GND und Normdaten für europäische Literatur? Personen und Werke in den multilingualen Korpora von ELTeC

Calvo Tello, José
calvotello@sub.uni-goettingen.de
Niedersächsische Staats- und Universitätsbibliothek Göttingen, Georg-August-Universität Göttingen

Rißler-Pipka, Nanette
nanette.rissler-pipka@gwdg.de
Gesellschaft für wissenschaftliche Datenverarbeitung mbH (GWDG)

Barth, Florian
florian.barth@uni-goettingen.de
Niedersächsische Staats- und Universitätsbibliothek Göttingen, Georg-August-Universität Göttingen

GND, Normdaten und die Digital Humanities

Viele Projekte in den Digital Humanities verwenden Identifier für Entitäten wie Personen, Werke, Körperschaften oder Orte, die auf eindeutige Einträge in Normdaten-Verzeichnissen oder in Knowledge Bases verweisen (Barth 2022 u. a.; Rosenkötter und Fischer 2020; Fischer und Jäschke 2018; Herrmann und Lauer 2018; Dieckmann, Hermes, und Neuefeind 2017). Zu diesen Ressourcen gehören u. a. die Gemeinsame Normdatei (GND), andere Normdaten von Nationalbibliotheken, Wikidata, VIAF, DBpedia, Getty oder CERL. Jede dieser Ressourcen ist nach verschiedenen Kriterien aufgebaut und bietet unterschiedliche Funktionalitäten. Dies bringt Vor- und Nachteile für die Projekte mit sich, die sie nutzen. Zum Beispiel hat Wikidata keinen echten Normierungscharakter im Gegensatz zu Normdaten-Verzeichnissen wie der GND. Wikidata hat jedoch den Vorteil, dass Nutzende selbständig neue Entitäten anlegen können. Dies ist bei den durch Bibliotheken verwalteten Normdaten-Verzeichnissen meist nur auf Antrag möglich (in Zukunft sollen, zumindest für die GND, die Eingabemöglichkeiten durch angepasste Webformulare und Redaktionsumgebungen erweitert werden, vgl. Kett u. a. 2022).

Nichtsdestotrotz ist die Sprache des Forschungsobjekts (z. B. von Textkorpora) und die Wahl der Normdaten-Ressource stark voneinander abhängig. Einige Projekte, die mit deutschsprachigen Texten arbeiten, haben sich für die GND entschieden, um Personen oder Werke zu identifizieren, u. a. die Digitale Bibliothek im TextGrid Repository, [1] das Deutsche Textarchiv [2] oder die deutschsprachigen Korpora aus DraCor (Fischer u. a. 2019) und ELTeC (Burnard, Schöch, und Odebrecht 2021). Projekte aus der deutschsprachigen Wissenschaftslandschaft, die im Bereich der nicht-deutschen Philologien forschen, entscheiden sich eher für andere Ressourcen, um Personen und Werke zu identifizieren. Die romanistischen Korpora der CLiGS-TextBox (Französisch, Spanisch, Italienisch und Portugiesisch; Schöch u. a. 2019), die spanischen Korpora CoNSSA (Calvo Tello 2021) und CONHA (Henny-Krahmer 2018), und die französischen Korpora in ELTeC und DraCor benutzen für die Identifikation ihrer Entitäten überwiegend Wikidata und VIAF.

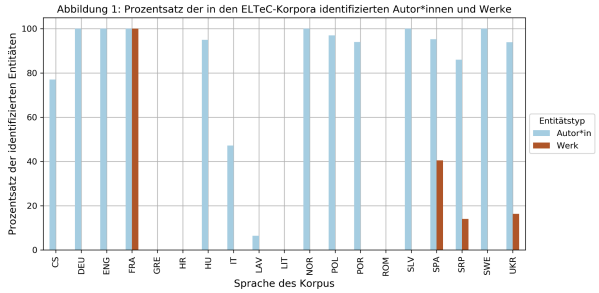
Die Bibliotheken und Fachinformationsdienste (FIDs) vieler Philologien in Deutschland verwenden die GND für die Sacherschließung ihrer Titel. Sie reichern umgekehrt die GND mit immer mehr Daten zu fremdsprachigen Autor*innen und deren Werke an. Es liegt nahe, dass die GND weiterhin hauptsächlich Autor*innen und Werke aus dem deutschsprachigen Raum verzeichnet. Das heißt jedoch nicht, dass die GND eine Quelle ist, die sich nur für die Germanistik eignet. Generell ist die GND in Form von Agenturen organisiert, die sich auf viele Bibliotheken und andere Institutionen in Deutschland verteilen und von der DNB koordiniert werden. [3] Seit 2020 spielt die GND außerdem eine neue Rolle durch das Projekt “GND für Kulturdaten” (Rosenkötter und Fischer 2020; Balzer u. a. 2019) und seit 2021 durch die Beteiligung in der NFDI [4] (Text+, NFDI4Culture). [5]

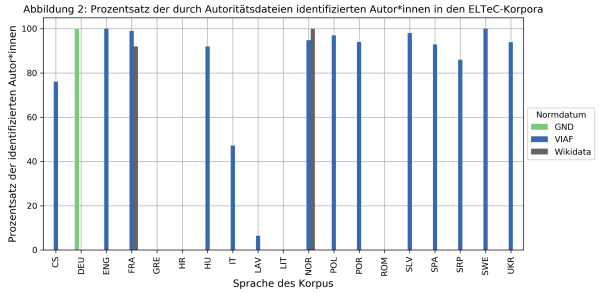
Daher fragen wir uns: Wie stark ist das Ungleichgewicht innerhalb der GND zwischen Einträgen zu deutsch- und fremdsprachiger Literatur? [6] Können die Anglistik, die Romanistik, die klassischen Philologien, die Slavistik und andere damit rechnen, ihre Entitäten in der GND zu finden?

ELTeC: mehrsprachige, vergleichbare Korpora für europäische Literatur

Unsere Frage beantworten wir anhand der multilingualen Korpora von ELTeC. Dabei handelt es sich um literarische Korpora in verschiedenen europäischen Sprachen, die in der COST-Action Distant Reading erstellt wurden (Schöch u. a. 2021; Odebrecht, Burnard, und Schöch 2021). Das Ziel des Projektes war die Zusammenstellung vergleichbarer Korpora mit 100 Romanen pro Sprache. Aktuell wurde dies für 11 Sprachen erreicht, während für 10 andere Sprachen weniger Romane vorliegen. Außerdem wurde jedes Dokument mit Metadaten zu Autor, Werk, Edition und Text ausgezeichnet. Auf Grundlage dieser Metadaten konnten wir unsere Analysen erstellen.

Die Personen und Werke in ELTeC sind teilweise bereits mit Wikidata, GND oder VIAF eindeutig identifiziert. Die Abbildungen 1 und 2 zeigen im Vergleich die Annotation mit Normdaten für Autor*innen und Werke pro Sprache und die Wahl der Normdatenressource für die Identifikation von Autor*innen.





Für die Sprachen Griechisch (GRE), Kroatisch (HR), Litauisch (LIT) und Rumänisch (ROM) konnten offenbar keine Normdaten verwendet werden. Um die Vergleichbarkeit der Ergebnisse zu gewährleisten, werden diese Sprachen bei den folgenden Analysen ausgeschlossen. Außer für das französische (FR) und in kleinen Teilen für das spanische (SPA), serbische (SRP) und ukrainische (UKR) Korpus wurden keine Werknormdaten eingetragen. Für die Autor*innen wurde überwiegend VIAF genutzt. Nur das französische und norwegische Korpus wurde auch mit Wikidata und das deutsche Korpus als einziges mit GND-IDs versehen.

Methode

Wir gehen in zwei Schritten vor, um ein vollständiges Bild über die mögliche Abdeckung mit Normdaten zu erhalten. Zunächst extrahieren wir die IDs der Autor*innen aus den TEI-Dokumenten der ELTeC-Korpora. Anhand der IDs werden die fehlenden Identifier aus Wikidata, GND und VIAF extrahiert. Das gelingt für die GND über die API von Lobid, [7] und für Wikidata und VIAF durch ihre native API. Nach diesem Schritt erhalten alle Autor*innen eindeutige Identifier aus allen drei Ressourcen, falls Mappings gefunden werden konnten.

Im nächsten Schritt werden die Werke mit Rückgriff auf die Autor*innen-ID identifiziert. Auch wenn für vier ELTeC-Korpora Werk-IDs (überwiegend mit VIAF) bereits vom Projekt erfasst wurden, ignorieren wir diese, um die gleiche Methode für alle Korpora anzuwenden. Wir führen drei parallele *Reconciliation*-Prozesse mit den drei Ressourcen für jedes Werk durch. Genauer, werden alle Werke aller Autor*innen abgerufen, um die Ähnlichkeit zwischen dem Titel des Werks in ELTeC und allen Titeln der Werke der jeweiligen Autor*in aus den Normdaten zu vergleichen. Für jede mögliche Paarung wird ein statistischer Wert für die Ähnlichkeit zwischen beiden Titeln berechnet (0 für Titel, die gar keine Gemeinsamkeiten haben, 1 für Titel, die deckungsgleich sind). Der Werktitel mit der höchsten Ähnlichkeit wird ausgewählt. Für die weitere Analyse werden nur die Werke berücksichtigt, deren Wert höher als 0.5 liegt. Für diese drei parallelen und automatischen *Reconciliation*- Prozesse werden die APIs von Lobid, Wikidata und VIAF benutzt.

Ergebnisse

Ausgehend von den bereits in ELTeC identifizierten Autor*innen (vgl. Abb. 1) wird überprüft, ob diese auch in den jeweils anderen Normdatenressourcen (GND, Wikidata, VIAF) vorhanden sind. Daher sind hier Sprachen, für die keine Normdaten in ELTeC existieren (Griechisch, Kroatisch, Litauisch, Rumänisch), nicht berücksichtigt.

Ergebnisse zu Autor*innen

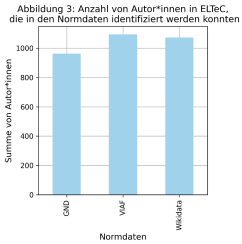
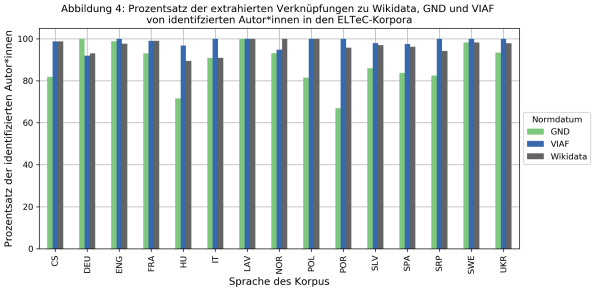


Abbildung 3 zeigt die Summe der Normdaten zu Autor*innen pro Ressource, die in den hier betrachteten Korpora gefunden oder ergänzt werden konnten. Für alle drei Ressourcen ist die Abdeckung hier sehr gut. Die GND liegt im Vergleich nur leicht zurück.

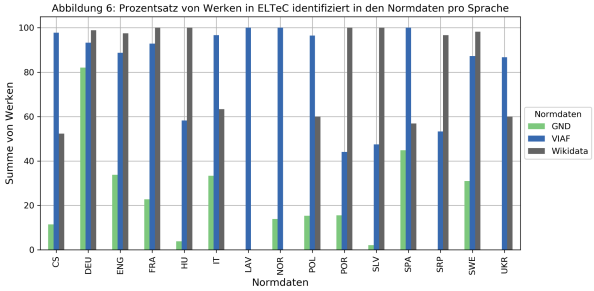
Die Verteilung dieser Daten pro Sprache wird in Abbildung 4 gezeigt. Das Bild entspricht der Zusammenfassung aus Abbildung 3. Erwartungsgemäß hat das deutsche Korpus die höchste Quote in der GND. Das norwegische Korpus kann mehr Treffer mit Wikidata als mit VIAF erzielen. Für Tschechisch und Polnisch erreichen sowohl VIAF als auch Wikidata sehr gute Ergebnisse. Neben dem Deutschen bietet die GND eine gute Abdeckung für Sprachen wie Englisch, Französisch, Italienisch, Norwegisch, Schwedisch und Ukrainisch.[8]



Ergebnisse zu Werken

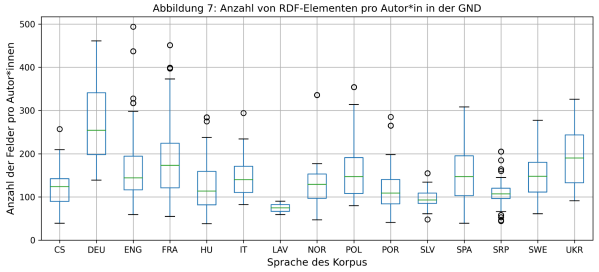
Das Bild ändert sich erwartungsgemäß, wenn nicht Autor*innen, sondern Werke in den drei Ressourcen gesucht werden (vgl. Abb.5). Während VIAF und Wikidata mehr als 700 Werke aus ELTeC verzeichnen, erreicht die GND nur knapp über 200.

Um diese Zahlen besser zu verstehen, zeigt Abbildung 6, dass nur das deutsche Korpus akzeptable Ergebnisse aus der GND erreicht (80 % der Werke). Alle anderen Sprachen bewegen sich zwischen null und knapp über 40 %. Die Abdeckung von Wikidata oder VIAF ist für viele Sprachen deutlich höher: von 60 % bis zu 100 %.

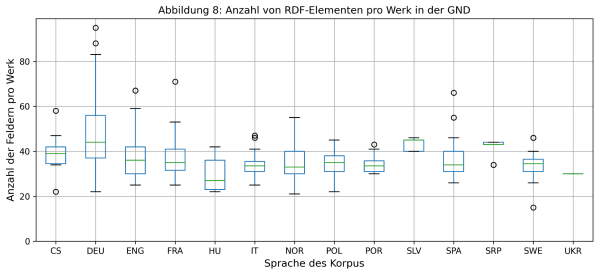


Anzahl der Informationen pro Entität

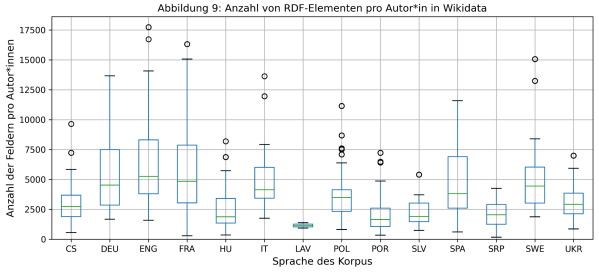
Entscheidend für eine Bewertung der Ressource ist nicht nur, ob eine Entität vorhanden ist, sondern wie gut sie mit Metadaten beschrieben ist. Für die GND ist zu erwarten, dass Entitäten aus dem deutschsprachigen Raum ausführlicher beschrieben werden als Entitäten aus anderen Regionen. Um dies zu messen, werden die Daten von allen ELTeC-Autor*innen und deren Werke als XML-RDF-Dokumente aus der GND heruntergeladen und die Anzahl der XML-Elemente quantifiziert. Ohne den semantischen Gehalt der Elemente zu bewerten, gehen wir davon aus, dass mehr Elemente auch mehr Informationen pro Entität bedeuten. Abbildung 7 zeigt daher für die GND die Anzahl der Elemente pro Autor*in. Während für Autor*innen aus dem deutschsprachigen Raum 200-330 Elemente vorhanden sind, werden nur 100-240 für andere Sprachen verzeichnet. Auch wenn Sprachen wie Französisch oder Ukrainisch mittlere Werte (bis 240) zeigen, ist der Abstand zwischen diesen Sprachen und dem Deutschen immer noch sehr groß.

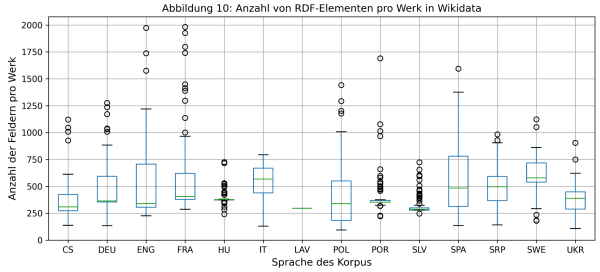


Für die Werke (vgl. Abb. 8) erreichen die deutschsprachigen Entitäten wieder deutlich höhere Werte als alle anderen Sprachen in der GND. Hier ist der Unterschied im Vergleich weniger groß, weil insgesamt für Werke weniger Elemente angelegt werden.



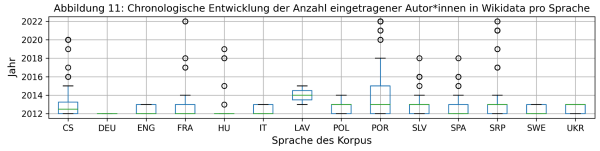
Für einen Vergleich wurden diese Daten auch aus Wikidata extrahiert (Abbildungen 9 und 10). Wir prüfen, ob in Wikidata ähnliche Verzerrungen gegenüber dem Englischen oder anderen Sprachen zu beobachten sind. Jedoch hat in Wikidata keine Sprache einen so klaren Vorsprung im Vergleich zu allen anderen Sprachen wie das Deutsche in der GND.

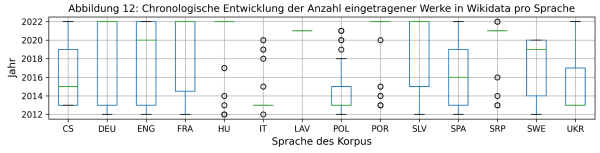




Anreicherung durch ELTeC

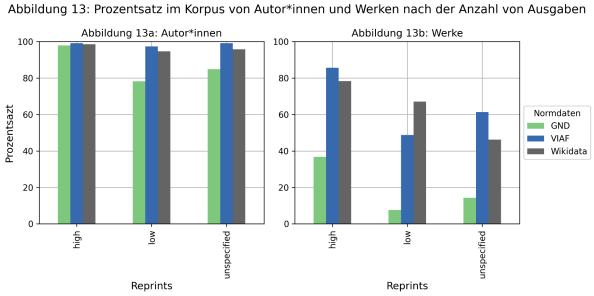
Insgesamt ist ein möglicher Grund für die bessere Abdeckung für Autor*innen und Werke in Wikidata und VIAF ist die Tatsache, dass die Teilnehmenden von ELTeC die Entitäten in Wikidata selbst eingetragen haben (Ne#i# u. a. 2022). Die Abbildungen 11 und 12 zeigen, dass zwar die Mehrheit der Autor*innen aus ELTeC bereits vor dem Start des Projekts (2017-2018) in Wikidata vorhanden waren, aber viele neue Werkeinträge entstanden. Die Metadaten zu Werken aus den sieben Sprachen, die Ne#i# u. a. (2022) ausgewählt haben, führen zu einer deutlichen Verbesserung der Abdeckung in den letzten Jahren.





Kanonisierungsgrad

Eine weitere Hypothese ist, dass der Kanonisierungsgrad die Abdeckung in den drei Ressourcen beeinflusst. In den ELTeC Korpora wurde dies anhand des Metadatenfelds “reprints” belegt. Für Autor*innen und Werke, die keine oder wenige “reprints” (“low”) haben, zeigt Abbildung 13, dass alle drei Ressourcen eine niedrigere Abdeckung haben. Dabei ist der Unterschied für die GND deutlich größer als bei VIAF und Wikidata. Die Daten deuten darauf hin, dass die GND stärker vom Kanonisierungsgrad beeinflusst ist als die anderen zwei Ressourcen. Besonders niedrig ist die Abdeckung von nicht kanonisierten Werken in der GND (Abb. 13, “low reprints”). Zu beachten ist, dass die Verteilung von solchen Metadaten in den ELTeC Korpora nicht gleichmäßig ist. Die Ergebnisse können dementsprechend allein durch die Zusammenstellung der Korpora und die Metadatenanreicherung beeinflusst sein.



Abschluss

Wie gut können nicht-germanistische Projekte aus dem deutschsprachigen Raum mit der GND Autor*innen und Werke identifizieren? Sollten sie lieber auf Wikidata oder VIAF zurückgreifen? Um das zu beantworten, wurden die multilingualen Korpora von ELTeC analysiert. Auch wenn diese Korpora nicht vollständig ausgewogen hinsichtlich Repräsentation der Inhalte und der Persistenz ihrer Identifier sein können, sind sie eine wertvolle Ressource von und für die Community. Weitere ähnliche Evaluationen könnten in Zukunft durchgeführt werden, wenn umfassenderes Vergleichsmaterial identifiziert oder zusammengestellt wird. Das ELTeC Korpus ist zeitlich (19. Jh.), quantitativ und sachlich (100 Romane pro Sprache) notwendig beschränkt und vor diesem Hintergrund sind auch die vorliegenden Ergebnisse zu betrachten.

Generell zeigt die GND eine gute Abdeckung von Personendaten und ist damit sehr nah an Wikidata oder VIAF. Jedoch füllt die GND deutlich mehr Felder (d.h. mehr Informationen) zu deutschen Autor*innen als zu anderen europäischen Autor*innen (für Personendaten außerhalb der Literatur mag das anders aussehen).

Hinsichtlich der Werknormdaten kann die GND nur für das Deutsche akzeptable Ergebnisse liefern. Auch für andere große Sprachen wie Französisch, Englisch oder Spanisch enthält die GND nur 40 % der enthaltenen Werke in ELTeC. Nicht nur die Abdeckung, sondern auch der Informationsgehalt ist für Werke deutschsprachiger Autor*innen höher als für alle anderen Sprachen. Darüber hinaus scheint die GND stärker vom Kanonisierungsgrad abhängig zu sein als VIAF oder Wikidata.

Wenn die GND damit als national-ausgerichtete Normdateninstitution erwartbar schlechter abschneidet, dann wäre zu prüfen, ob die Normdaten anderer Einrichtungen (vor allem von Nationalbibliotheken) eine ähnliche oder sogar stärkere Favorisierung der eigenen Sprache verzeichnen.

Durch die Einbindung der GND (der DNB und GND-Agenturen in anderen Bibliotheken) in die NFDI öffnet sich die GND nicht nur der Community, sondern es werden auch wichtige Diskussionen zu Multilingualität (GNDmul) [9] und Internationalität geführt. Wir sind zuver-

sichtlich, dass die GND mithilfe der Community den Anteil fremdsprachiger Werke und Autor*innen in Zukunft erhöhen kann. Innerhalb von Text+ versuchen wir, Normdaten und Forschungsdaten zu verknüpfen. Die in dieser Analyse verwendeten Skripte werden auch für die Entwicklung von Pipelines zur Datenanreicherung im TextGrid Repository genutzt. So können die neu identifizierten Personen und Werke aus den ELTeC-Korpora mit den entsprechenden IDs zu den Daten aus dem TextGrid Repository hinzugefügt werden. Umgekehrt konnte damit auch ELTeC neue Daten für die Korpora gewinnen.

Fußnoten

1. https://textgridrep.org/.
2. https://www.deutschestextarchiv.de/."
3. Vgl. GND-Partner: https://gnd.network/Webs/gnd/DE/UeberGND/Partner/partner_node.html; https://prezi.com/p/unl16mzwubbs/gndzoom/.
4. Nationale Forschungsdateninfrastruktur: https://www.nfdi.de/.
5. Text+: https://www.text-plus.org/; NFDI4Culture: https://nfdi4culture.de/.
6. Von den 501.913 Einzelwerken in der GND sind 57.857 Deutsch, 10.476 Englisch und 6.918 Französisch, 5.918 Italienisch, 2.419 Spanisch, 449 Portugiesisch, 390 Rumänisch, aber auch 12.329 Latein, 6.072 Griechisch und der Großteil ohne Sprachangabe: 379.872 (https://explore.gnd.network/search?f.satzart=Werk&f.land.limit=80&rows=25)
7. Vgl. Lobid-API, mit der GND-Einträge abgerufen werden können: https://lobid.org/gnd/api.
8. Das lettische Korpus (LAV) zählen wir nicht mit, weil der Balken in Abb. 4 zwar 100% für alle drei Ressourcen anzeigt, aber es sich insgesamt nur um 5 Romane handelt (vgl. Abb. 1).
9. https://gnd.network/Webs/gnd/DE/UeberGND/Partner/partner_node.html

Bibliographie

Barth, Florian, Varachkina, Hanna, Dönicke, Tillmann, und Luisa Gödeke. Levels of Non-Fictionality in Fictional Texts. In *Proceedings of ISA-18 Workshop at LREC2022*, 27–32. Marseille, 20 June 2022. http://www.lrec-conf.org/proceedings/lrec2022/workshops/ISA-18/pdf/2022.isa18-1.4.pdf.

Balzer, Detlev, Barbara K. Fischer, Jürgen Kett, Susanne Laux, Jens M. Lill, Jutta Lindenthal, Mathias Manecke, Martha Rosenkötter, und Axel Vitzthum. 2019. „Das Projekt ‚GND für Kulturdaten' (GND4C)". *o-bib. Das offene Bibliotheksjournal / Herausgeber VDB* 6 (4): 59–97. https://doi.org/10.5282/o-bib/2019H4S59-97.

Burnard, Lou, Christof Schöch, und Carolin Odebrecht. 2021. „In search of comity: TEI for distant reading". *Journal of the Text Encoding Initiative* , Nr. Issue 14 (März). https://doi.org/10.4000/jtei.3500.

Calvo Tello, José. 2021. *The Novel in the Spanish Silver Age: A Digital Analysis of Genre Using Machine Learning* . Digital Humanities Research 5. Bielefeld: transcript.

Dieckmann, Lisa, Jürgen Hermes, und Claes Neuefeind. 2017. „Bild, Beschreibung, (Meta)Text. Automatische inhaltliche Erschließung und Annotation kunsthistorischer Daten". In *Digitale Nachhaltigkeit* , 103–7. Bern: DHd. https://zenodo.org/record/3684825#.YuoMEBzP1aQ.

Fischer, Frank, Ingo Börner, Mathias Göbel, Angelika Hechtl, Christopher Kittel, Carsten Milling, und Peer Trilcke. 2019. „Programmable Corpora - Die digitale Literaturwissenschaft zwischen Forschung und Infrastruktur am Beispiel von DraCor". In *6. Tagung des Verbands Digital Humanities im deutschsprachigen Raum, DHd 2019, Frankfurt & Mainz, Germany, March 25-29, 2019* , herausgegeben von Patrick Sahle und Patrick Helling. https://doi.org/10.5281/zenodo.4622061.

Fischer, Frank, und Robert Jäschke. 2018. „Liebe und Tod in der Deutschen Nationalbibliothek". In *DHd2018: "Kritik der digitalen Vernunft"* , 261–66. Cologne, Germany: Digital Humanities im deutschsprachigen Raum. https://hal.archives-ouvertes.fr/hal-01787558.

Henny-Krahmer, Ulrike. 2018. „Exploration of Sentiments and Genre in Spanish American Novels". In *Puentes/Bridges* . México DF: ADHO. https://dh2018.adho.org/exploration-of-sentiments-and-genre-in-spanish-american-novels/.

Herrmann, J. Berenike, und Gerhard Lauer. 2018. „Korpusliteraturwissenschaft. Zur Konzeption und Praxis am Beispiel eines Korpus zur literarischen Moderne". *Osnabrücker Beiträge zur Sprachtheorie* , Nr. 92: 127–56.

Kett, Jürgen, Christoph Kudella, Andrea Rapp, Regine Stein, und Thorsten Trippel. 2022. „Text+ und die GND – Community-Hub und Wissensgraph". *Zeitschrift für Bibliothekswesen und Bibliographie* 69 (1–2): 37–47. https://doi.org/10.3196/1864295020691262.

Nešić, Milica Ikonić, Ranka Stanković, Christof Schöch, und Mihailo Skoric. 2022. „From EL-TeC Text Collection Metadata and Named Enti-ties to Linked-Data (and Back)". In *8th Workshop on Linked Data in Linguistics* , 7–16. Marseille: LREC. http://www.lrec-conf.org/proceedings/lrec2022/workshops/LDL/pdf/2022.ldl2022-1.2.pdf.

Odebrecht, Carolin, Lou Burnard, und Christof Schöch. 2021. „European Literary Text Collection (ELTeC): April 2021 release with 14 collections of at least 50 novels." Zenodo. https://doi.org/10.5281/zenodo.4662444.

Rosenkötter, Martha, und Barbara Fischer. 2020. „Normdaten der Faktenanker für Qualität im semantischen Retrieval. Der Ausbau der Gemeinsamen Normdatei (GND) im Projekt GND für Kulturdaten (GND4C)." In *Spielräume: Digital Humanities zwischen Modellierung und Interpretation* , 344–45. Paderborn: DHd. https://zenodo.org/record/3666690#.YuoNlRzP1aQ.

Schöch, Christof, Tomaz Erjavec, Roxana Patras, und Diana Santos. 2021. „Creating the European Literary Text Collection (ELTeC): Challenges and Perspectives". *Modern Languages Open* , Mai. https://doi.org/10.5281/ZENODO.4742419.

Schöch, Christof, José Calvo Tello, Ulrike Henny-Krahmer, und Stefanie Popp. 2019. „The CLiGS Textbox: Building and Using Collections of Literary Texts in Romance Languages Encoded in TEI XML". *Journal of the Text Encoding Initiative* , August. https://doi.org/10.4000/jtei.2085.