Berichte aus den DH-Förderprogrammen der Deutschen Nationalbibliothek

Maximilian Würz

Gattungsspezifische Metadaten identifizieren, extrahieren und weiterverarbeiten

Der bibliografische Gesamtbestand der DNB als Datenquelle am Beispiel der ›kurzen Erzählformen‹

Impressum

Herausgeberin:
Deutsche Nationalbibliothek (Leipzig, Frankfurt am Main)
Adickesallee 1
60322 Frankfurt am Main
Deutschland
www.dnb.de

Vertreten durch den Generaldirektor Frank Scholze

2025
ISSN 2944-0904
<urn:nbn:de:101-20250321156>

CC BY-SA 3.0 DE

Erstellt von Maximilian Würz

Die Arbeit wurde von der Deutschen Nationalbibliothek über das Forschungsstudienprogramm von HERMES – Humanities Education in Research, Data, and Methods gefördert. Das dieser Publikation zugrunde liegende Vorhaben wurde mit Mitteln des Bundesministeriums für Forschung, Technologie und Raumfahrt unter dem Förderkennzeichen 16DKZ2009 gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt bei der Autorin/beim Autor.

DEUTSCHE NATIONAL BIBLIOTHEK

HERMES
Humanities Education in Research, Data, and Methods

Gefördert durch:
Bundesministerium für Forschung, Technologie und Raumfahrt

Finanziert von der Europäischen Union
NextGenerationEU

Das HERMES-Forschungsstudienprogramm an der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek als eine der größten Gedächtnisinstitutionen Deutschlands bietet ihre Datenbestände und digitalen Sammlungen für Wissenschaft und Forschung sowie für experimentelles und kreatives Arbeiten an. Damit leistet sie ihren Beitrag zur Sicherung des archivierten Kulturerbe Deutschlands und Wahrung eines lebendigen Platzes in der Gesellschaft.

Am BMFTR-geförderten Projekt HERMES – Humanities Education in Research, Data, an Methods (Laufzeit 2023 bis 2026) ist die Deutsche Nationalbibliothek federführend beteiligt und fördert den Aufbau eines Datenkompetenzzentrums für die Geistes- und Kulturwissenschaften. Im Zentrum von HERMES steht der digitale Wandel, der in der Forschung zu einem veränderten Bedarf an Expertise im Umgang mit digitalen Kulturdaten führt. Über das HERMES-Forschungsstudienprogramm fördert die Deutsche Nationalbibliothek den Umgang mit und den Erwerb von Datenkompetenzen und unterstützt während der Projektlaufzeit bis zu vier Arbeiten jährlich. Für die Studien stellt die Deutsche Nationalbibliothek ihren reichhaltigen Datenbestand zur Verfügung, der von bibliografischen Metadaten bis hin zu digitalen Volltexten reicht.

Die Studien, die in einem Bearbeitungszeitraum von drei Monaten entstanden sind, wurden eigenständig bearbeitet. Sie zeugen von aktuellen Fragestellungen der Forschung und geben einen Einblick in die Herangehensweisen und Methoden, die in den Geisteswissenschaften angewandt werden. Die Projektergebnisse zeigen zudem die Herausforderungen im Umgang mit Daten auf, geben einen Eindruck von vorhandenen Datenkompetenzen und leisten einen Beitrag dazu, unterstützende Angebote für Wissenschaft und Forschung an deren Bedarfe adäquat anzupassen.

Der hier vorliegende Bericht bildet den Abschluss der jeweiligen Studienförderung durch das HERMES-Forschungsstudienprogramm. Die Gestaltung des Berichts ist den Geförderten freigestellt, um die individuellen Anforderungen an das Projekt berücksichtigen zu können. Die Berichte werden daher von der Deutschen Nationalbibliothek nicht redaktionell betreut und stellen eigenständige Werke in der alleinigen Verantwortung der Geförderten dar.

Gattungsspezifische Metadaten identifizieren, extrahieren und weiterverarbeiten: Der bibl. Gesamtbestand der DNB als Datenquelle am Beispiel der ›kurzen Erzählformen‹

Inhalt

Abstract I
1. Projektbeschreibung und Studiendesign 1
2. Das Fallbeispiel: Kurze Erzählformen im 21. Jahrhundert 3
2.1 Das Zielsystem 3
2.2 Der bibliografische Gesamtbestand als Datenquelle 4
2.3 Herausforderungen 6
2.4 Der Arbeitsprozess: Extract-Transform-Load 7
2.4.1 Identifikation und Extraktion 7
2.4.2 Bereinigung der Datensätze 15
2.4.3 Anreicherung und Import in das Zielsystem 20
3. Prozessreflexion und Generalisierung der Erkenntnisse 22
Literaturverzeichnis 25

Abstract

Die vorliegende Forschungsstudie untersucht, inwiefern sich der bibliografische Gesamtbestand der Deutschen Nationalbibliothek (DNB) als Ausgangspunkt eines Prozesses zur Identifikation, Extraktion und Weiterverarbeitung gattungsspezifischer Metadaten eignet. Unter Einbezug des Fallbeispiels der ›kurzen Erzählformen‹ wurde die Maßgabe formuliert, möglichst alle genuin deutschsprachigen Titel, die kurze narrative Prosagattungen enthalten und in einem bestimmten Beobachtungszeitraum (2008-2023) publiziert wurden zu identifizieren, die dazugehörigen Metadaten in einer festgelegten Form zu exportieren und in ein differentes Zielsystem (postgreSQL-Datenbank) zu integrieren. Der herausgearbeitete, technisch dokumentierte Arbeitsprozess umfasst im Wesentlichen drei Schritte: die Identifikation und Extraktion, die Bereinigung sowie die Anreicherung mit anschließendem Import der relevanten Datensätze. Im Zuge der Identifikation und Extraktion geraten verschiedene Online-Schnittstellen und Exportformate in den Fokus, wobei der Bezug gattungsspezifischer Metadaten im Exportformat MARC21-xml über die SRU-Schnittstelle als besonders nutzbringend erkannt wurde. Die Bereinigung widmet sich notwendigen formalen und inhaltlichen Anpassungen der exportierten Metadaten. Über die Kodierung und Ausführung mehrerer Python-Skripte werden die einzelnen Bereinigungsschritte umgesetzt und die dabei entfernten Datensätze zur weiteren Reflexion technisch in Exporttabellen dokumentiert. Wie sich zeigt, haben die von der DNB genutzten Gattungsbegriffe einen hohen Mehrwert als Instrument zur gattungsbezogenen Filterung, auch wenn aufgrund der abweichenden Konkretheit zwischen bibliografischer Klassifizierung und dem literaturwissenschaftlichen Gattungsverständnis manuelle Nacharbeiten erforderlich sein können. Im letzten Prozessschritt werden die in den bereinigten Metadaten enthaltenen Personen- und Körperschaftsnamen mit den Normdaten der GND abgeglichen, mit weiteren Daten angereichert und in das Zielsystem importiert. Abschließend erfolgt eine Generalisierung der im Zuge der Prozessreflexion erarbeiteten Erkenntnisse.

1. Projektbeschreibung und Studiendesign

Möchte man das wechselseitige Verhältnis von Literaturen und Öffentlichkeiten quantitativ untersuchen und dabei ergründen unter „welchen sozialen, politischen, ökonomischen und medialen Bedingungen“ bestimmte Gattungen entstehen oder welche potenziellen Effekte diese auf die sie umgebenden Umwelten haben können (vgl. Kley/Niefanger 2022), besteht eine wesentliche Aufgabe in der Beschaffung, Aufbereitung und Auswertung von gattungsspezifischen Daten. Je nach Ausrichtung der Analyse können unterschiedliche Informationen von Belang sein: Daten zu Artefakten, die die beobachteten Gattungen beinhalten, zu den Akteuren, die an deren Entstehung und Verbreitung beteiligt sind oder zu den medialen Kontexten, in welchen sie publiziert und rezipiert werden. Allein aus ökonomischen Gründen ist es sinnvoll bei solchen Vorhaben möglichst auf bestehende Datenquellen zurückzugreifen und zusammenzuführen, statt die benötigten Informationen in aufwendigen Verfahren eigenständig neu zu erheben.

Die vorliegende Forschungsstudie fokussiert auf den bibliografischen Gesamtbestand der Deutschen Nationalbibliothek[1] als potenzielle Datenquelle für gattungsspezifische Forschungsarbeiten und widmet sich dabei der Frage, wie gut dieser zur Identifikation, Extraktion und Weiterverarbeitung von Metadaten mit expliziten Gattungsbezügen geeignet ist. Anhand des Fallbeispiels der kurzen Erzählformen, welche schriftlich fixierte narrative Kurzprosa wie Kurzgeschichten und ihr nahverwandte Gattungen umfassen (vgl. Würz 2023), wird analysiert, wie effektiv relevante (d. h. die entsprechenden Textformen enthaltende) Manifestationen von Buchtiteln recherchiert, samt der dazugehörigen Titel- bzw. Metadaten separiert und in neue Zielsysteme integriert werden können.

Im Wesentlichen liegen der Forschungsstudie dabei mehrere Fragestellungen zugrunde, die sich in zwei Komplexe unterteilen lassen. Ersterer umfasst Fragen, die einen klaren Gattungsbezug aufweisen: Wie eignet sich der Gesamtdatenbestand der DNB als Datengrundlage für Untersuchungen zu spezifischen Gattungen oder Textsorten? Wie lassen sich relevante Artefakte effektiv recherchieren und filtern, wenn die gesuchten Gattungen in ihrer Konkretheit von den Gattungsbegriffen, die von der DNB verwendet werden, wie etwa „Erzählende Literatur: Gegenwartsliteratur ab 1945“ und „Erzählende Literatur: Anthologien“ (DNB 2025e) abweichen? Inwiefern lassen sich gattungsspezifische Fragestellungen an den Gesamtdatenbestand der DNB umsetzen, gerade auch mit Blick auf weniger etablierte Sammelbegriffe wie den hier verwendeten Ausdruck der ›kurzen Erzählformen‹?

Der zweite Fragenkomplex nähert sich dem Bestand dagegen aus technischer Richtung und versammelt Fragestellungen mit einem klaren Datenbezug: Wie lassen sich als relevant identifizierte Datenmengen aus dem Gesamtdatenbestand der DNB destillieren und in andere Systeme übertragen?

[1] Im Folgenden mit DNB abgekürzt.

Welche Zwischen- und Bereinigungsschritte vom Export der Daten bis zum Import in einen neuen Datenbankbestand sind notwendig? Wie hoch ist der Aufwand bei der Aufbereitung der exportierten Daten und wie weit lässt sich der Prozess automatisieren, um auch bei etwaigen Folgeuntersuchungen reliable, valide und objektive Ergebnisse zu erzielen?

Um diese Fragen im gesteckten Rahmen der Forschungsstudie gewissenhaft beantworten zu können, wurde folgendes Studiendesign erarbeitet: Zunächst wird das Fallbeispiel der ›kurzen Erzählformen‹ erläutert sowie das Zielsystem und die dafür benötigten Daten beschrieben. Demgegenüber wird der bibliografische Gesamtbestand der deutschen Nationalbibliothek als Datenquelle ausgeleuchtet, Diskrepanzen zwischen den Anforderungen des Zielsystems und den Gegebenheiten der Datenquelle aufgezeigt und in klar umrissene Herausforderungen übersetzt. In der Folge wird die eine Vorgehensweise zur Identifikation, Extraktion und Weiterverarbeitung relevanter Informationen in Form eines Extract-Transform-Load-Prozesses (kurz: ETL-Prozess) schematisch erarbeitet und konkret so weit umgesetzt, dass die Potenziale des Gesamtbestandes der DNB zur Identifikation, Extraktion und Weiterverarbeitung von Metadaten mit expliziten Gattungsbezügen überblicksartig erkennbar und mögliche Herausforderungen sowie verschiedene Praktiken im Umgang damit herausgestellt werden. Eng am Fallbeispiel orientiert wird der Versuch unternommen, alle genuin deutschsprachigen Buchtitel, die kurze Erzählformen enthalten und im festgelegten Beobachtungszeitraum (2008-2023) veröffentlicht wurden, über einen möglichst weit automatisierten Ablauf zu identifizieren, ausgewählte Titel- und Metadaten derselbigen zu extrahieren und in eine Relationale Datenbank zu integrieren. Dieser Prozess, welcher eine vergleichende Analyse des DNB-Katalogs[2] und der SRU-Schnittstelle[3] auf Ebene der Datenauslese und die Bereinigung mit Python ebenso umfasst, wie eine konzipierte Datenkonsolidierung und -angleichung an die Gemeinsame Normdatei durch den Einsatz des Open Source Tools Open-Refine[4], wird dabei technisch dokumentiert und fortwährend reflektiert. Abschließend nimmt sich die Forschungsstudie einer zusammenfassenden Prozessreflexion und der Generalisierung der am Fallbeispiel gewonnenen Erkenntnisse an.

[2] DNB 2025d. URL: https://portal.dnb.de/opac.htm.
[3] DNB 2025h. URL: https://www.dnb.de/DE/Professionell/Metadatendienste/Datenbezug/SRU/sru_node.html.
[4] GND. URL: https://lobid.org/gnd/reconcile/; OpenRefine Documentation 2025. URL: https://openrefine.org/docs/manual/reconciling.

2. Das Fallbeispiel: Kurze Erzählformen im 21. Jahrhundert

Das Fallbeispiel der ›kurzen Erzählformen‹ ist dem laufenden Dissertationsprojekt *Kurzes Erzählen im 21. Jahrhundert – Entstehungsbedingungen und Aneignungspotenziale kurzer Erzählformen in der Gegenwart*[5] entnommen. Dieses macht es sich zur Aufgabe, die grundlegenden Faktoren hinter der Genese kurzer Erzählformen sowie mögliche Effekte dieser Schreibweisen auf die sie umgebenden Umwelten zu untersuchen. Im Zuge dieses Projekts werden Hypothesen, welche ebensolche Faktoren und Effekte anhand eines Datenkorpus empirisch überprüft. Da bisher keine Bezugsquelle existiert, die explizit Metadaten zu kurzen Erzählformen, zu Artefakten, die diese beinhalten oder zu Akteuren, die an ihrer Entstehung beteiligt sind, umfassend in sich versammelt, stellt der Aufbau des Datenkorpus in Form einer Relationalen Datenbank einen wesentlichen Teil dieses Dissertationsprojekts dar. In der Art einer synkretistischen Zusammenführung finden Daten aus vier für kurze Erzählformen als besonders relevant eingestuften Publikationssegmenten (Bücher, Periodika, Literaturwettbewerbe, Social-Reading-Plattformen) Eingang in die geplante Datenbank. Nach der entsprechenden Aufbereitung wird es möglich, operationalisierte Hypothesen zu den Entstehungsbedingungen und Aneignungspotenzialen von kurzen Erzählformen mittels SQL-Abfragen datenbasiert zu überprüfen. Für die Umsetzung dieser Aufgabe ist der Gesamtbestand der bibliografischen Daten der Deutschen Nationalbibliothek von besonderer Relevanz, da sich die Erhebung der Metadaten zu den in Buchform publizierten kurzen Erzählformen in weiten Teilen darauf stützt.

Das Lemma ›kurze Erzählformen‹ kann als gattungsneutraler Sammelbegriff für schriftlich fixierte narrative Kurzprosa wie Kurzgeschichten und ihnen nahverwandten Gattungen verstanden werden. Folglich liegt das Interesse der vorliegenden Forschungsstudie darin zu analysieren, inwieweit sich der bibliografische Gesamtbestand der DNB dazu eignet, Titel zu identifizieren, die Kurzgeschichten, Anekdoten, Stories, kurze Erzählungen oder Geschichten enthalten, Metadaten zu diesen Titeln und den daran beteiligten Personen und Körperschaften zu extrahieren und in die angesprochene Relationale Datenbank zu integrieren.

2.1 Das Zielsystem

Um besser zu verstehen, welche Anforderungen an den Datenbestand der DNB gestellt werden, ist eine Beschreibung des Zielsystems, sowie der benötigten Informationen und Formate unerlässlich. Konkret handelt es sich bei dem Zielsystem, um eine relationale postgreSQL-Datenbank[6] (im Folgenden

[5] Würz 2022. URL: https://www.literaturundoeffentlichkeit.phil.fau.de/person/kurzes-erzaehlen-im-21-jahrhundert/.
[6] Release 18.0. Siehe: PostgreSQL 2025. URL: https://www.postgresql.org/.

mit RDB abgekürzt) mit dem Namen „kurzes_erzaehlen“. Diese wird auf einem eigens dafür eingerichteten Server aufgebaut und enthält Informationen zu kurzen Erzählformen, den Artefakten, in welchen diese publiziert werden und den Akteuren, die an ihrer Entstehung oder Verbreitung beteiligt sind. Im Wesentlichen sind es drei Entitäten in dieser RDB, die auf Metadaten aus dem Gesamtbestand der DNB angewiesen sind: die Tabellen ›Buch‹, ›Körperschaft‹, ›Person‹. Unter der Entität ›Buch‹ finden sich Informationen zu den Manifestationen von Buchtiteln, die kurze Erzählformen beinhalten. Ähnlich zu den Titeldaten, wie sie in bibliografischen Beständen bei der Titelaufnahme notiert werden, fallen darunter die ISBN, der Buchtitel, die Auflage des Buchs, der Publikationsort, die verlegende Person oder Körperschaft, das Publikationsjahr, die Seitenzahl sowie die Namen der weiteren an der Entstehung beteiligten Personen und Körperschaften in ihren Rollen als Verfasser:innen, Herausgeber:innen, Übersetzer:innen usw. In die Tabellen ›Person‹ und ›Körperschaft‹ fließen die Namen und weitere Informationen von Personen und Körperschaften, welche an der Entstehung und Verbreitung der relevanten Bücher beteiligt sind, ein. Hierzu gehören demografische respektive unternehmerische Angaben, wie das Alter, das Geschlecht, das Geburts- bzw. Gründungsdatum, ein etwaiges Sterbe- bzw. Auflösungsdatum, bekannte Alternativnamen, Websites und die GND-Nummer, also der eindeutige Identifikationsschlüssel in der Gemeinsamen Normdatei[7]. Damit die benötigten Daten in das Zielsystem integriert werden können, müssen sie als UTF-8-codierte CSV-Dateien vorliegen, die formal mit den Datencontainern der Tabellen ›Buch‹, ›Person‹ und ›Körperschaft‹ übereinstimmen. Sobald die Daten den Konfigurationen der einzelnen Entitäten und ihrer Spalten entsprechen, erfolgt der Import der Daten kommandozeilenorientiert über die Konsolenanwendung *psql* oder über die grafische Benutzeroberfläche des auf dem Clientrechner installierten Programms pgAdmin 4[8].

2.2 Der bibliografische Gesamtbestand als Datenquelle

Wie zu Beginn dargestellt, dient der bibliografische Gesamtbestand der DNB im eingebrachten Fallbeispiel als Datenquelle für die in Buchform publizierten kurzen Erzählformen, den Titeln, welche sie beinhalten und den Akteuren, die an deren Entstehung und Verbreitung beteiligt sind. Als „die zentrale Archivbibliothek Deutschlands“ sammelt die DNB „alle Medienwerke in Schrift, Bild und Ton, die seit 1913 in und über Deutschland oder in deutscher Sprache veröffentlicht werden. Ob Bücher, Zeitschriften, CDs, Schallplatten, Karten oder Online-Publikationen […] ohne Wertung, im Original und lückenlos“.[9] Dabei werden die einzelnen Objekte nicht nur gesammelt, sondern auch erschlossen, sodass sie auf Basis von Metadaten, mit welchen sie beschrieben wurden, im Datenbestand auffindbar sind.

[7] GND 2022. URL: https://gnd.network/Webs/gnd/DE/UeberGND/ueberGND_node.html.
[8] Version: pgAdmin 4 v8.9. Siehe: https://www.pgadmin.org/.
[9] DNB 2025a.

Diese „Erschließungsdaten“ umfassen im Fall der DNB sowohl formale als auch inhaltliche Informationen: „Die Formalerschließung ermöglicht die Suche nach einzelnen Medienwerken – mit Angaben wie Personen, Titel, Verlag, ISBN, ISSN, URN, DOI. Die Inhaltserschließung strukturiert unsere Bestände thematisch mit Schlagwörtern, DDC-Sachgruppen, DDC-Notationen und Gattungsbegriffen“ (DNB 2025b). Somit kann angenommen werden, dass der Gesamtbestand der DNB einen adäquaten Ausgangspunkt für quantitativ ausgerichtete Forschungsarbeiten, die versuchen, sich einer gattungsspezifischen Grundgesamtheit anzunähern und über die Auswertung von Metadaten Erkenntnisse zu den Entstehungsbedingungen und Aneignungen dieser Gattung(en) zu erlangen, darstellt. Die Bereitstellung der Daten in verschiedenen Formaten und zur freien Nutzung ist dabei ein weiterer Vorteil:

> Alle Titeldaten der Deutschen Nationalbibliothek und Normdaten der Gemeinsamen Normdatei sind kostenfrei unter „Creative Commons Zero”-Bedingungen (CC0 1.0) zur freien Nachnutzung verfügbar. Norm- und Titeldaten können in folgenden Formaten bezogen werden: MARC 21, MARC21-XML und RDF (verschiedene Serialisierungen). Nur für Titeldaten sind folgende Formate geeignet csv (csv_title) und DNB Casual (oai_dc) (DNB 2025d).

Bezogen werden können diese Metadaten über verschiedene Online-Schnittstellen (vgl. DNB 2025g), wobei im Wesentlichen zwei Bezugswege für das bearbeitete Fallbeispiel in Frage kommen: einerseits der im DNB-Portal eingebettete Katalog inklusive der Möglichkeit des Datenexports über den sogenannten Datenshop sowie andererseits die SRU-Schnittstelle. Der Katalog bietet registrierten Nutzer:innen die Möglichkeit Informationen zu über 53,1 Millionen Medienwerken zu recherchieren und die Metadaten zur Weiterverarbeitung über den Datenshop zu exportieren (vgl. DNB 2025i). Auf diese Weise können Datensätze zu einzeln ausgewählten Manifestationen ebenso heruntergeladen werden, wie ganze Trefferlisten nach Suchanfragen im Katalog, wobei eine breite Auswahl an Exportformaten zur Verfügung steht, jedoch mit Limitationen hinsichtlich der Gesamtmenge an Datensätzen pro Export umgegangen werden muss: „Die maximale Anzahl für eine Auswahl aus einzelnen Datensätzen beträgt 200 Treffer, die Größe einer kompletten Trefferliste einer einzelnen Suchanfrage darf 10.000 Datensätze nicht überschreiten“ (DNB 2025d).

Auch die SRU-Schnittstelle kann für die gezielte Recherche und Entnahme von Titel- und Normdaten verwendet werden. Mittels Search/Retrieve via URL (SRU) können Anwender:innen „gezielt in den Daten der Deutschen Nationalbibliothek suchen und diese anschließend in die eigene Umgebung übernehmen. Die Ergebnisse der Suche stehen in einem definierten XML-Format zur Verfügung“ (DNB 2025e), wie etwa MARC21, MARC21plus, RDF/XML oder MODS-xml. Ein direkter Export der Ergebnisse im CSV-Format ist zwar nicht möglich, allerdings lassen sich aus den verfügbaren XML-Dateien in einem Extraschritt – etwa durch die Ausführung eines Python-Skripts – relativ unkompliziert CSV-Dateien erzeugen. Die hier eingesetzte Retrievalsprache ist CQL (Contextual Query Language), wobei die Syntax

der Suchanfrage in eine URL eingepflegt und über einen beliebigen Webbrowser ohne weitere Zugangsvoraussetzungen abgeschickt werden kann. Somit ist es möglich Abfragen über die SRU-Schnittstelle an automatisierte Prozesse anzubinden und Datenexporte ohne Mengenlimitationen umzusetzen.

2.3 Herausforderungen

Die Anforderungen des Zielsystems und die Ausrichtung des Forschungsinteresses geben vor, welche Daten bezogen werden müssen. Die Gegebenheiten der Datenquelle bestimmen indes inwieweit der Zugriff auf die gesuchten Informationen möglich ist und welche Optionen es für den Bezug und die Weiterverarbeitung dieser Daten gibt. Für Forschende ergeben sich hieraus verschiedene Herausforderungen, die von der Wahl der passenden Schnittstelle und des adäquaten Exportformats bis zur Identifikation der relevanten Datensätze durch die Formulierung einer zutreffenden Syntax bei den Suchanfragen reicht. Die zu Grunde liegende Frage lautet dabei: Wie lässt sich ein möglichst effizienter Prozess entwickeln, in dem Sinne, dass die Forschungsinteressen adäquat bearbeitet, Aufwände in Grenzen gehalten und dabei die wissenschaftlichen Gütekriterien der Objektivität, Reliablität und Validität eingehalten werden? Im Nachfolgenden finden diese Herausforderungen in knapper Form für den konkreten Fall des Beispiels ihre Ausformulierung, bevor in den Folgekapiteln der daraus hervorgegangene Arbeitsprozess näher vorgestellt wird.

Die wohl grundlegendste Herausforderung stellt die Identifikation relevanter Titel dar. Wie wird es möglich in der Datenquelle *all* und gleichzeitig *nur* jene Datensätze zu identifizieren, die von Relevanz für die geplante Untersuchung sind? Hierfür gilt es nicht bloß einzelne Relevanzkriterien zu formulieren, wie etwa den Beobachtungszeitraum, die obligatorisch enthaltene Gattung, die Materialart oder die Sprache der Interessensobjekte, sondern auch Identifikationsmechanismen zu erarbeiten mit welchen diese Relevanzkriterien möglichst effektiv in die Suchanfragen bei der Recherche eingeflochten werden können. Zwei Ansätze bilden hier die Pole: erstens, ein möglichst weit gefasstes, d. h. nur geringfügig eingeschränktes Suchvorgehen, das im Sinne des eines „Catch-All“-Prinzips erst einmal alle möglicherweise relevanten Titel abgreift, um danach in verschiedenen Bereinigungsschritten irrelevante Datensätze auszusieben oder zweitens, eine sehr starke Ausformulierung und Konkretisierung der Suchanfragen, die weitaus weniger redundante Datensätze einbezieht, im Ergebnis jedoch möglicherweise mit hohen Verlusten von relevanten Titeln, welche nicht den exakten Suchkriterien entsprachen, einhergeht. Im Fallbeispiel der kurzen Erzählformen steht besonders die Frage im Vordergrund, wie sich Buchtitel, die bestimmte Gattungen enthalten, zuverlässig identifizieren lassen und ob es generell möglich ist, zu gewährleisten, dass alle Titel, die diesem gattungsspezifischen Kriterium entsprechen, aufgefunden werden. Dabei kommt erschwerend hinzu, dass es sich bei den kurzen

Erzählformen um keinen etablierten Sammelbegriff handelt, hier also zusätzlich gattungstheoretische Vorarbeit geleistet werden muss (vgl. Würz 2023).

Neben der Identifikation der relevanten Titel stellt das Auffinden der vom Zielsystem benötigten Informationen in der Datenquelle eine weitere wesentliche Herausforderung dar. Hierbei stehen Entscheidungen hinsichtlich der passenden Schnittstelle und des richtigen Exportformats an. Die Auswahl sollte so getroffen werden, dass die Schnittstelle den Bezug eines Datenformats ermöglicht, das die konkret gesuchten Informationen enthält und möglichst kompatibel mit dem Zielsystem bzw. dem weiteren Verarbeitungsprozess ist. Zusätzlich spielt in diesem Zusammenhang der benötigte Automatisierungsgrad eine Rolle und damit die Frage, ob es sich um einen einmaligen Datenexport handelt oder die Daten möglicherweise wiederholt und in regelmäßigen Abständen über den festgelegten Prozess bezogen werden sollen. Nicht zuletzt sollte der ausgearbeitete Prozess mit möglichst geringem Aufwand nachvollziehbar und reproduzierbar sein, also bei gleicher Umsetzung, zu gleichen Ergebnissen führen und eine hohe Validität hinsichtlich der gestellten Forschungsfragen aufweisen.

2.4 Der Arbeitsprozess: Extract-Transform-Load

ETL-Strecken können von einfachen und einmaligen Arbeitsprozessen bis hin zu periodisch wiederholten und langfristig implementierten Systemen verschiedene Komplexitäten annehmen. Dabei ist der grundsätzliche Nutzen stets der gleiche: die Extraktion von Daten aus einer einzelnen oder mehreren, heterogenen Datenquellen, die Transformation der extrahierten Daten in eine vom Zielsystem benötigte Form, um sie konsistent und nutzbar zu machen sowie der Upload dieser Daten in ebenjenes Zielsystem (vgl. Albrecht 2013, S. 1; Vaisman/Zimányi 2014, S. 285; Kramer 2019, S. 136; Seenivasan 2023, S. 40). Zumeist ist dabei ein intensiver Identifikationsprozess der benötigten Informationen vorgelagert, wie auch im hier bearbeiteten Fallbeispiel. In den nachfolgenden Unterkapiteln wird diese erweiterte ETL-Strecke schrittweise erläutert und auf die dabei verwendeten Skripte[10] verwiesen.

2.4.1 Identifikation und Extraktion

Entsprechend des Fallbeispiels werden all jene Titel im bibliografischen Gesamtbestand der DNB gesucht, die kurze Erzählformen enthalten und in Buchform auf deutscher Sprache im Beobachtungszeitraum 2008-2023 erschienen sind. Als kurze Erzählformen gelten jene schriftlich fixierten Gattungen, die gleichzeitig kurz, narrativ und prosaisch sind. Dies können Kurzgeschichten, Stories oder Anekdoten, aber auch mit weniger konkreten Gattungsbezeichnungen belegte Texte wie Geschichten oder kurze Erzählungen sein. Die DNB vergibt in ihrem Gesamtbestand literarische Gattungsbegriffe „für alle Titel der Belletristik und der Kinder- und Jugendliteratur“ (DNB 2025h). Die Grundlage ist dafür

[10] Die Skripte wurden im GitHub-Repository „kurze-Erzaehlformen“ hinterlegt. Siehe Würz 2025. URL: https://github.com/maximilianwuerz/kurze-Erzaehlformen/tree/main/Python-Skripte.

eine Liste von derzeit (Stand: 01. Oktober 2025) 90 festgelegten Gattungsbegriffen, die an die „Warengruppen des Verzeichnisses für lieferbare Bücher angeglichen“ (DNB 2025c) sind. Die Zusammenstellung der Gattungsbegriffe folgt dabei keiner einheitlichen Logik, sodass neben primär inhaltlich definierten Genrebegriffen, etwa zu den Subgenres „New Adult“ und „Young Adult“ auch zwei der drei weitergefassten und primär formal definierten Hauptbegriffe der „Naturformen“ (Goethe[11]) bzw. der klassischen Gattungstrias[12], „Lyrik“ und „Dramatik“ darin vorkommen. Diese „Uneinheitlichkeit der Differenzierungskriterien“ (Zymner 2003, S. 105) ist in der Literaturwissenschaft kein neues Phänomen, sondern vielmehr die Regel (vgl. Neumann/Nünning 2007, S. 2), was auch daran liegt, dass Gattungen „keine vorgängigen Entitäten sind, die einfach existieren, sondern als Produkte oder, noch stärker, als *Konstrukte* spezifischer Praktiken zu begreifen sind“ (Gencarelli 2024, S. 6). In der Klassifikationspraxis der DNB finden für den Bereich der „Erzählende[n] Literatur“ fünf hierarchisch untergeordnete Begriffe Verwendung: „Erzählende Literatur: Hauptwerk vor 1945“, „Erzählende Literatur: Gegenwartsliteratur ab 1945“, „Historische Romane und Erzählungen“, „Erzählende Literatur: Anthologien“ sowie „Romanhafte Biographien“ (DNB 2025i). Neben der Nutzung dieser, bezogen auf narrative Literaturen sehr weit gefassten Gattungsverweise, können Manifestationen, die bestimmte Gattungen enthalten, im Gesamtbestand der DNB auch über die Nutzung von Suchbegriffen und -schlüsseln recherchiert werden. Wie effektiv dies tatsächlich ist, wird bei der Umsetzung des nachfolgenden Arbeitsprozesses anhand von manuellen Analysen der erhaltenen Sucherergebnisse und der Erstellung von Exporttabellen im Zuge der einzelnen Bereinigungsschritte dokumentiert und reflektiert. Das Vorgehen beginnt dabei mit zwei Prozesszweigen, deren Vor- und Nachteile gegeneinander abgewogen werden: Erstens, der Recherche im Online-Katalog der DNB inklusive des Exports über den daran angebundenen Datenshop und zweitens, der Identifikation und Extraktion über die SRU-Schnittstelle. Darüber hinaus werden die beiden Exportformate CSV und MARC21-xml in die vergleichende Analyse einbezogen, und dabei das Für und Wider der Nutzung dieser Dateien im Fall von gattungsbezogenen Fragestellungen taxiert.

2.4.1.1 Identifikation über den DNB-Katalog und die SRU-Schnittstelle

Die gesuchten Medienwerke müssen mehrere Relevanzkriterien erfüllen. Zum einen muss es sich um Bücher handeln, die auf deutscher Sprache im Beobachtungszeitraum 2008-2023 erschienen sind. Zum anderen müssen sie kurze narrative Prosa enthalten. Um nun nur solche Artefakte im Gesamtbestand der DNB zu identifizieren, die diesen Kriterien entsprechen, müssen zweckgemäße Suchanfragen formuliert werden. Sowohl die Expertensuche des DNB-Katalogs als auch die SRU-Schnittstelle setzen eine

[11] Goethe, Johann Wolfgang von: Noten und Abhandlungen zum besseren Verständniß des West-östlichen Divans. In: Werke, Bd. 3/1. Hgg. von Hendrik Birus. Frankfurt am Main 1994. Hier S. 206f.
[12] Siehe zur Herausbildung der „Dreiteilung der Dichtung in Epik, Lyrik und Drama“ Zymner 2003, S. 14-32.

Suchanfrage in der Syntax der Retrievalsprache CQL voraus, auch wenn sich die konkreten Ausformulierungen aufgrund der differenten technischen Voraussetzungen der Zugriffsmethoden über die Suchmaske des Katalogs und via URL unterscheiden. Wie bei der Vorstellung des Fallbeispiels bereits ausgeführt, weisen die typologischen Merkmale kurz, narrativ und prosaisch auf die Kurzgeschichte und ihr nahverwandte Gattungen hin. In der bisherigen Theorie wurde die Kurzgeschichte wiederholt mit Anekdoten, Kalendergeschichten, Erzählungen, Novellen, Short Stories und Geschichten verglichen und vielfältige Versuche unternommen, sie gegen dieselben abzugrenzen (vgl. Doderer 1980; Durzak 1994, 1980/2002; Marx 2005; Meyer 2014). Während Novellen in der Regel eine mittlere Länge aufweisen und somit nicht das Relevanzkriterium der Kürze erfüllen und im engeren Sinn verstandene Kalendergeschichten gegenwärtig keine größere Rolle mehr in der literarischen Produktion einnehmen, werden die verbliebenen Gattungsbeispiele als relevant für das Fallbeispiel der kurzen Erzählformen eingestuft. Da die von der DNB verwendeten Gattungsbegriffe für erzählende Literatur wahlweise zu weit gefasst, dabei das Kriterium der Kürze nicht berücksichtigen und somit Romane neben kurzen Erzählformen versammeln[13] oder aber zu spezifisch sind und damit eine Vielzahl relevanter Sammelbände ausschließen[14] und zudem nicht direkt in Suchanfragen integriert werden können, eignen sich diese lediglich zur Filterung unpassender Datensätze und zur Analyse in einem späteren Schritt, nicht aber zur beginnenden Recherche im Gesamtbestand. Für die zu formulierende Suchanfrage bedeutet dies, dass mehrere selbstgewählte Gattungsbegriffe in die Suchanfrage eingehen, wobei die Entscheidung auf die Bezeichnungen „Kurzgeschichten“, „Stories“, „Erzählungen“, „Geschichten“, „Anekdoten“ und „Anthologie“ fiel. Unter Einbezug dieser Begriffe und des Indexkürzels „tit“ in die Suchanfrage wird es möglich, alle Titelfelder der im Gesamtbestand enthaltenen Datensätze zur durchsuchen. Hierzu gehören „zum Beispiel Haupttitel, Titelzusätze, Serientitel und Titel von Werken“ (DNB 2022). Im Fall der kurzen Erzählformen enthalten diese Felder in der Regel eine der oben genannten Gattungsbezeichnungen, weshalb sie sich für die Identifikation relevanter Artefakte im Besonderen eignen. Hinzu kommen weitere notwendige Einschränkungen hinsichtlich der Materialart, der Sprache, des Beobachtungszeitraums und der Sachgruppe. Um lediglich Bücher einzubeziehen, wird der Index *mat=books* verwendet, zur Einschränkung der Sprache auf deutschsprachige Titel *spr=ger*, für die Einschränkung des Veröffentlichungszeitraum and *jhr>2007 and jhr<2024* und für die Filterung auf belletristische Erzeugnisse, der Sachgruppenindex *sgt=B*. In vollständiger Form wird bei der Recherche im Online-Katalog der DNB demnach die Expertensuche

[13] Dies war bei einer stichprobenartigen Untersuchung von Suchergebnissen vor allem bei Gattungsbegriffen „Erzählende Literatur“, „Erzählende Literatur: Hauptwerk vor 1945“, „Erzählende Literatur: Gegenwartsliteratur ab 1945“, „Historische Romane und Erzählungen“ zu beobachten.

[14] Dies gilt für den DNB-Gattungsbegriff „Erzählende Literatur: Anthologien“. Eine stichprobenartige Überprüfung ließ erkennen, dass eine Vielzahl von relevanten Sammelbänden andere Gattungsbezeichnungen aufwies, auch wenn die literaturwissenschaftliche Unterscheidung zwischen Anthologie und Sammelband nicht konsistent eingehalten wurde.

tit all "Kurzgeschichte*" or tit all "Geschichten" or tit all "Erzählungen" or tit all "Stories" or tit all "Anekdoten" or tit all "Sammelband" or tit all "Anthologie" and jhr>2007 and jhr<2024 and spr all "ger" and mat all "books" and sgt all "B"

verwendet. Im Fall der SRU-Schnittstelle wird diese Syntax in eine Anfrage mit URL-Kodierung übersetzt:

https://services.dnb.de/sru/dnb?version=1.1&operation=searchRetrieve&query=TIT%3DKurzgeschichte*%20or%20TIT%3DGeschichten%20or%20TIT%3DErz%C3%A4hlungen%20or%20TIT%3DStories%20or%20TIT%3DAnekdoten%20or%20TIT%3DSammelband%20or%20TIT%3DAnthologie%20and%20sgt=B%20and%20spr=ger%20and%20(jhr%3E=2008%20and%20jhr%3C=2023)%20and%20mat=books&recordSchema=MARC21-xml[15].

Auf diese Weise werden über beide Schnittstellen die gleichen 22.976 Datensätze identifiziert[16], welche voraussichtlich Manifestationen mit kurzen Erzählformen beschreiben, damit eine Relevanz für die Forschungsfragen aufweisen und in der Folge zur weiteren Auswertung exportiert werden müssen. Im Online-Katalog der Deutschen Nationalbibliothek ist es über den DNB-Shop möglich, Datensätze zu den erhaltenen Trefferlisten herunterzuladen. Wie in Kapitel 2.2 angerissen, liegt die downloadbare Höchstzahl bei 10.000 Treffern, sodass die ursprüngliche Suchanfrage mit 22.976 Ergebnissen auf mindestens drei Anfragen aufgeteilt werden muss, damit die Maximalmenge nicht überschritten wird. Im konkreten Fall der oben formulierten Anfrage bietet es sich an, die Aufteilung anhand des Zeitraums vorzunehmen und diesen in drei etwa gleich große Intervalle zu gliedern (z.B. 2008-2013; 2014-2019; 2020-2023). Dies ermöglicht den Bezug aller identifizierten Datensätze, hat jedoch zur Folge, dass die exportierten Daten nach dem Export entweder manuell oder per Skript zusammengeführt werden müssen. Dieser Schritt kann unter Umständen und je nach Dateiformat Probleme bereiten, etwa wenn es bei MARC21-xml-Daten die verzeichneten Werte im Datenfeld ›numberOfRecords‹ nicht mehr mit der tatsächlichen Gesamtzahl übereinstimmen oder sich die ›recordPosition‹-Tags in der konsolidierten Datei wiederholen. Im Gegensatz dazu erlaubt die SRU-Schnittstelle den automatisierten Download der kompletten Trefferliste. Zwar beträgt die Anzahl der gelieferten Datensätze standardmäßig nur 10 und maximal 100 Datensätze pro Antwort, sodass in der Browseransicht nicht mehr Treffer gleichzeitig einsehbar sind, allerdings können weitere Datensätze über den Einbau der Logik &startRecord=“ in die Anfrage abgerufen werden (vgl. DNB 2025h). Dies bietet den Vorteil, dass die Übernahme der Datensätze in die eigene Arbeitsumgebung direkt in ein Skript integriert werden und so

[15] Für Informationen zu den Formulierungsstandards der SRU-Schnittstelle siehe: https://www.dnb.de/DE/Professionell/Metadatendienste/Datenbezug/SRU/sru_node.html.
[16] Stand: 30.10.2025.

automatisiert erfolgen kann. Auf diesem Weg ist es jedoch nicht möglich die Daten im vom Zielsystem eingeforderten Format CSV zu exportieren, sodass die relevanten Informationen aus einer der downloadbaren XML-Dateien extrahiert bzw. die XML-Dateien in das CSV-Format konvertiert werden müssen. Die Frage, welche Schnittstelle gewählt werden sollte, hängt also von der Menge an Suchtreffern, der Teilbarkeit der Suchanfrage und von dem Grad der gewünschten Prozessautomatisierung (und damit der zu erwartenden Wiederholungsrate des Extraktionsprozesses) ab. Daneben spielt das benötigte Dateiformat eine nicht unerhebliche Rolle, was direkt zu den Ausführungen im nachfolgenden Kapitel führt.

2.4.1.2 Extraktion in den Formaten CSV und MARC21-xml

Im Zuge der „Extraktionsphase […] werden alle benötigten Daten aus den verschiedenen internen und externen Datenquellen geholt“ (Kramer 2019, S. 136). Auch wenn darauf in der Regel noch eine Phase der Bereinigung und Anreicherung folgt, ist es von Vorteil bereits hier auf eine hohe Kompatibilität der akquirierten Daten zu achten, um die Folgeaufwände möglichst gering zu halten. Damit die Kompatibilität der beiden Exportformate für das vorliegende Fallbeispiel eingeordnet werden kann, ist es sinnvoll noch einmal die vom Zielsystem benötigten Informationen und Datentypen aufzulisten. In den Tabellen 1-3 sind die einzelnen Entitäten der Relationalen Datenbank, die darin enthaltenen Spalten samt den Datentypen und einer kurzen Beschreibung der Inhalte aufbereitet. Die Entität ›Buch‹ beinhaltet in der Hauptsache Metainformationen zu den in der Tabelle aufgenommenen Buchtiteln. Die Entität ›Person‹ umfasst verschiedene Informationen zu den Personen, die an der Entstehung der relevanten Artefakte beteiligt sind und die Entität ›Körperschaft‹ versammelt selbiges für die entsprechenden Körperschaften. Während letztere Tabellen lediglich die Personen- bzw. Körperschaftsnamen

Tabelle	Spalte	Datentyp	Inhalt
Buch	ISBN	varchar	Internationale Standardbuchnummer, wie im Katalog vermerkt.
	Buchtitel	varchar	Der Hauptsachtitel des Buches.
	Auflage	varchar	Information zur Auflage des Buches.
	Publikationsort	varchar	Der Ort, an dem das Buch veröffentlicht wurde.
	Publisher	varchar	Der Name der verlegenden Instanz des Buches.
	Publikationsjahr	integer	Das Jahr der Veröffentlichung des Buches.
	Seitenzahl	integer	Die Anzahl der Seiten des Buches.
	Verfasser	varchar	Die Namen der Autor/innen.
	Herausgeber	varchar	Die Namen der Herausgeber/innen.
	Übersetzer	varchar	Die Namen der Übersetzer/innen.
	keineRolle	varchar	Die Namen der Personen, ohne spez. Rollenzuweisung.
	sonstigeRolle	varchar	Die Namen der Personen, mit sonstigen Rollen.

Tabelle 1: Metainformationen zur Entität ›Buch‹ in der Relationalen Datenbank des Fallbeispiels.

Tabelle	Spalte	Datentyp	Inhalt
Person	Person_ID	varchar	Eindeutiger Identifikator der Person.
	Person_Name	varchar	Der Name der Person.
	Person_NameGND	varchar	Der standardisierte Name der Person aus der GND.

	Beruf_Beschäftigung	text	Details zu Beruf oder Beschäftigung der Person gemäß GND.
	Geburtsdatum	varchar	Das Geburtsdatum der Person gemäß GND.
	Sterbedatum	varchar	Das Sterbedatum der Person gemäß GND (falls vorhanden).
	Geschlecht	varchar	Das Geschlecht der Person gemäß GND.
	GND_Nummer	varchar	Die zugehörige GND-Nummer der Person.
	Person_Homepage	varchar	Homepage oder persönliche Website der Person.

Tabelle 2: Metainformationen zur Entität ›Person‹ in der Relationalen Datenbank des Fallbeispiels.

Tabelle	**Spalte**	**Datentyp**	**Inhalt**
Körperschaft	Körperschaft_ID	varchar	Eindeutiger Identifikator der Körperschaft.
	Körperschaft_Name	varchar	Der Name der Körperschaft.
	KörperschaftNameGND	varchar	Der Name der Körperschaft gemäß GND.
	Sitz	varchar	Der Hauptsitz der Körperschaft gemäß GND.
	Gründungsdatum	varchar	Datum der Gründung der Körperschaft gemäß GND.
	Auflösungsdatum	varchar	Datum der Auflösung gemäß GND (falls zutreffend).
	GND_Nummer	varchar	Die zugehörige GND-Nummer der Körperschaft.
	Homepage	varchar	Website der Körperschaft.

Tabelle 3: Metainformationen zur Entität ›Buch‹ in der Relationalen Datenbank des Fallbeispiels.

aus den Datensätzen des Gesamtbestandes der DNB übernehmen und mit weiteren Informationen aus den Normdaten der der Gemeinsamen Normdatei (GND) sowie aus anderen Datenquellen vereinen[17], umfasst die Tabelle ›Buch‹ Informationen, die sich in den Titeldaten der jeweiligen Manifestationen finden lassen. Über den Datenshop der DNB besteht die Möglichkeit „Trefferlisten oder individuelle Selektionen von Titeldaten aus dem Katalog der Deutschen Nationalbibliothek im CSV-Format (csv_title) herunterzuladen“ (DNB 2024). Dabei werden „vorhandene Normdaten automatisch entfernt und somit nur die Titeldaten geliefert“ (ebd.). Diese reduzierten Dateien bieten den Vorteil, dass sie im vom Zielsystem benötigten Dateiformat vorliegen und aufgrund der geringen Zahl von Datenfeldern weitaus übersichtlicher sind. Die Informationen für die Spalten der Entität ›Buch‹ finden sich in der CSV-Exportdatei in folgenden Feldern: die ISBN in *ISBN*, der Buchtitel in *title*, die Auflage in *edition*, der Publikationsort und der Publisher in *publisher*, das Publikationsjahr in *year*, die Seitenzahl in *format* und die Namen und Rollenzuordnungen für die Spalten Verfasser, Herausgeber, Übersetzer, keineRolle und sonstigeRolle im Datenfeld *creator*. Alle weiteren exportierten Datenfelder der CSV-Datei weisen hinsichtlich der Forschungsfragen keinen Nutzen auf und können somit entfernt werden. Zeichengetreu übernehmen lassen sich die Informationen aus den Datenfeldern *ISBN*, *edition* und *year*. In den anderen Datenfeldern finden sich dagegen Abweichungen, in Form von redundanten Zeichenketten, falschen Datentypen oder Vermischungen benötigter Informationen, welche folglich entfernt, umgewandelt oder logisch aufgeschlüsselt werden müssen. So beinhaltet die Spalte *title* neben dem benötigten Hauptsachtitel noch Zusätze, Paralleltitel sowie Verfasserangaben (vgl. DNB 2024) und die Spalte *format* eine Angabe zur physischen Einheit, in welcher der Umfang angegeben wird (z. B. Seiten oder

[17] Siehe Kapitel 2.4.3.

Blätter) sowie teilweise exakte Gewichts- und Größenangaben. In den anderen Spalten sind indes wichtige Informationen für die Entität ›Buch‹ in vermischter Form zu finden, wie etwa der Verlagsort und der Verlagsname separiert durch die Zeichenkette „ : “ in der Spalte *publisher.* Auch sind die Informationen für die Rollenspalten (Verfasser, Herausgeber etc.), also die „Namen der Verfasser, Körperschaften und sonstigen beteiligten Personen und Körperschaften in Ansetzungsform“ (DNB 2024) in der Spalte *creator*, mit dem Verweis auf ihre jeweilige Rolle in eckigen Klammern dahinterstehend und vermischt, zu finden, sodass die Namen in einem aufwändigen Verfahren separiert und entsprechend der Rollenverweise korrekt auf die jeweiligen Spalten verteilt werden müssen. Der Umgang mit den CSV-Titeldaten fordert demnach bereits im Zuge der Extraktion Anpassungen der Inhalte ein, die in ihrer Komplexität von geringen Adaptionen, wie dem Entfernen einzelner Zeichenketten, bis zu Trenn- und Verteilungsfunktionen, die aufwändige Kodierungen erfordern, reichen. Umsetzen lässt sich dies etwa durch die Erstellung und Ausführung eines Python-Skripts.

Bei der Extraktion von Daten aus divergenten Quellen sind jedoch nicht nur solche Informationen von Interesse, die sich in der Endtabelle im Zielsystem wiederfinden, sondern auch solche, die zur Erhöhung der Datenvalidität eingesetzt werden können. So ermöglicht etwa der Rückgriff auf die von der DNB verwendeten Gattungsbegriffe eine weitere Einschränkung der bis dato erhaltenen Suchergebnisse. Ebenso können die zugewiesenen Sprachcodes, die eindeutig identifizieren, welche Manifestationen im Original in deutscher Sprache erschienen sind beziehungsweise neben deutschsprachigen Texten auch noch anderssprachige Erzeugnisse enthalten, eine Konkretisierung des Datenbestandes in Bezug auf die Forschungsfragen leisten. Zu finden sind diese Informationen in den weitaus umfangreicheren MARC21-xml-Repräsentationen der Datensätze[18]. „MARC-Formate sind Standards für die Repräsentation und den Tausch von Daten in maschinenlesbarer Form“ (DNB 2025f), wobei diese in auch in einer XML-Struktur zur Verfügung gestellt werden. Wie sich der „Arbeitstabelle Feldbeschreibung Titeldaten im Format MARC21 (Stand: 26. Mai 2025)“ entnehmen lässt, handelt es sich bei dem Feld 655 um den Indexierungsterm für das Genre bzw. die Form, wobei die zur weiteren Einschränkung benötigten Gattungsbegriffe vom Unterfeld „$a“ ausgegeben werden. Weist eine Manifestation mehrere Gattungszuweisungen auf, so wird jeweils ein neues Feld „655“ erzeugt (vgl. DNB 2025f). In ähnlicher Funktion weist das Feld „041“ die Sprachinformationen auf, wobei das Unterfeld „$a“ die Sprache(n), in der das Medienwerk verfasst ist, ausgibt und das Unterfeld „$h“ etwaige Angaben zur Originalsprache enthält. Für alle weiteren, vom Zielsystem benötigten Informationen, können die in Tabelle 4 enthaltenen Datenfelder aufgerufen werden. So wird es möglich, die zunächst aufgrund der Informationsvielfalt unübersichtlich wirkende MARC21-xml-Datei durch den Einsatz eines Python-

[18] Neben den CSV- und MARC21-xml-Dateien bietet die DNB noch weitere Exportformate an, auf die im Rahmen der vorliegenden Forschungsstudie nicht weiter eingegangen wird. Siehe hierzu ausführlich: DNB 2025k. URL: https://www.dnb.de/DE/Professionell/Metadatendienste/Datenbezug/metadatenBezugUndFormate.html.

Skripts in das vom Zielsystem benötigte Format zu konvertieren, indem die Inhalte der relevanten XML-Felder in die Spalten einer neuen CSV-Datei überführt werden.

Spalte der Tabelle ›Buch‹ im Zielsystem	**MARC21-Feld**
ISBN	020 $a
Buchtitel	245 $a
Auflage	250 $a
Publikationsort	264 $a
Publisher	264 $b
Publikationsjahr	264 $c
Seitenzahl	300 $a
Verfasser	700 $a (Zuordnung nach 700 $e)
Herausgeber	700 $a (Zuordnung nach 700 $e)
Übersetzer	700 $a (Zuordnung nach 700 $e)
keineRolle	700 $a (Zuordnung nach 700 $e)
sonstigeRolle	700 $a (Zuordnung nach 700 $e)
Gattung_Genre	655 $a
Sprache	041 $a
Originalsprache	041 $h

Tabelle 4: Relevante MARC21-Felder für die Tabelle ›Buch‹.

Da sich die Inhalte der Datenfelder in der MARC21-xml und der exportierbaren CSV-Datei im Datenshop teilweise unterscheiden, können verschiedene Vor- und Nachteile der beiden Exportformate für den Einsatz im Zuge des geschilderten Beispielprozesses herausgestellt werden: So entspricht der Buchtitel im MARC21-Feld „020 $a“ den Anforderungen des Zielsystems und muss nicht weiter bereinigt werden. Auch sind die Informationen zum Publikationsort und Publisher in unterschiedliche Felder geschrieben, sodass hier kein Trennmechanismus notwendig ist. In beiden Exportformaten analog verhält es sich indes für die Spalten *isbn*, *edition* und *seitenzahl*. Während die beiden erstgenannten Spalten ohne Anpassung übernommen werden können, müssen die Ziffern der Seitenzahl formal exkludiert werden. Mit einem Mehraufwand gegenüber dem direkten Export im CSV-Form ist bei der Spalte Publikationsjahr über das MARC21-Feld „264 $c“ zu rechnen. Dieses liefert im Vergleich zum Datenfeld *year* nicht zuverlässig vierstellige Ziffernfolgen, sondern beinhaltet in einigen Fällen Klammern, sowie Angaben zum Monat oder der Jahreszeit, sodass es sich hierbei nicht um den erforderlichen Datentyp integer handelt und die obsoleten, d. h. über die vierstellige Jahreszahl hinausgehenden Zeichen entfernt werden müssen. Zusammenfassend ergibt sich aus der Arbeit mit dem Datentyp MARC21-xml jedoch ein geringerer Anpassungsaufwand, der sich in einer einfacheren Kodierung niederschlägt. Gleichzeitig ermöglichen die zusätzlichen Datenfelder mit den Gattungsinformationen und Sprachcodes inhaltliche Bereinigungen, die zu einer Erhöhung der Datenvalidität führen können.

Da der Export der insgesamt 22.976 Datensätze über die SRU-Schnittstelle vollautomatisiert und ohne Mengenbegrenzung erfolgen kann, wurde die Entscheidung getroffen ein Python-Skript zur Arbeit mit dieser Schnittstelle und den MARC21-xml-Repräsentationen zu konzipieren. Das Ergebnis stellt das im Github-Repository abgelegte Skript „01_SRU_DNB_Datenexport_MARC21_to_CSV.py“

dar[19]. Die auf diese Weise erzeugte CSV-Datei „DNB_KE_2008_2023.csv“[20], stellt die Grundlage für alle weiteren Bereinigungs- und Anreicherungsschritte dar.

2.4.2 Bereinigung der Datensätze

Bereits nach flüchtiger Durchsicht der 22.976 extrahierten Datensätze wird erkennbar, dass sich darunter einige unvollständige, fehlerhafte, doppelte und mit Blick auf das Forschungsinteresse irrelevante Einträge befinden. Zur Verbesserung der Datenqualität ist es daher notwendig, einige formale und auch inhaltliche Bereinigungsschritte vorzunehmen[21]. Hierfür wurde das Python-Skript „02_Bereinigung_DNB-Daten.py“[22] erstellt, mithilfe dessen die Datensätze der „DNB_KE_2008_2023.csv“ schrittweise an die Anforderungen des Zielsystems angepasst und mit dem Ziel reduziert wurden, dass darin ausschließlich Manifestationen beschrieben werden, die kurze Erzählformen enthalten, im Beobachtungszeitraum 2008-2023 veröffentlicht und im Original auf deutscher Sprache verfasst wurden. Wie dem Skript zu entnehmen ist, wird dabei mit der Python-Bibliothek *pandas*[23] gearbeitet. Die Grundstruktur des Skripts folgt einer einheitlichen Logik: nach dem Einlesen der erzeugten CSV-Datei „DNB_KE_2008_2023.csv“ werden einige Output-Pfade und -Dateien definiert, die auf die final bereinigte Datei sowie weitere, im Bereinigungsprozess erzeugte Exporttabellen verweisen. Bevor auf diese Exporttabellen, deren Zweck es ist, inhaltliche Bereinigungsschritte technisch zu dokumentieren und reflektierbar zu machen, näher eingegangen wird, werden in aller Kürze die formalen Anpassungen geschildert, welche notwendig sind, um eine statistische Auswertbarkeit der Daten zu ermöglichen: Als Erstes werden im Skript die Inhalte der Spalten *Publikationsort*, *Publisher* und *Auflage* konsolidiert. Konkret werden dabei die eckigen Klammern, die einige Einträge aufweisen, entfernt. Danach werden die Inhalte der Spalte *Seitenzahl* so reduziert, dass nur mehr die Ziffern, welche die Seitenzahl angeben verbleiben. Mit dem letzten notwendigen formalen Schritt wird die vierstellige Jahreszahl aus den Angaben in der Spalte *Publikationsjahr* extrahiert.

Nach diesen, auf formale Gleichheit abzielenden Bereinigungsschritten, folgen im Skript einige Bereinigungsschritte, die eine Erhöhung der Datenvalidität bezwecken, indem unvollständige, redundante oder inhaltlich unzutreffende Datensätze entfernt werden. Damit diese Schritte und ihre Auswirkungen nachvollziehbar sind, erzeugt das Skript zu jedem Schritt eine Exporttabelle, welche jene

[19] Siehe: https://github.com/maximilianwuerz/kurze-Erzaehlformen/tree/main/Python-Skripte.
[20] Siehe: https://github.com/maximilianwuerz/kurze-Erzaehlformen/tree/main/Outputs.
[21] Streng genommen ließen sich die bis zu dieser Stelle geschilderten Anpassungen an den Inhalten der Datenfelder bereits als Bereinigungsschritte bezeichnen. Da diese jedoch im Zuge des Extraktionsprozesses erfolgen und sozusagen als Extraktion zweiter Ordnung, Informationen aus den MARC21-xml-Repräsentationen, welche aus dem Gesamtbestand der DNB extrahiert wurden, in eine neue CSV-Datei segregieren, wurden sie im vorangegangenen Kapitel untergebracht.
[22] Siehe: https://github.com/maximilianwuerz/kurze-Erzaehlformen/tree/main/Python-Skripte.
[23] Pandas ist ein “open source data analysis and manipulation tool, built on top of the Python programming language” (Pandas 2025. URL: https://pandas.pydata.org/). Genutzt wurde die Version 2.3.3.

Datensätze dokumentiert, die bei den jeweiligen Schritten aus der Gesamttabelle entfernt werden[24]. So wird nicht nur erkennbar welche bzw. wie viele redundante Datensätze separiert wurden; es wird auch möglich Reflexionen über die Treffsicherheit der formulierten Suchanfrage, die Datenqualität des Gesamtbestandes der DNB sowie die Eignung desselben als Datengrundlage für quantitative Forschungsarbeiten mit gattungsbezogenen Fragestellungen anzustellen. Der nächste Bereinigungsschritt im Skript nimmt sich also der Entfernung von Duplikaten in der Datei „DNB_KE_2008_2023.csv“ an. Als Duplikat gilt dabei jeder Datensatz, der mit einem anderen Datensatz über alle vorhandenen Spalten hinweg identisch ist. Dies ist insofern möglich, als nur ein geringer Teil aus den MARC21-xml-Repräsentationen der identifizierten Daten übernommen wurde. Datensätze, die im Gesamtbestand der DNB nicht als Duplikate gelten, da sie sich in einem der vielen vorhandenen Datenfelder unterscheiden oder einen eindeutigen Identifikator aufweisen, der bei der Extraktion nicht übernommen wurde, können somit in der hier bearbeiteten reduzierten Fassung von Datenfeldern Duplikate darstellen[25]. Wie der Exporttabelle „Entfernte_Duplikate.csv“ zu entnehmen ist, wurden in diesem Schritt 381 Datensätze entfernt, deren Daten über alle in Tabelle 4 dargestellten Spalten und MARC21-Datenfelder identisch sind. Die beiden darauffolgenden Bereinigungsschritte fokussieren hingegen auf die Spalte *ISBN*. Zuerst werden alle Datensätze entfernt, die keine ISBN aufweisen (= 1.401 Datensätze). Dieser Schritt wird unternommen, da eine Durchsicht der Daten zuvor ergab, dass Datensätze ohne ISBN oft fehlerhaft sind oder auf ein Medienwerk verweisen, das kein Buch im klassischen Sinn, sondern ein Kalender, Gesellschaftsspiel, Magazin oder anderweitiges Druckerzeugnis ist. Exemplarisch können die in der Exporttabelle „Entfernte_ohne_ISBN.csv“ enthaltenen Datensätze der Heftromane vom Martin Kelter Verlag oder der Karten-/Rätselspielreihe „black stories“ vom moses. Verlag angeführt werden. Danach werden Datensätze entfernt, die entsprechend ihrer *ISBN* ein Duplikat zu einem anderen Eintrag sind (= 171 Datensätze). Dies erfolgt aus dem Grund, da die ISBN im Zielsystem den eindeutigen Identifikator in der Tabelle ›Buch‹ darstellt und davon ausgegangen wird, dass sich Manifestationen, mit der gleichen ISBN, nicht wesentlich unterscheiden, die doppelte Aufnahme im Datenbestand aus literaturwissenschaftlicher Perspektive also keinen Mehrwert bietet[26]. Mit dem nächsten Bereinigungsschritt nimmt sich das Skript der Ausrichtung des Fallbeispiels auf deutschsprachige Texte an. Die Nutzung

[24] Die Exporttabellen sind zu finden unter: https://github.com/maximilianwuerz/kurze-Erzaehlformen/tree/main/Outputs.

[25] Die Verwendung des Begriffs Duplikat erfolgt daher weniger im streng bibliografischen Sinn und bezogen auf den Katalogisierungsstandard des Gesamtbestandes der DNB, sondern bezogen auf jene Informationen, die zur Beantwortung der literaturwissenschaftlich ausgerichteten Fragestellungen im Fallbeispiel, von Relevanz sind. Dabei ist es bspw. interessant, wie viele Kurzgeschichtenbände in einem bestimmten Jahr von einer bestimmten Autorin erschienen sind, aber weniger, wie viele Datensätze zu Manifestationen ein und desselben Titels im Verzeichnis der DNB technisch angelegt wurden.

[26] Diese Konzentration auf Buchtitel statt Manifestationen ließe sich freilich noch weitertreiben, indem z.B. nur Erstauflagen berücksichtigt oder bei Datensätzen mit gleichem Buchtitel nur jene beibehalten werden, die am frühesten publiziert wurden. Da auch diese Vorgehensweisen mit Unschärfen verbunden sind, wird dieser Aspekt in der vorliegenden Forschungsstudie nicht weiter ausgeleuchtet.

des Sprachindex in der Suchanfrage („spr=ger“) hat bereits dafür gesorgt, dass ausschließlich Titel ausgegeben wurden, die auf deutscher Sprache publiziert wurden. Hiermit ist jedoch nicht garantiert, dass es sich um genuin deutschsprachige Texte und Bücher handelt; stattdessen finden sich in den Datensätzen auch vielfach Übersetzungen aus anderen Sprachen. Um diese nun ausschließen zu können, wurde die Spalte *SpracheOriginal* aus dem Datenfeld „041 $h“ erzeugt und der Bereinigungsschritt im Skript eingebaut. Als Resultat finden sich 2.082 entfernte Datensätze in der Exporttabelle Entfernte_Originalsprache.csv“. Da die Forschungsarbeit zu kurzen Erzählformen darüber hinaus ein besonderes Interesse an den Akteuren, die an der Entstehung und Verbreitung dieser Schreibweisen beteiligt sind, hat, wurden in einem weiteren Schritt all jene Datensätze entfernt, die in keiner der Rollenspalten (also auch nicht in der Spalte *creator* der Titeldaten oder dem MARC21-Datenfeld „700 $a“) einen Eintrag haben. Exkludiert wurden auf diese Weise weitere 758 Zeilen. Zu guter Letzt erfolgt im Skript noch ein weiterer Bereinigungsschritt hinsichtlich der im Datenset vorhandenen Gattungen, um zu gewährleisten, dass ausschließlich solche Titel berücksichtigt werden, die kurze Erzählformen enthalten. Da die Eignung des Gesamtbestandes der DNB für quantitative Forschungsarbeiten mit gattungsbezogenen Fragestellungen im Fokus der vorliegenden Studie liegt, nimmt dieser Aspekt im Folgenden etwas mehr Raum ein.

Zur Erinnerung: Es wurde bereits festgestellt, dass die von der DNB verwendeten Gattungsbegriffe für die Recherche relevanter Titel in einem ersten Schritt nur bedingt eignen, da sie entweder zu weitläufig oder aber zu spezifisch sind. Dazu wurde die These aufgestellt, dass sich der Einbezug der Gattungsbegriffe zur Filterung an späterer Stelle als fruchtbar erweisen könnte. Dies ist es, was mit dem nun behandelten Bereinigungsschritt im Skript versucht wird. Um vor der Kodierung des Bereinigungsschritts ein Gefühl dafür zu bekommen, welche Gattungsbegriffe im Datenset vorkommen, wurden die nach den vorangegangenen Säuberungen verbliebenen 18.291 Datensätze zunächst einmal manuell durchgesehen. Mehrere Aspekte stachen bei dieser ersten Datenexploration hervor: Erstens, sind viele der Titel mit mehreren Begriffen versehen, wobei sich die Bezeichnungen in einigen Fällen zu ergänzen scheinen und so als konkretere Verweise gedeutet werden können. Dies ist der Fall, wenn etwa „Science Fiction“ mit „Erzählende Literatur: Anthologien“ kombiniert wird, was relativ eindeutig auf eine narrative Sci-Fi-Anthologie hindeutet und damit auf eine inhaltliche Relevanz für das hier bearbeitete Fallbeispiel. Zweitens konnten einige verwendete Gattungsbegriffe identifiziert werden, die eindeutig auf eine inhaltliche Irrelevanz hindeuten, wie zum Beispiel „Hauptwerk vor 1945“ aufgrund des gesteckten Beobachtungszeitraums oder „Historische Kriminalromane“, da Romane *per definitionem* das Kriterium der Kürze verfehlen. Auch die Begriffe der „Lyrik“ und „Dramatik“ können dazu gezählt werden, wenn auch mit einer Einschränkung: Werden sie mit Bezeichnungen kombiniert, die auf eine gemischte Anthologie oder zugleich erzählende Literatur hindeuten, ist es denkbar, dass es sich um ein Sammelwerk handelt, dass neben lyrischen oder dramatischen Beiträgen auch solche fasst,

die als kurze Erzählformen verstanden werden können. Drittens und in Kontrast zum zweiten Aspekt, lässt sich die eindeutige Relevanz eines Datensatzes nur über einen verwendeten Gattungsbegriff herleiten, und zwar „Erzählende Literatur: Anthologien“, da dieser impliziert, dass narrative Formen gebündelt darin vorkommen und in der Regel kurze Texte assembliert werden (vgl. Althaus/Bunzel/Göttsche 2007, XXIII). Alle weiteren verwendeten Begriffe können lediglich mit wechselnder Genauigkeit als Indikatoren für eine mögliche Relevanz herangezogen werden. Bei der konkreten Umsetzung des Bereinigungsschrittes im Skript wurden deshalb nur die unter dem zweiten Aspekt angeführten Gattungsbegriffe verwendet, also alle Datensätze exkludiert, die in der Spalte Gattung_Genre die Worte „Lyrik“ oder „Dramatik“, aber nicht die Strings „gemischt“ oder „hlende Literatur“27 enthalten. Zudem wurden jene mit den Worten „Hauptwerk vor 1945“, „Romanhafte“ oder „Kriminalroman“ entfernt und jene mit dem Wort „Bildband“, so denn sie nicht gleichzeitig den String „hlende Literatur“ enthielten. Wie die Exporttabelle dieses Bereinigungsschrittes „Entfernte_falsche_Gattung.csv“ erkennbar macht, wurden auf diese Weise weitere 1.435 Datensätze entfernt; erhalten blieben somit 16.856. Um einen Einblick zu geben, welche Gattungszuschreibungen in dem reduzierten Datenset mit welchen Häufigkeiten vorzufinden sind, wurden in Abbildung 1 die 20 meistgenutzten Bezeichnungen diagrammatisch dargestellt. In 3.136 Fällen – und damit am häufigsten –

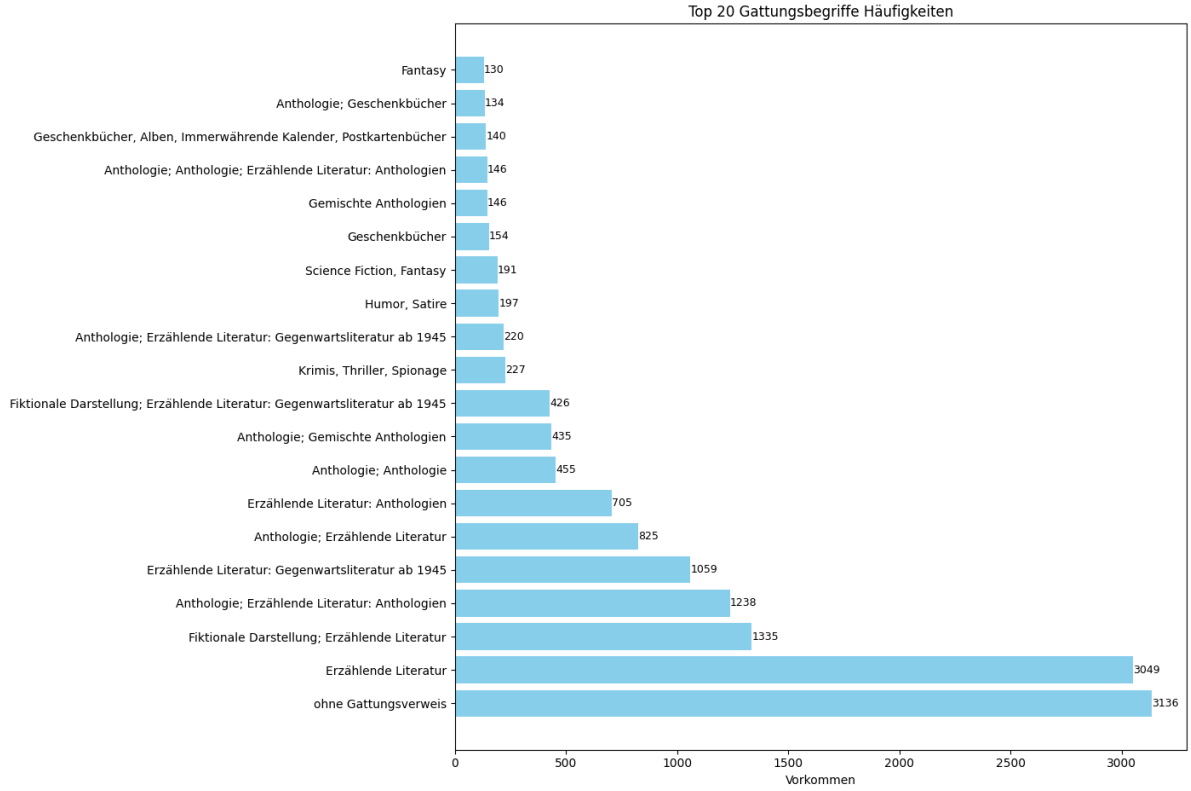

Abbildung 1: Top 20 Gattungsbegriffe nach Häufigkeiten im bereinigten Datenset.

[27] Da es bei der Filterung auf Basis des Strings „Erzählende Literatur“ wiederholt zu fehlerhaften Umsetzungen wegen des Umlauts kam, wurde der String soweit reduziert, dass er noch eindeutig auf die Gattungsbegriffe der Erzählenden Literatur verweist ohne den Umlaut einzubeziehen.

wird überhaupt keine Gattungszuschreibung verwendet. Bereits dieser Aspekt verdeutlicht, dass eine Recherche im Datenbestand, die sich ausschließlich auf die im Datenfeld „655“ festgehaltenen Gattungsbegriffe stützt, eine Vielzahl potenziell relevanter Datensätze unberücksichtigt lässt. Danach ist der mit Abstand am häufigsten verwendete Gattungsverweis „Erzählende Literatur“. Dies ergibt insofern Sinn, als dieser Begriff in der offiziellen „Liste der in der DNB verwendeten Gattungsbegriffe“ hierarchisch über den inkludierten Unterbegriffen „Erzählende Literatur: Hauptwerk vor 1945“, „Erzählende Literatur: Gegenwartsliteratur ab 1945“, „Historische Romane und Erzählungen“, „Erzählende Literatur: Anthologien“ und „Romanhafte Biographien“ steht. Ist die Auszeichnung eines Titels etwa aufgrund gewisser Textkomplexitäten mit Unsicherheiten verbunden, kann der Rückgriff auf einen weniger konkreten Oberbegriff eine Lösung sein. Knapp die Hälfte der hier aufgelisteten Gattungsbezeichnungen weisen auf einen narrativen Inhalt der Titel hin, wobei die Zusätze „Anthologie“, „Anthologien“ und „Gegenwartsliteratur ab 1945“ den Aussagewert erhöhen und stärker darauf hindeuten, dass in diesen Werken kurze Erzählformen enthalten sind. Daneben finden sich mit „Krimis, Thriller, Spionage“, „Humor, Satire“, „Science Fiction; Fantasy“ und „Fantasy“ auch vier Gattungsverweise, die weniger auf „formale“, als auf „thematisch- inhaltliche“ (Neumann/Nünning 2007, S. 2) Aspekte verweisen. Hierdurch wird zwar markiert, dass die Bücher fantastische oder futuristische Stoffe verarbeiten oder aber Elemente aufweisen, die typischerweise in Krimis verarbeitet werden, nicht aber, ob dies in kurzer, prosaischer und narrativer Form erfolgt (auch wenn dies in der Literatur wohl die gängigste Art der Verarbeitung solcher Inhalte darstellt). Zuletzt finden sich in der Abbildung mit den Beschreibungen „Geschenkbücher, Alben, Immerwährende Kalender, Postkartenbücher“ bzw. „Geschenkbücher“ noch zwei Gattungsbezeichnungen, die eher „außerliterarische Kriterien“ (Neumann/Nünning 2007, S. 2) heranziehen und dabei auf den Funktionalisierungskontext dieser Titel, jedoch ebenfalls nicht eindeutig auf kurze narrative Texte als Inhalt verweisen.

Zurückgeblickt auf die in Kapitel 2.3 formulierte Grundherausforderung, wonach in der Datenquelle *all* und gleichzeitig *nur* jene Datensätze identifiziert werden sollten, die aufgrund der beinhalteten kurzen Erzählformen von Relevanz für die exemplarisch angeführte Untersuchung sind, lässt sich anhand der genutzten Gattungsbezeichnungen eine erste vorsichtige Bewertung des bisher geschilderten ETL-Prozesses vornehmen. So scheinen die Gattungsbezeichnungen der verbliebenen Datensätze erstmal nicht im Widerspruch zu dem ausgegebenen Ziel zu stehen, da sie nicht eindeutig auf eine Irrelevanz verweisen. Mit jeder genutzten Bezeichnung kann theoretisch ein Titel mit kurzen Erzählformen beschrieben werden. Dies ist umso wahrscheinlicher, da zu Beginn des Prozesses bereits weitaus spezifischere Stichwörter genutzt wurden, um die Titel, Untertitel und Gesamttitel der Medienwerke im Gesamtbestand zu durchsuchen. So handelt es sich bei den 3.136 leeren sowie den 3.049 Datensätzen, die mit dem Gattungsbegriff „Erzählende Literatur“ ausgezeichnet wurden, gleich-

sam um Datensätze, die mindestens eines der gesuchten Stichwörter, wie „Kurzgeschichte“, „Anekdote“ etc. in den Titelfeldern aufweisen. Wie Schubert aufzeigt, ist der automatisierten Bezug von Metadaten aus bibliografischen Beständen grundsätzlich mit Unsicherheiten verbunden. So sind die „Voraussetzungen von Etikettierungen oder Labels oder Indexeinträgen […] immer begriffsbezogene Datenkategorien, die aus fachspezifischen Kontexten stammen, die durch automatisierte Extraktion nicht zuverlässig erfaßt werden können“ (Schubert 2021, S. 42). Verstärkt werden diese Unsicherheiten noch, wenn selbst innerhalb einer Disziplin Uneinigkeit zu bestimmten Klassifizierungskriterien herrscht, wie dies in der Literaturwissenschaft klassischerweise bei Gattungsdiskussionen der Fall ist. Trifft also ein bereits uneinheitliches Verständnis von Gattungen auf eine abweichende bibliografische Klassifizierungspraxis, kann es zu fehlerhaften Ergebnissen im Extraktionsprozess kommen. Die Frage ist an dieser Stelle also mit wie viel Unsicherheit umgegangen werden kann, welche Fehlerquote vertretbar ist und ob eine manuelle Prüfung der extrahierten Datensätze im Nachgang erforderlich ist. Da die vorliegende Forschungsstudie in der Hauptsache einen Überblick über die Potenziale des Gesamtbestandes der DNB zur Identifikation, Extraktion und Weiterverarbeitung von Metadaten mit expliziten Gattungsbezügen liefern und mit der Beantwortung der formulierten Forschungsfragen primär eine Herausstellung möglicher Herausforderungen und verschiedener Praktiken im Umgang damit leisten möchte, wird die Bereinigung des Datenbestandes nicht weiter konkretisiert. Vielmehr soll nun noch in Kürze aufgezeigt werden, wie sich die Daten zu Personen und Körperschaften, die nachweislich eine Rolle bei der Entstehung und Verbreitung von kurzen Erzählformen einnehmen, über einen Abgleich mit der Gemeinsamen Normdatei anreichern lassen und gemeinsam mit den extrahierten Datensätzen zu den relevanten Buchpublikationen in das Zielsystem integriert werden können.

2.4.3 Anreicherung und Import in das Zielsystem

Die bisher erhobenen Personen- und Körperschaftsdaten, beschränken sich auf die Namen und Rollen, wie sie in den Datensätzen der „DNB_KE_2008_2023_bereinigt.csv“ zu finden sind. Auf Basis der darin zu findenden Namen müssen alle weiteren, vom Zielsystem benötigten Daten (siehe Tabelle 2 & 3) angereichert werden. Hierfür ist es notwendig, die Namen der Personen und Körperschaften zu extrahieren und in jeweils eigene CSV-Dateien zu überführen. Für die Extraktion der Personennamen werden zunächst jene Spalten herangezogen, die sich durch Inhalte mit Personennamen auszeichnen, also die Rollenspalten *Verfasser*, *Herausgeber*, *Übersetzer*, *keineRolle* und *sonstigeRolle* aus der Datei „DNB_KE_2008_2023_bereinigt.csv“. Mittels eines Python-Skripts „03_Extraktion_Person_Koerperschaft_Namen.py“[28] werden die Namen automatisch separiert und in die neu angelegte Datei „Person.csv“ mit den Spalten P*erson_ID* und *Person_Name* übertragen. Analog zu diesem Vorgehen,

[28] Siehe: https://github.com/maximilianwuerz/kurze-Erzaehlformen/tree/main/Python-Skripte.

werden die Körperschaftsnamen, d. h. alle Bezeichnungen von Verlagen, Unternehmen, Organisationen, Institutionen etc., die in der Spalte *Publisher* der Datei „DNB_KE_2008_2023_bereinigt.csv“ vorkommen, in einem gesonderten Schritt im Python-Skript extrahiert und in die Spalten *Koerperschaft_ID* und *Koerperschaft_Name* der neuen Datei „Koerperschaft.csv“ eingepflegt. Da die gesammelten Namen trotz der einheitlichen Bezugsquelle teilweise differente Schreibweisen aufweisen, müssen im nächsten Schritt Konsolidierungsmaßnahmen getroffen werden. Mehrere Optionen stehen hierfür bereit, wobei an dieser Stelle nur zwei Varianten angerissen werden. Zum einen können die neu erzeugten, alphabetisch sortierten Dateien (manuell) durchgesehen sowie auf Dubletten[29] überprüft werden. Die so identifizierten Dubletten werden danach in ein Mapping überführt, in ein Python-Skript integriert und über die Ausführung dieses Skripts konsolidiert[30]. Alternativ können die beiden CSV-Dateien auch mit der Cluster-Methode „Nearest Neighbor“ des Open-Source-Tools OpenRefine konsolidiert werden. Dabei werden über die Anwendung der Levenshtein Distanz ähnliche Zeichenketten zusammengeführt und im Anschluss manuell oder automatisiert konsolidiert[31]. Welche Option gewählt wird, ist für die Erkenntnisse der vorliegenden Forschungsstudie unerheblich. Stattdessen gilt es festzuhalten, dass die Arbeit mit Personen- und Körperschaftsnamen, die dem Gesamtbestand der DNB entnommen werden, aufgrund der häufig voneinander abweichenden Namensschreibweisen nach der Extraktion weitere Konsolidierungsschritte notwendig macht, die je nach Menge der Datensätze teils erhebliche manuelle Mehraufwände mit sich zieht. Liegen die Datensätze in der bereinigten Form vor[32], können nun alle im Datenset vorkommenden Namen über den Reconciliation-Dienst der Gemeinsamen Normdatei (GND) in OpenRefine mit den Normdaten der GND abgeglichen[33] und basierend darauf mit weiteren Daten angereichert werden. Auf diese Weise wird nicht nur den Vorgaben der Deutschen Forschungsgemeinschaft, die „[f]ür die Erfassung personenbezogener, biografischer und geografischer Information […] die Verwendung der Gemeinsamen Normdatei (GND) der Deutschen Nationalbibliothek erwartet“ (Altenhöner et al 2022), entsprochen; vielmehr können den Datensätzen hierdurch bereits geprüfte Normdaten angefügt werden. So können die bisher zusammengetragenen Personen- respektive Körperschafsnamen – wenn sie beim Abgleich

[29] Dubletten sind Daten oder Datensätze, die identisch sind bzw. auf eine identische Entität verweisen, aber aufgrund teils minimaler Abweichungen etwa in der Schreibweise nicht als identisch erkannt werden (vgl. Schneider 1999, S. 60ff.).

[30] Siehe: https://github.com/maximilianwuerz/kurze-Erzaehlformen/tree/main/Python-Skripte.

[31] Zu einer ausführlichen Erläuterung der Cluster-Methoden von OpenRefine siehe: https://openrefine.org/docs/manual/cellediting#cluster-and-edit.

[32] „Person_bereinigt.csv“ und „Koerperschaft_bereinigt.csv“ unter https://github.com/maximilianwuerz/kurze-Erzaehlformen/tree/main/Outputs.

[33] Zu den Diensten und einer einführenden Erläuterung dem Reconsiliation-Dienst der GND siehe: https://gnd.network/Webs/gnd/DE/Entdecken/entdecken_node.html#gndExplorer. Zu einer ausführlichen Erläuterung des Reconsiling mit OpenRefine siehe: https://openrefine.org/docs/manual/reconciling. Zur Nachvollziehbarkeit des hier angewendeten Verfahrens zur GND-Reconsiliation via OpenRefine siehe: Pohl, Adrian/Steeg, Fabian: GND reconciliation for OpenRefine. URL: https://blog.lobid.org/2018/08/27/openrefine.html. [Zugriffsdatum: 29.05.2025]. 27 Aug 2018 – 19 Aug 2019.

erkannt werden – um in der GND vorhandene Informationen erweitert werden. Ist diese Erweiterung erfolgt, sodass die Dateien in der vom Zielsystem benötigten Form vorliegen, können sie in OpenRefine heruntergeladen werden. Beim Umgang mit neueren Daten bzw. zukünftig dürfte sich der Reconsiliation-Prozess sogar noch vereinfachen und dabei immer besser funktionieren, denn wie Glagla-Dietz et. al betonen, hat sich die „Arbeit mit Personenangaben in Titeldaten [..] in der DNB in den letzten Jahren stark verändert“ (Glagla-Dietz et. al 2024, S. 15). Seit 2017 werden demnach vermehrt maschinelle Prozesse eingesetzt, um die Personendaten im Gesamtbestand der DNB automatisiert mit GND-Datensätzen zu verknüpfen (ebd.). Hierdurch wird es möglich, neben den Namen und der Rolle weitere Informationen zu Personen direkt aus den MARC21-xml-Repräsentationen auszulesen, was perspektivisch gar den Einsatz von OpenRefine obsolet machen könnte. Da der Umsetzungsstatus dieses Vorhabens dem Verfasser der vorliegenden Forschungsstudie bei Ausfertigung jedoch nicht bekannt war, wurde der Prozess wie dargestellt umgesetzt. Nach Abschluss aller Anreicherungen und Transformationen, wenn also die drei CSV-Dateien für in ein für die Relationale Datenbank standardisiertes Format vorliegen, beginnt die „Ladephase“ (Kramer 2019, S. 136). In dieser Phase, die zugleich den Abschluss des Prozesses darstellt, werden die Dateien kommandozeilenorientiert oder über das GUI[34] von pgAdmin 4 in die zuvor erstellten Datencontainer der RDB importiert. Wurde dieser Schritt erfolgreich umgesetzt, können die importierten Daten mit den anderen Entitäten der RDB verknüpft und über SQL-Abfragen statistisch ausgewertet werden.

3. Prozessreflexion und Generalisierung der Erkenntnisse

Quantitative Forschungsstudien mit gattungsbezogener Ausrichtung, wie sie vor allem im Bereich der Computational Literary Studies oder den Digital Humanities bearbeitet werden, müssen trotz der zunehmenden Etablierung der Forschungsbereiche[35] damit umgehen, dass für die Beantwortung ihrer Fragestellungen in der Regel passgenau nutzbare Text- und Datenbestände fehlen oder nur mit hohen Aufwänden selbstständig erhoben werden können. Dabei gäbe es durchaus Datenquellen, die über eine Fülle von Metadatenbeständen verfügen, es fehlt jedoch oftmals an standardisierten Prozessen zur Nutzbarmachung der darin enthaltenen Informationen. In den vorangegangenen Kapiteln wurde deshalb der Versuch unternommen, einen solchen Prozess anhand eines festgelegten Szenarios zu erarbeiten, fortwährend zu reflektieren und dabei Erkenntnisse über die Eignung des bibliografischen Gesamtbestandes der Deutschen Nationalbibliothek als Datenquelle zur Identifikation, Extraktion und

[34] Steht für „Graphical User Interface“.
[35] Hierzu exemplarisch: Svenja Guhr, Trends in den Computational Literary Studies bei den DHd-Jahrestagungen, 13. Mai 2025, in: DHd Blog, https://dhd-blog.org/?p=22316. Bundesministerium für Forschung, Technologie und Raumfahrt: Digital Humanities. 2025. URL: https://www.geistes-und-sozialwissenschaften-bmbf.de/de/Digital-Humanities-1710.html (Zugriff: 28.10.2025).

Weiterverarbeitung von gattungsbezogenen Metadaten zu erlangen. Unter Einbezug des Fallbeispiels der ›kurzen Erzählformen‹ wurde die Maßgabe formuliert, möglichst alle Titel im Gesamtbestand der DNB zu identifizieren, die kurze narrative Prosagattungen enthalten, die dazugehörigen Metadaten in einer festgelegten Form zu exportieren und in eine Relationale Datenbank zu integrieren. Der so ausgearbeitete Prozess umfasste die Datenauslese über zwei verschiedene Schnittstellen, den DNB-Katalog inklusive des Datenshops sowie die SRU-Schnittstelle und eine vergleichende Analyse der Exportformate CSV und MARC21-xml. Im Zuge dieses Identifikations- und Extraktionsschritts wurde festgestellt, dass der Einbezug von Gattungsbegriffen in der Suchanfrage mit „tit all“ zunächst solide Ergebnisse liefert, die jedoch eine hohe Zahl redundanter und irrelevanter Daten beinhalten und weiter gefiltert werden müssen. Die Wahl der Schnittstelle sollte abhängig davon getroffen werden, wie hoch die Menge der identifizierten Datensätze (Suchtreffer) ist, wie gut sich die Suche in mehrere Anfragen aufteilen lässt und wie hoch der gewünschte Grad an Automatisierung ist. Übersteigt die Zahl gelieferter Suchergebnisse den Wert von 10.000, lässt sich die genutzte Syntax nur schwer zerlegen und soll der Extraktionsprozess mehrfach wiederholt werden, ist die Nutzung der SRU-Schnittstelle zu empfehlen. Daneben kann auch das vom Zielsystem benötigte Dateiformat eine Rolle spielen, wobei der Vergleich von CSV-Titeldaten und MARC21-xml-Repräsentationen ergab: Auch wenn das vom Zielsystem eingeforderte Dateiformat CSV ist und die exportierten MARC21-xml-Repräsentationen in einem Extra Schritt ausgelesen und in CSV konvertiert werden müssen, ist die Arbeit mit letzteren zu empfehlen. In den MARC21-xml-Dateien ist das Datenfeld „655“ enthalten, in dem die von der DNB verwendeten Gattungsbegriffe verzeichnet werden. Durch den Einbezug dieser Gattungsbegriffe wird eine inhaltliche, gattungsbezogene Filterung ermöglicht, wodurch sich die Datenvalidität erheblich erhöht lässt. Auch wenn die verwendeten thematisch-inhaltliche, formale und außerliterarische Aspekte vermischen (vgl. Neumann/Nünning 2007, S. 2) und gerade mit Blick auf die Länge respektive Kürze von literarischen Formen, keinen großen Mehrwert liefern, können sie doch als zusätzliches Reflexions- und Filterinstrument eingesetzt werden. Zusätzlich waren bei der Verwendung der MARC-21-xml-Repräsentationen auch weniger formale Anpassungen notwendig, da etwa die Buchtitel bereits korrekt im Datenfeld „245 a“ enthalten sind. Wie versucht wurde aufzuzeigen, lassen sich vielfältige Bereinigungen recht unkompliziert über ein zusätzliches Python-Skript oder den Einsatz des Open Source Tools OpenRefine umsetzen. Dieses Tool kann zudem zur Anreicherung von Personen und Körperschaftsnamen aus der Gemeinsamen Normdatei genutzt werden, indem die Personennamen aus den Metadaten der extrahierten Titel extrahiert in OpenRefine abgeglichen und um bibliografische oder anderweitige Daten erweitert werden. Für die Integration in ein neues Zielsystem muss letztlich nur darauf geachtet werden, dass die neu erzeugten Dateien dem benötigten Importformat und damit den Tabellen in der RDB entsprechen. Grundsätzlich gilt es zu beachten, dass im Rahmen der vorliegenden Forschungsstudie nur eine von vielen möglichen Umsetzungen des ETL-Prozesses geschildert wurde

und es mannigfaltige Alternativen gibt. Die stärkere Einbindung von OpenRefine bereits bei den formalen Bereinigungsschritten, der Einbezug einer größeren Menge von MARC21-Datenfeldern (wie den Unterfeldern zu den Personendaten von Datenfeld 700) oder auch der Einbezug der digitalisierten Inhaltsverzeichnisse zum Ausschluss längerer narrativer Formen, können hier exemplarisch genannt werden. Letzten Endes ist die Standardisierung eines solchen Prozesses aufgrund der Individualität von Forschungsinteressen immer nur bis zu einem bestimmten Grad möglich und auch die Automatisierung stößt an Grenzen, sodass eine Vollautomatisierung solcher Prozesse nur auf Kosten der Datenqualität und -validität funktioniert. Da die Klassifikationspraktiken in Bezug auf Gattungen von dem Verständnis von Gattungen in der Literaturwissenschaft abweichen, ist eine manuelle Prüfung für zuverlässige und qualitativ hochwertige Ergebnisse unerlässlich.

Abschließend lässt sich festhalten, dass der bibliografische Gesamtbestand der DNB in hohem Maß für die Identifikation, Extraktion und Weiterverarbeitung von gattungsspezifischen Metadaten geeignet ist, was an der umfangreichen Auszeichnung der Medienwerke, der kompatiblen Anbindung an die GND liegt, wodurch die Anreicherung mit Normdaten vereinfacht wird, aber auch an den vielfältigen Angeboten von Tutorials und Schulungen vonseiten der DNB[36]. Verbesserungspotenzial ist aus literaturwissenschaftlicher Perspektive lediglich hinsichtlich der Konkretheit der verwendeten Gattungsbegriffe und der nicht immer vorliegenden formalen Einheitlichkeit in einzelnen Datenfeldern zu erkennen.

[36] DNB 2024a. URL: https://www.dnb.de/DE/Professionell/Services/Wissenschaftund-Forschung/DNBLab/dnblabTutorials.html.

Literaturverzeichnis

Albrecht, Alexander: Understanding and Managing Extract-Transform-Load Systems. [Dissertation]. Universität Potsdam 2013.

Althaus, Thomas / Bunzel, Wolfgang / Göttsche, Dirk: Ränder, Schwellen, Zwischenräume. Zum Standort Kleiner Prosa im Literatursystem der Moderne. In: Kleine Prosa. Theorie und Geschichte eines Textfeldes im Literatursystem der Moderne. Hgg. von Thomas Althaus, Wolfgang Bunzel, Dirk Göttsche. Tübingen: Max Niemeyer Verlag 2007.

Bundesministerium für Forschung, Technologie und Raumfahrt: Digital Humanities. 2025. URL: https://www.geistes-und-sozialwissenschaften-bmbf.de/de/Digital-Humanities-1710.html (Zugriff: 28.10.2025).

Deutsche Nationalbibliothek: CSV – Comma-separated values. 2024. URL: https://www.dnb.de/csv. (Zugriff: 20.10.2025).

Deutsche Nationalbibliothek: DnbLab: Tutorials. 2024a. URL: https://www.dnb.de/DE/Professionell/Services/WissenschaftundForschung/DNBLab/dnblabTutorials.html.

Deutsche Nationalbibliothek: Die Deutsche Nationalbibliothek im Porträt. 2025a. URL: https://www.dnb.de/portraet (Zugriff: 28.10.2025).

Deutsche Nationalbibliothek: Erschließung. 2025b. URL: https://www.dnb.de/erschliessung (Zugriff: 19.10.2025).

Deutsche Nationalbibliothek: Erschließung von Medienwerken. 2025c. URL: https://www.dnb.de/DE/Professionell/Erschliessen/erschliessen_node.html (Zugriff: 22.10.2025).

Deutsche Nationalbibliothek: Hilfe: Expertensuche. 2022. URL: https://www.dnb.de/DE/Service/Hilfe/Katalog/kataloghilfeExpertensuche.html?nn=493900#codes (Zugriff: 29.10.2025).

Deutsche Nationalbibliothek: Willkommen im Datenshop. 2025d. URL: https://portal.dnb.de/metadataShopHome#top (Zugriff: 22.10.2025).

Deutsche Nationalbibliothek: Liste der in der DNB verwendeten Gattungsbegriffe (Stand: 01. Oktober 2025). 2025e. URL: https://www.dnb.de/SharedDocs/Downloads/DE/Professionell/Erschliessen/listeGattungsbegriffe.html (Zugriff: 29.10.2025).

Deutsche Nationalbibliothek: MARC 21. 2025f. URL: https://www.dnb.de/marc21 (Zugriff: 22.10.2025).

Deutsche Nationalbibliothek: Metadatendienste. 2025g. URL: https://www.dnb.de/DE/Professionell/Metadatendienste/metadatendienste_node.html#sprg186916 (Zugriff: 21.10.2025).

Deutsche Nationalbibliothek: SRU-Schnittstelle. 2025h. URL: https://www.dnb.de/DE/Professionell/Metadatendienste/Datenbezug/SRU/sru_node.html (Zugriff: 20.10.2025).

Deutsche Nationalbibliothek: Über den Katalog der Deutschen Nationalbibliothek. 2025i. URL: https://www.dnb.de/DE/Benutzung/Katalog/katalogDNB_node.html (Zugriff: 22.10.2025).

Deutsche Nationalbibliothek: Bezugswege und Exportformate. 2025k. URL: https://www.dnb.de/DE/Professionell/Metadatendienste/Datenbezug/metadatenBezugUndFormate.html (Zugriff: 22.10.2025).

Doderer, Klaus: Die Kurzgeschichte in Deutschland. Ihre Form und Ihre Entwicklung. Darmstadt: Wissenschaftliche Buchgesellschaft 1980 [1953].

Durzak, Manfred: Die Kunst der Kurzgeschichte: zur Theorie und Geschichte der deutschen Kurzgeschichte. 2. verb. Auflage. München: Fink 1994.

Durzak, Manfred: Die deutsche Kurzgeschichte der Gegenwart: Autorenporträts, Werkstatt-gespräche, Interpretationen. Würzburg: Königshausen & Neumann 2002.

Gencarelli, Angela: Doing Genre. Grundlagen, Begriffe und Perspektiven einer praxeologischen Gattungsforschung. In: Doing Genre: Praxeologische Perspektiven auf Gattungen und Gattungsdynamiken. Hgg, von Angela Gencarelli. Berlin/Boston: De Gruyter 2024. S. 1-22. DOI: https://doi.org/10.1515/9783111217741-001.

Glagla-Dietz, Stephanie/Grund-Davidov, Stefan/Habermann, Nicole/Vorndran, Angela: Personen in GND und Titeldaten. Vorschläge, Anreicherungen und Verknüpfungen. In: o-bib. Das offene Bibliotheksjournal, Bd. 11 Nr. 1 (2024). DOI: https://doi.org/10.5282/o-bib/5998.

GND – Gemeinsame Normdatei: Entdecken Sie die Dienste und Angebote rund um die GND von A bis Z. 2024. URL: https://gnd.network/Webs/gnd/DE/Entdecken/entdecken_node.html#gndExplorer (Zugriff: 20.10.2025).

GND – Gemeinsame Normdatei: GND Reconciliation. URL: https://lobid.org/gnd/reconcile/ (Zugriff: 20.10.2025).

GND – Gemeinsame Normdatei: Über die GND. 2022. URL: https://gnd.network/Webs/gnd/DE/UeberGND/ueberGND_node.html (Zugriff: 20.10.2025).

Goethe, Johann Wolfgang von: Noten und Abhandlungen zum besseren Verständniß des West-östlichen Divans. In: Werke, Bd. 3/1. Hgg. von Hendrik Birus. Frankfurt am Main 1994.

Guhr, Svenja: Trends in den Computational Literary Studies bei den DHd-Jahrestagungen. 2025. In: DHd Blog, https://dhd-blog.org/?p=22316 (Zugriff: 20.10.2025).

Kley, Antje/Niefanger, Dirk: Vorstellung Graduiertenkolleg 2806/1 - Literatur und Öffentlichkeit in differenten Gegenwartskulturen. 2023. URL: https://www.literaturundoeffentlichkeit.phil.fau.de/ (Zugriff: 20.10.2025).

Kramer, Frank: Ein allgemeiner Ansatz zur MetadatenVerwaltung. Berlin: Logos Verlag 2019.

Marx, Leonie: Die deutsche Kurzgeschichte. Stuttgart/Weimar: Metzler 2005.

Meyer, Anne-Rose: Die deutschsprachige Kurzgeschichte: eine Einführung. Berlin: Erich Schmidt 2014.

Neumann, Birgit/Nünning, Ansgar: Einleitung: Probleme, Aufgaben und Perspektiven der Gattungstheorie und Gattungsgeschichte. In: Gattungstheorie und Gattungsgeschichte. Hgg. von Marion Gymnich, Birgit Neumann und Ansgar Nünning. Trier: WVT Wissenschaftlicher Verlag Trier 2007.

OpenRefine Documentation (Last updated by Lohmeier, Felix): Reconsiling. 2025. URL: https://openrefine.org/docs/manual/reconciling (Zugriff: 20.10.2025).

OpenRefine Documentation (Last updated by Morris, Tom): URL: https://openrefine.org/docs/manual/cellediting#cluster-and-edit (Zugriff: 20.10.2025).

Pandas: pandas. 2025. URL: https://pandas.pydata.org/ (Zugriff: 20.10.2025).

pgAdmin: pgAdmin - PostgreSQL Tools. 2025. URL: https://www.pgadmin.org/ (Zugriff: 20.10.2025).

Pohl, Adrian/Steeg, Fabian: GND reconciliation for OpenRefine. 2019. URL: https://blog.lobid.org/2018/08/27/openrefine.html (Zugriff: 20.10.2025).

PostgreSQL: PostgreSQL: The World's Most Advanced Open Source Relational Database. 2025. URL: https://www.postgresql.org/ (Zugriff: 20.10.2025).

Schneider, Wolfram: Ein verteiltes Bibliotheks-Informationssystem auf Basis des Z39.50-Protokolls“. ZIB, 1999.

Schubert, Charlotte: Digital Humanities auf dem Weg zu einer Wissenschaftsmethodik: Transparenz und Fehlerkultur. In: Digital Classics Online, Bd. 7. 2021. DOI: https://doi.org/10.11588/dco.2021.7.82371.

Seenivasan, Dhamotharan: ETL (Extract, Transform, Load) Best Practices. In: International Journal of Computer Trends and Technology (IJCTT), vol. 71, no. 1, pp. 40-44, 2023. URL: http://dx.doi.org/10.2139/ssrn.5148200.

Vaisman, Alejandro/Zimányi, Esteban: Data Warehouse Systems. Data-Centric Systems and Applications. Berlin/Heidelberg: Springer-Verlag 2014. DOI: https://doi.org/10.1007/978-3-642-54655-6.

Würz, Maximilian: Arbeitsbericht: Die Kurzgeschichte und kurze Erzählformen in Theorie und Praxis. In: Erlanger Beiträge zur Medienwirtschaft, Institut für Buchwissenschaft, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU). No. 16/2023. 2023.

Würz, Maximilian: kurze-Erzaehlformen [GitHub Repository]. 2025. URL: https://github.com/maximilianwuerz/kurze-Erzaehlformen. Initial Release: DOI: 10.5281/zenodo.17511770.

Würz, Maximilian: Projektbeschreibung „Kurzes Erzählen im 21. Jahrhundert - Entstehungsbedingungen und Aneignungspotenziale von Kurzen Erzählformen in der Gegenwart“. 2022. URL: https://www.literaturundoeffentlichkeit.phil.fau.de/person/kurzes-erzaehlen-im-21-jahrhundert/ (Zugriff: 20.10.2025).

Zymner, Rüdiger: Gattungstheorie. Probleme und Positionen der Literaturwissenschaft. Paderborn: mentis Verlag GmbH 2007.