardie, Andrew. 2012. "CQPweb – combining power, flexibility and usability in a corpus analysis tool." In *International Journal of Corpus Linguistics* 17(3): 380 –409.

IIPC. n. d. "The WARC Format." https://iipc.github.io/warc-specifications/specifications/warc-format/warc-1.1/ (zugegriffen: 14. Dezember 2022).

Jürgensen, Christoph. 2011. "Ins Netz gegangen – Inszenierungen von Autorschaft im Internet am Beispiel von Rainald Goetz und Alban Nikolai Herbst." In *Schriftstellerische Inszenierungspraktiken – Typologie und Geschichte*, hg. von Christoph Jürgensen und Gerhard Kaiser, 405–422. Heidelberg: Winter.

Knapp, Lore. 2012. "Christoph Schlingensiefs Blog. Multimediale Autofiktion im Künstlerblog." In *Narrative Genres im Internet: Theoretische Bezugsrahmen, Mediengattungstypologie und Funktionen*, hg. von Ansgar Nünning und Jan Rupp, 117–132. Trier: Wissenschaftlicher Verlag.

Knapp, Lore. 2014. "Künstlerblogs. Zum Einfluss der Digitalisierung auf literarische Schreibprozesse (Goetz, Schlingensief, Herrndorf)." Berlin: Ripperger & Kremers.

Manning, Christopher D., Prabhakar Raghavan und Hinrich Schütze. 2010. "Introduction to information retrieval." In *Information retrieval 13*: 192–195. https://doi.org/10.1007/s10791-009-9115-y.

Michelbach, Elisabeth. 2019. "Poetik des autobiografischen Blogs." Dissertation, Universität Göttingen.