SDC4Lit – Science Data Center for Literature. Aufbau eines nachhaltigen Datenlebenszyklus für Literaturforschung und -vermittlung

Jan Hess[1], Alexander Holz[1], Nina Buck[2], Andreas Ganzenmüller[2], Volodymyr Kushnarenko[2], Björn Schembera[2], André Blessing[4], Pascal Hein[3], Kerstin Jung[4], Heinz Werner Kramski[1], Claus-Michael Schlesinger[3], Mona Ulrich[1], Thomas Bönisch[2], Andreas Kaminski[2], Roland S. Kamzelak[1], Jonas Kuhn[4] and Gabriel Viehhauser[3]

[1]Deutsches Literaturarchiv Marbach
[2]Höchstleistungsrechenzentrum Stuttgart
[3]Institut für Literaturwissenschaft/Digital Humanities, Universität Stuttgart
[4]Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart

Das Science Data Center for Literature (SDC4Lit) hat sich das Ziel gesetzt, die Anforderungen, die (Digitale) Literatur an ihre Archivierung, Erforschung und Vermittlung stellt, systematisch zu reflektieren und entsprechende Lösungen für einen nachhaltigen Datenlebenszyklus für Literaturforschung und -vermittlung langfristig umzusetzen. Im Zentrum stehen dabei der Aufbau langzeitverfügbarer Repositories für (Digitale) Literatur und die Entwicklung einer Forschungsplattform.

1 Projektziel

Die Digitalisierung verändert die Bedingungen für die Produktion, Distribution, Rezeption und damit auch für die Erforschung von Literatur. Die veränderten medialen Bedingungen führen nicht nur zur Übersetzung von gedruckten Texten in digitale Objekte, sondern bringen selbst produktiv neue Literaturformen und -gattungen hervor. Hierzu zählen etwa literarische Hypertexte, Blog-Formate, literarische Tweets und Twitter-Bots, aber auch Texte und Textgeneratoren, die auf computerlinguistische Methoden setzen. Zum einem scheinen sich diese Texte zur Anwendung computergestützter Analysemethoden besonders anzubieten, da sie genuin in elektronischer Form vorliegen. Zum anderen bringt diese Form für ihre Archivierung und Bereitstellung eine Reihe von besonderen Anforderungen mit sich.

Das hier beschriebene Poster ist in der Open Access-Plattform der Universität Heidelberg heiDOK unter der DOI: https://doi.org/10.11588/heidok.00030001 veröffentlicht.

Publiziert in: Vincent Heuveline, Nina Bisheh (Hg.): E-Science-Tage 2021. Share Your Research Data. Heidelberg: heiBOOKS, 2022.
DOI: https://doi.org/10.11588/heibooks.979.c13748 (CC BY-SA 4.0)

So führen die hochfrequenten Erneuerungszyklen digitaler Technik dazu, dass die ursprünglichen Darbietungsformen historischer elektronischer Texte teils aufwendig rekonstruiert werden müssen, da die entsprechende Hard- oder Software schnell veraltet ist.

Das Science Data Center for Literature (SDC4Lit)[1] hat sich das Ziel gesetzt, die Anforderungen, die Digitale Literatur an ihre Archivierung, Erforschung und Vermittlung stellt, systematisch zu reflektieren und entsprechende Lösungen für einen nachhaltigen Datenlebenszyklus für Literaturforschung und -vermittlung langfristig umzusetzen. Im Zentrum stehen dabei der Aufbau eines langzeitverfügbaren Repositoriums für (Digitale) Literatur und die Entwicklung einer Forschungsplattform, die die Möglichkeit zum computergestützten Arbeiten mit den Beständen der Repositorien bietet. Da eine solche Repositoriumsstruktur, die Sammeln, Archivieren und Analysieren miteinander verzahnt, nur in der interdisziplinären Zusammenarbeit zu bewerkstelligen ist, sind mit dem Deutschen Literaturarchiv Marbach (DLA), dem Höchstleistungsrechenzentrum Stuttgart (HLRS), dem Institut für Maschinelle Sprachverarbeitung (IMS) sowie dem Institut für Literaturwissenschaft/Digital Humanities (ILW) an der Universität Stuttgart im Projekt Partner mit Expertisen in den verschiedenen Bereichen miteinander vereint.

2 Datenmaterial

Die Daten, die im Repositorium gesammelt, archiviert und – sofern rechtlich möglich – zur Verfügung gestellt werden, stammen aus der Sammlung des Deutschen Literaturarchivs Marbach und werden stetig durch neue Bestände und Objekte ergänzt. Die Datenmenge lässt sich in drei Themenbereiche gliedern.

Den ersten Teil bildet der Bereich Literatur im Netz. Hierbei handelt es sich um archivierte Webseiten, literarische Blogs oder Online-Magazine mit Bezug zur Neueren deutschen Literatur. Dieses Korpus wurde am DLA von 2008 bis 2018 zusammengestellt und kuratiert. In dieser Zeit wurden etwa 500 Internet-Quellen einmalig oder wiederholt gespeichert und so insgesamt ca. 3.840 Speicherungen vollzogen. Zur Archivierung der Webseiten wurden über die Jahre unterschiedliche Capturing-Tools und -Techniken eingesetzt (u. a. HTTrack[2] und Heritrix[3]), sodass aktuell ca. 75% der Speicherungen im Zielformat WARC[4] und die restlichen als offene Ressourcen in Verzeichnissen vorliegen. Beide Speicherungsformen werden ins Repositorium übernommen. Der zweite Teil besteht aus genuin digitalen Vor- und Nachlässen. Hierbei handelt es sich um Born-digitals[5], die dem DLA von Autorinnen und Autoren oder Institutionen zur Archivierung überlassen wurden. Diese

[1] SDC4Lit Homepage, `https://www.sdc4lit.de/`, letzter Abruf: 10.05.2021.

[2] Roche, Xavier et al., "HTTrack Website Copier 3.49-2",`https://www.httrack.com/`, letzter Abruf: 10.05.2021.

[3] "Heritrix3", Internet Archive, `https://github.com/internetarchive/heritrix3/wiki`, letzter Abruf: 10.05.2021.

[4] "The WARC Format 1.1", IIPC,`https://iipc.github.io/warc-specifications/specifications/warc-format/warc-1.1/`, letzter Abruf: 10.05.2021.

[5] Kramski, Heinz Werner, "Stichwort Born-digitals", `https://edlex.de/index.php?title=Born-digitals`, letzter Abruf: 10.05.2021.

Vor- und Nachlässe können alle technisch möglichen Datenträger, Dateiformate und Datenmengen enthalten und müssen mehrfach bearbeitet werden, um sie strukturiert und nutzbar im Repositorium ablegen zu können. Aktuell werden Born-digital Vor- und Nachlässe von etwa 75 Bestandsbildnern am DLA archiviert. Diese Vor- und Nachlässe befinden sich in unterschiedlichen Bearbeitungsstadien und stellen eine Gesamt-Datenmenge von ca. 2,8 TB, verteilt auf etwa 2000 Datenträger, dar.

Den dritten und bisher kleinsten Teil bilden die Computerspiele mit Bezug zur (deutschsprachigen) Literatur. Die Sammlung reicht hier von einfachen, textbasierten Spielen bis hin zu aufwändigen 3D-Adventures und weist somit auch hier eine weite Spanne von unterschiedlichen technischen Anforderungen, Datenformaten und Zugangsmöglichkeiten auf. Diese Breite an Sammlungsobjekten aus unterschiedlichen Zeiten, Quellen und Systemen stellt in mehrfacher Hinsicht eine große Herausforderung dar. Bereits die Vorbereitung dieser Daten für das Repositorium gestaltet sich aufgrund der Obsoleszenz von Datenträgern und -formaten, möglicher unlesbarer bzw. defekter Dateien, der Formatmigration etc. anspruchsvoll. Neben den technischen Herausforderungen erschwert die heterogene Datenlage auch die Auswahl einer geeigneten Repositoriumssoftware bzw. die Auswahl und Anwendung verschiedener Metadaten-Standards. Ein Ziel von SDC4Lit besteht darin, die beschriebenen Daten nutzbar und in ihrer Ästhetik authentisch für die Forschung und Vermittlung zu Verfügung zu stellen.

3 Architektur-Entwurf

Zu den Kernaufgaben des Projekts gehört der Aufbau einer Plattform, die archivierte digitale Objekte nicht nur passiv zur Verfügung stellt, sondern auch eine Interaktion mit ebendiesen Objekten erlaubt, um weitergehende Forschung und Analyse zu ermöglichen (Abb. 1). Im Gegensatz zu bisherigen generischen Lösungen wie [1], [8] wird in SDC4Lit ein disziplinspezifischer Ansatz mit starker Integration in die Forschungsinfrastruktur verfolgt.

Eine der Kernkomponenten der SDC4Lit-Architektur ist das Primärdaten-Repositorium, in dem alle Objekte der digitalen Literatur langfristig gespeichert werden. Dazu werden die Archivmaterialien zunächst entsprechend vorbereitet und in das Repositorium eingefügt. Kuratiert werden die Bestände des Primärdaten-Repositoriums ausschließlich von DLA-Mitarbeiterinnen und -mitarbeitern. Nutzerinnen und Nutzer des Repositoriums sollen gezielt nach Objekten suchen und mit gefundenen Objekten weiterarbeiten können.

Für die weitere Arbeit mit den über die Plattform zur Verfügung gestellten Archivalien ist der Aufbau einer Forschungs- und Analyseumgebung geplant, die niedrigschwellig zu bedienende digitale Analysemethoden und -werkzeuge zusammenführt, dokumentiert und in Form modularer Pipelines für die Forschung bereitstellt. Die Ergebnisse der Analysen können von den Nutzerinnen und Nutzern als Forschungsdaten in einem separaten Repositorium gespeichert werden, um sie für die weitere Nachnutzung zur Verfügung zu stellen.

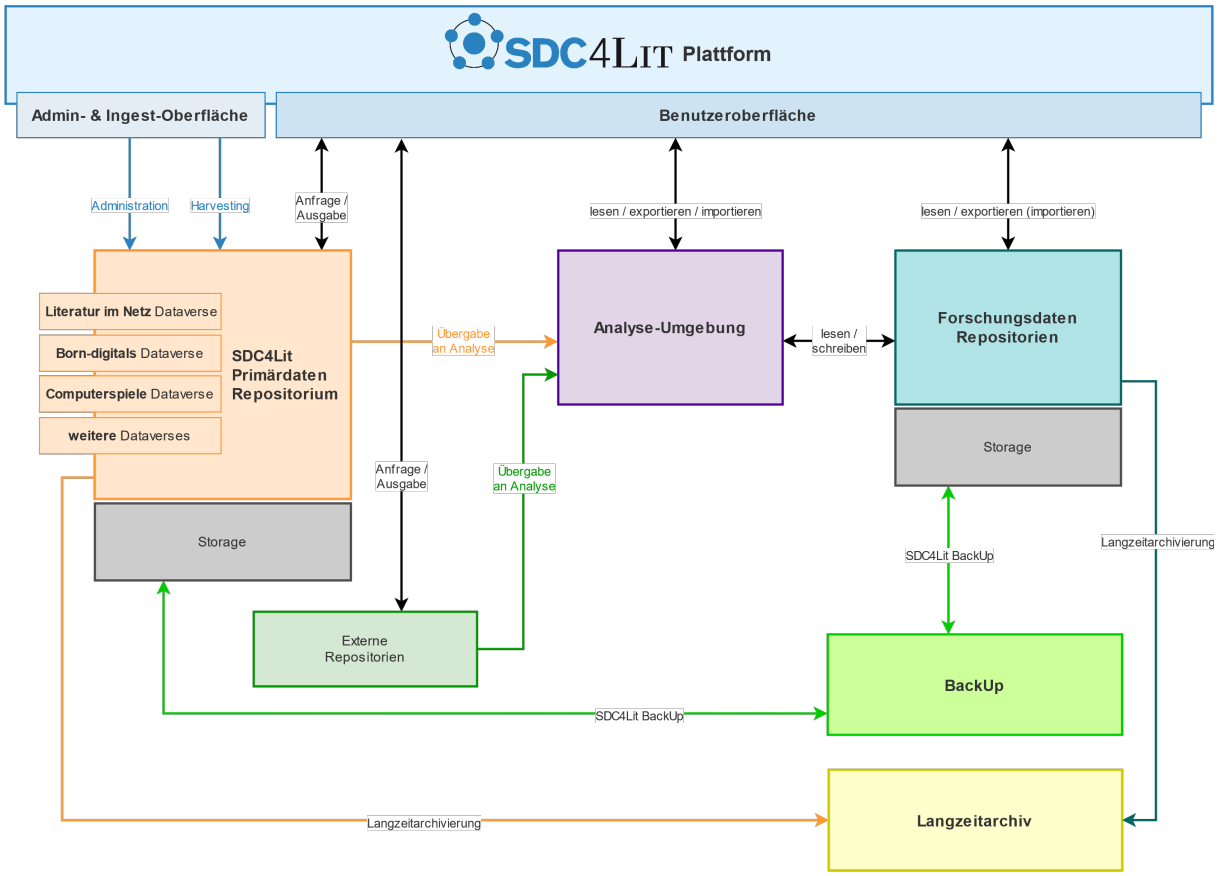

Abbildung 1: Entwurf der SDC4Lit-Architektur.

Um die Nachhaltigkeit und Sicherheit der Daten gewährleisten zu können, ist ein Tape-Backup für beide Repositorien sowie die Anbindung an ein Langzeitarchiv vorgesehen. Da sich die Frage der Nachhaltigkeit nicht allein in der Datenspeicherung erschöpft, sondern auch die Vermittlung von Methodenkompetenzen sowie die Rückkopplung mit den Bedürfnissen der Fachcommunity betrifft, verfolgt das SDC4Lit einen forschungs- und vermittlungsorientierten Ansatz, der auf Fallstudien zur (digitalen) Literatur basiert und in dessen Kontext bereits vorhandene literaturwissenschaftliche Methoden und Werkzeuge erprobt oder (weiter-)entwickelt werden. Neben verschiedenen anderen Fallstudien und methodischen Ansätzen wurden und werden beispielsweise verschiedene Entitätenerkenner hinsichtlich ihrer Verwendbarkeit für literarische Blogs und Zeitschriften getestet. [7] Zur Unterstützung der Erschließung und der textgenetischen Erforschung digitaler Vor- und Nachlässe werden Textähnlichkeitsmaße und Text-Reuse-Werkzeuge erprobt und weiterentwickelt. In einer Beta-Version zur Verfügung steht bereits ein in Zusammenarbeit mit Studierenden der Universität Stuttgart im Rahmen eines Projektseminars entwickeltes Software-Modul zur Erforschung nicht-linearer narrativer Strukturen von Netzlite-

ratur.[6] Dieses warc2graph[7] benannte Python-Modul, mit dem sich WARC- und Webobjekte graphbasiert modellieren lassen, wurde im Rahmen der E-Science-Tage 2021 in einem Tandem-Talk vorgestellt. Zusammen mit anderen bereitgestellten Methoden und Werkzeugen wird warc2graph in der späteren Projektphase in die Forschungsumgebung eingebunden und über die SDC4Lit-Plattform zur Verfügung stehen, um die im SDC4Lit-Primärdaten-Repositorium bereitgestellten, aber auch eigene Korpora analysieren zu können.

Die beschriebenen Infrastrukturkomponenten sollen über die gemeinsame SDC4Lit-Plattform bedient werden können, auf der alle Arbeitsschritte für die Nutzerinnen und Nutzer transparent und selbsterklärend aufbereitet werden. Geplant ist, zukünftig auch andere externe Primärdaten-Repositorien anzubinden und so verschiedene Communitys auf einer Plattform zusammenzubringen.

4 Herausforderungen

Die Obsoleszenz der Dateiformate und Datenträger, der große Umfang an Daten der digitalen Vor- und Nachlässe sowie die Diversität der Formate aller Primärmaterialien sind große Herausforderungen bei der Archivierung und Analyse der Daten. Eine daraus resultierende Kernaufgabe war die Auswahl einer geeigneten Repositoriumssoftware. Viele vorhandene Softwarelösungen zum Aufbau von Repositorien bringen zum Teil einen sehr unterschiedlichen Funktionsumfang mit. Um ein geeignetes Softwarepaket auswählen zu können, wurden anhand von User Storys ein Katalog von „Anforderungen an Repositorien“ formuliert und die einzelnen Anforderungen jeweils priorisiert und gewichtet. Anhand dessen wurden unterschiedliche Softwarelösungen, namentlich DSpace, Dataverse, MyCoRe, Fedora bzw. Islandora, Invenio und AtoM, in mehreren Schritten analysiert und evaluiert. Da keine der vorhandenen Open-Source-Produkte sämtliche unserer Anforderungen [2] vollumfänglich erfüllen konnte, müssen für bestimmte Funktionalitäten Eigenentwicklungen geleistet werden. Die Entscheidung fiel letztendlich auf Dataverse[8], da hier die meisten Kriterien erfüllt wurden und auf eigene Erfahrungen, beispielsweise aus dem Projekt DIPL-ING [7], zurückgegriffen werden kann.

Eine weitere Herausforderung, die derzeit im Fokus steht, ist die (Meta-)Datenmodellierung. Da bisher keine Standards existieren, die die SDC4Lit-Primärdatenbestände annähernd vollständig abdecken, werden in SDC4Lit eigene Datenmodelle entwickelt, die auf Vorarbeiten und Erfahrungen der Projektpartner basieren und den FAIR-Prinzipien

[6] Weitere Informationen sind dem Full Paper zum Vortrag zu entnehmen, das ebenfalls im Tagungsband der EST-21-Konferenz unter dem Titel ”Nicht-lineare narrative Strukturen in Netzliteratur: Speicherung und Nachnutzung von Forschungsdaten aus der computergestützten Extraktion von Verweisstrukturen in Hypertexten”veröffentlicht wird.

[7] Der Programm-Code von warc2graph wird auf `https://github.com/dla-marbach/warc2graph` veröffentlicht. Das Paket kann über den Python Package Index installiert werden: `https://pypi.org/project/warc2graph/`. Ein aktueller Snapshot (Version 0.1.1) ist über Zenodo verfügbar: DOI:10.5281/zenodo.4742254 (Hein et al. 2021 [6]).

[8] “The Dataverse Project”, Dataverse, `https://dataverse.org/`, letzter Abruf: 10.05.2021.

entsprechen [10]. Um Interoperabilität und Nachnutzbarkeit der Daten zu gewährleisten, orientiert sich SDC4Lit an Metadatenstandards wie METS [3], MODS [5] und PREMIS [4], die im Bereich des Bibliotheks- und Archivwesens etabliert sind.

Zurzeit[9] wird ein Repositoriumsprototyp auf Basis von Dataverse aufgesetzt und demnächst evaluiert. Im nächsten Schritt werden die ersten Daten in das Repositorium eingepflegt und für die weitere Bearbeitung bereitgestellt. Die Analysemethoden und -werkzeuge werden ausgewählt oder (weiter-)entwickelt und in die Forschungsumgebung integriert, sodass das Hauptziel von SDC4Lit – der Aufbau eines nachhaltigen Datenlebenszyklus für Literaturforschung und -vermittlung – langfristig umgesetzt werden kann.

Danksagung

SDC4Lit wird gefördert vom Ministerium für Wissenschaft, Forschung und Kunst in Baden-Württemberg.

Literaturverzeichnis

[1] Felix Bach, Björn Schembera, and Jos Van Wezel, “Design and Implementation of the first Generic Archive Storage Service for Research Data in Germany”, International Journal of Digital Curation 15.1 (2020).

[2] Felix Bach et al., “Kriterien für die Auswahl einer Softwarelösung für den Betrieb eines Repositoriums für Forschungsdaten”. Bausteine Forschungsdatenmanagement (`https://bausteine-fdm.de`) (eingereicht; Publikation vorauss. 2021).

[3] Linda Cantara, “METS: The metadata encoding and transmission standard”, Cataloging & classification quarterly 40.3-4 (2005): 237-253.

[4] Priscilla Caplan, “Understanding Premis”, Washington DC, USA: Library of Congress, 2009.

[5] Richard Gartner, “MODS: Metadata object description schema”, JISC Techwatch report TSW (2003): 3-6.

[6] Pascal Hein et al., „Warc2graph,“ Zenodo, 2021. DOI: `https://doi.org/10.5281/zenodo.4742254`, `https://zenodo.org/record/4742254`.

[7] Kerstin Jung et al., Workshop “Ensemble-Methoden aus menschlichen und maschinellen Bewertungen - Ein Entitäten Abstimmungsexperiment für 3 Tools und N Forschende”, `https://vdhd2021.hypotheses.org/197`, letzter Abruf: 10.05.2021.

[8] Angelina Kraft et al., “The RADAR Project – A Service for Research Data Archival and Publication”, ISPRS International Journal of Geo-Information 5.3 (2016): 28.

[9] Stand: April 2021.

[9] Björn Schembera et al., “Datenmanagement in Infrastrukturen, Prozessen und Lebenszyklen für die Ingenieurwissenschaften: Abschlussbericht des BMBF-Projektes Dipl-Ing”, (2019). https://doi.org/10.2314/KXP:1693393980.

[10] Mark D.Wilkinson et al., “The FAIR Guiding Principles for scientific data management and stewardship”, Scientific data 3.1 (2016): 1-9.