

Wikipedia als Hallraum der Kanonizität: »1001 Books You Must Read Before You Die«

Rohe, Jonas

jonas.rohe@fu-berlin.de
Freie Universität Berlin, Deutschland
ORCID: 0009-0002-8541-8520

Illmer, Viktor J.

v.illmer@fu-berlin.de
Freie Universität Berlin, Deutschland
ORCID: 0000-0002-7334-781X

Poggel, Lisa

l.poggel@fu-berlin.de
Freie Universität Berlin, Deutschland
ORCID: 0000-0003-1051-8148

Fischer, Frank

fr.fischer@fu-berlin.de
Freie Universität Berlin, Deutschland
ORCID: 0000-0003-2419-6629

Einführung

Wikipedia-Sitelinks (auch: Interwiki-Links) sind Verbindungen zwischen einem Wikipedia-Artikel und den entsprechenden Artikeln in anderen Sprachversionen der Wikipedia. Diese Links ermöglichen es, schnell zwischen verschiedenen Sprachversionen eines Lemmas zu wechseln. Für unsere Untersuchung ist die Anzahl verschiedener Sprachversionen zu einem Lemma von besonderem Interesse, wie anhand eines Beispiels demonstriert werden soll.

Derzeit gibt es aktive Wikipedia-Versionen für 331 verschiedene Sprachen (Wikipedia statistics 2024). Zum englischen Schriftsteller Charles Dickens gibt es etwa 162 Sitelinks, das heißt, es gibt über den Autor derzeit Artikel in 162 verschiedenen Sprachen in der Wikipedia. Zu seinem Roman »A Tale of Two Cities« gibt es momentan 51 Sitelinks (Wikidata 2024a).

Im Weltliteratur-Diskurs verbreitet sich seit einiger Zeit die Idee, die Anzahl von Wikipedia-Sitelinks als Teil der »Metrics of World Literature« zu verwenden (Robinson 2017), als »a simple measure of canonicity« (Kukkonen 2020, S. 244; vgl. auch Fischer et al. 2023). Das bloße Vorhandensein von mehreren Sprachversionen kann dabei als Kanonizitätsmarker verstanden werden. Diese Idee greifen

wir auf und möchten sie anhand eines konkreten Kanonprojekts entwickeln.

Der Sitelink-Metrik an die Seite stellen möchten wir darüber hinaus eine weitere Bewertungseinheit für Entitäten im Wikiversum, den sogenannten QRank. Hier werden die Seitenaufrufe zu einem Wikidata-Item, also Aufrufe von Wikipedia, Wikitravel, Wikibooks etc., zusammengerechnet und ein QRanking vergeben (<https://qrank.wmcloud.org/>). Demnach hätte, um bei unserem vorherigen Beispiel zu bleiben, Charles Dickens einen QRank von 4.158.626, sein Roman »A Tale of Two Cities« einen QRank von 1.238.166 (was der schlechtere Wert ist, das heißt, der Autor steht in diesem Fall besser da als sein Roman).

Das Korpus

Für unsere Analyse greifen wir auf das von Peter Boxall herausgegebene Kanonisierungsprojekt *1001 Books You Must Read Before You Die* (2006) zurück. Dieses Werk, veröffentlicht als dickleibige Bände in verschiedenen Ausgaben und Auflagen, präsentiert einen Kanon (genauer gesagt: bildungsbürgerlichen Kanon, vgl. Winko 2007) aus ein-tausendundeins Büchern, zumeist Romanen, die von über 150 Beitragenden zusammengestellt und mit jeweils einer Kurzzusammenfassung betextet wurden.

Werke der anspruchsvollen Literatur werden in der chronologischen Aufstellung neben Werke der Populärkultur gereiht: Zwischen Umberto Ecos *The Name of the Rose* und Klassikern der englischen Literatur wie *Jane Eyre* oder *Wuthering Heights* der Brontë-Schwestern finden sich Titel der Populärliteratur wie Douglas Adams' *Dirk Gently's Holistic Detective Agency* oder Bret Easton Ellis' Roman *American Psycho*. In den bisher fünf Auflagen von *1001 Books You Must Read Before You Die* (2006, 2008, 2010, 2012, 2018) wurde die Liste der Werke, welche man nach Boxall »gelesen haben sollte, bevor das Leben vorbei ist«, peu à peu aktualisiert. Einige Titel wurden gestrichen, um andere hinzufügen zu können, wobei jeder Band nach wie vor 1001 Werke präsentierte. Auf diese Weise ist die Zahl der 1001 Bücher des ersten Bandes über die verschiedenen Ausgaben hinweg auf insgesamt 1318 Werke gewachsen, die das (Metadaten-)Korpus für unsere Analyse bilden.

Die Änderungen an der Textauswahl in den verschiedenen Editionen lassen sich auch auf die Kritik zurückführen, welche an diesem Kanonprojekt an verschiedenen Stellen geäußert wurde. Philip Hensher etwa beschrieb das Werk 2006 als einen »very loose canon«; in seinem Review im *Spectator* kritisierte er ausführlich die Textauswahl: »All the same, this list could be a little less terrible. It is so amazing a series of obvious omissions, weird inclusions and horrid middle-aged attempts at grooviness that you wonder who on earth it is intended for.« Einige der Titel, deren Auslassung Hensher moniert, wie der japanische Klassiker *The Tale of Genji* oder die Romane *Midaq Alley* und *Miramar* von Naguib Mahfouz finden sich ab der zweiten Ausgabe im Band.

Datensatz und Workflow

Wie bei anderen Kanoninitiativen hat sich auch um die *1001 Books* eine Lese-Community gebildet, die bis heute aktiv ist. Beobachten lässt sich dies unter anderem auf der zu Amazon gehörigen Plattform Goodreads. Zu den fünf Ausgaben des Kanons wurde eine »Listopia«-Liste erstellt und alle 1318 Bücher dort eingruppiert, inklusive Goodreads-internen Links auf die einzelnen Werke. Besser dient unseren Zwecken aber ein dort verlinktes Google-Spreadsheet, in welchem die Autor*innen, die Originaltitel, das Vorkommen bzw. Nicht-Vorkommen der Werke in den fünf Ausgaben der *1001 Books* sowie weitere Metadaten verzeichnet sind.

Diese Tabelle ist unser Ausgangspunkt für die Anreicherung mit Normdaten mithilfe von OpenRefine, das wir in der Version 3.8.2 genutzt haben. Über den Reconciling Service konnten wir alle Autor*innen und eine Großzahl der Werke mit ihren Wikidata-Einträgen verknüpfen. Insgesamt haben von den 1318 Werken bis dato 1257 einen Wikidata-Eintrag. Die mit Wikidata-IDs angereicherte Tabelle haben wir neben anderen Materialien und dem Code in unserem GitHub-Repositorium veröffentlicht (Rohe et al. 2024 bzw. <https://github.com/temporal-communities/1001-books>).

Die Verknüpfung mit Wikidata ermöglicht den Zugriff auf weitere Metadaten. Dadurch kann der *1001 Books* -Kanon hinsichtlich des Geschlechts und der Nationalität der vertretenen Autor*innen sowie der Sprach- und Erstveröffentlichungsdaten der literarischen Werke statistisch beschrieben werden.

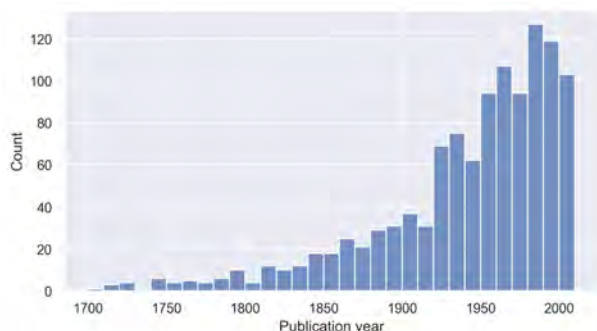


Abbildung 1. Histogramm der Publikationsjahre aus Wikidata (n = 1173). Zur besseren Lesbarkeit des Diagramms wurden die 27 Werke, die vor 1700 veröffentlicht wurden, hier ausgelassen.

Die Visualisierung der Publikationsdaten verdeutlicht, dass der Großteil der Werke im 20. Jahrhundert erschienen ist (Abb. 1). Ein Blick auf das Geschlecht der Verfasser*innen zeigt eine deutlich ungleiche Verteilung: Werke von 281 Autorinnen stehen Werke von 1031 Autoren gegenüber; ein Roman wurde von einer non-binären Person geschrieben (die Differenz zu den insgesamt 1318 Werken ist durch einige Anonymate bedingt). Am stärksten vertreten sind Charles Dickens und J. M. Coetzee mit jeweils 10 Wer-

ken, gefolgt von Virginia Woolf (9), Graham Greene, Don DeLillo, Ian McEwan, Philip Roth und Samuel Beckett (jeweils 8). Schon diese Auflistung legt nahe, dass im Kanon der *1001 Books* englischsprachige Werke überrepräsentiert sind. Für ein klareres Bild haben wir die einzelnen Werke mit Geocodes korreliert und dabei mit der Gemeinsamen Normdatei (GND) auf eine weitere Datenquelle zurückgegriffen.

In den GND-Datensätzen zu literarischen Werken sind *Geographic Area Codes* (Ländercodes) enthalten, die den Entstehungsraum des jeweiligen Werks beschreiben und entweder direkt oder über das Herkunftsland der Verfasser*innen oder die Sprache des Textes ermittelt wurden. Wenn das Werk eine*n Verfasser*in hat, wird als Ländercode bei geografischer Übereinstimmung das Herkunftsland der Verfasser*innen verwendet und sonst der Ursprungsort des Werks (Deutsche Nationalbibliothek 2024, S. 19). Für Werke ohne Verfasser*in oder bekannten Ursprungsort wird der Ländercode aufgrund der Sprache des Originaltextes zugeordnet. Insofern entspricht die Zuordnung der Wikidata-Property P495 »country of origin« für (literarische) Werke (Wikidata 2024b).

Zur Abfrage der Ländercodes wurde die durch lobid bereitgestellte Schnittstelle zur GND (lobid-gnd) genutzt. Um die gesuchten Einträge zu identifizieren, wurden die Ergebnisse der Datenabfrage mit den Wikidata-Werk-IDs oder verschiedenen Namensvarianten der Verfasser*innen und Titel der Werke abgeglichen. Falls für ein Werk kein Datensatz in der GND gefunden werden konnte, wurde die erwähnte Wikidata-Property P495 zur geografischen Zuordnung der Werke herangezogen und das jeweilige Land einem GND-Ländercode zugeordnet. Die GND-Ländercodes sind mit GeoNames-URIs verknüpft, die zur Abfrage der Längen- und Breitengrade über die Schnittstelle der geografischen Datenbank GeoNames verwendet wurden. Insgesamt konnten GeoNames-IDs für 1144 Werke ermittelt werden, 857 davon über die GND und 287 über Wikidata.

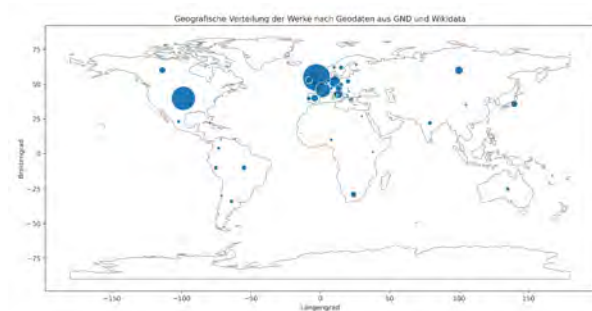


Abbildung 2. Geografische Verteilung der Werke (Geodaten aus GND und Wikidata).

Trotz der von Boxall im Vorwort zur zweiten Auflage angekündigten Diversifizierung des Kanons durch Erweiterung um nicht-englischsprachige Werke (vgl. Boxall 2008) wird nach Betrachtung der Geocodes offenkundig, dass sich

die Auswahl weiterhin stark auf englischsprachige Texte konzentriert und sich insgesamt ein eurozentristischer Blick auf Weltliteratur manifestiert, der sich auch in den Neuauflagen nicht grundlegend ändert (Abb. 2). Es dominieren Texte aus Großbritannien (347), gefolgt von den USA (266), Frankreich (106), Deutschland (57) sowie Italien und Russland (je 30).

Boxall schreibt: »Does a body of writing, a canon of essential texts, emerge from a national context, or does it in some way transcend nationality, rising above the contexts that generate it? What does it mean to try to respond to all of these different national contexts at the same time? Is it possible to produce a list that can speak at once to the readers in Turkey and in Greece, in Serbia and Croatia?« (Boxall 2008)

Dies kann anhand der Metadaten zu den ausgewählten Texten verneint werden. Der Kanon der *1001 Books* bleibt seinem nationalen Kontext – der Herausgeber Boxall selbst lehrt an der britischen Universität von Sussex – durchaus verhaftet. Griechische, türkische oder serbische Leser*innen gleichermaßen anzusprechen, erscheint schwierig, wenn nur sechs griechische, zwei serbische und ein einziges türkisches Werk (Orhan Pamuks *Snow*) in die Anthologie aufgenommen wurden.

Sitelinks und QRanks

Nachdem die Zusammensetzung des Korpus beschrieben wurde, wenden wir uns nun der eigentlichen Analyse zu. Dafür betrachten wir die Wikipedia-Sitelinks und den QRank der aufgeführten Werke und Autor*innen als »simple measure of canonicity«.

Mithilfe eines Python-Jupyter-Notebooks wurden entsprechende Metadaten von Wikidata abgefragt: zum einen alle Wikipedia-Artikel über Autor*innen und Werke aller Spracheditionen (*Sitelinks*), zum anderen das Publikationsjahr der Werke. Zusätzlich wurden sowohl für Werke als auch für Autor*innen der Wikidata QRank (Brawer 2024) ergänzt (Stand: 16.03.2024). Dieser Ranking-Wert, der auf einer Kombination aus Sitelinks und Pageviews basiert, soll zusätzlich Robustheit gegenüber saisonalen und kurzfristigen Popularitätseffekten bieten.

Werk- und Autor*innendaten sind somit mit zwei verwandten langfristigen »Relevanzmaßen« angereichert. Die vollständige Tabelle mit den Relevanzmaßen kann in unserem GitHub-Repository eingesehen bzw. live berechnet werden.

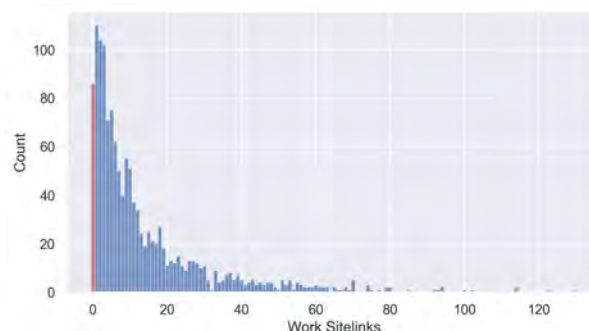


Abbildung 3. Histogramm der Anzahl an Wikipedia-Sitelinks für Werke (n = 1318, h = 1).

Ganz links in Abb. 3 zeigt sich, dass 90 der insgesamt 1318 Werke zwar einen Wikidata-Eintrag, aber keinen einzigen zugehörigen Wikipedia-Artikel haben. Die größte Gruppe von Werken hat eine n einzigen Wikipedia-Artikel (109). Die Top-4 bilden: *The Thousand and One Nights* (130), *Don Quixote* (123), *A Dream of Red Mansions* (102), *The Tale of Genji* (100). Werke besitzen im Mittel 13,7 und im Median 7 Sitelinks.

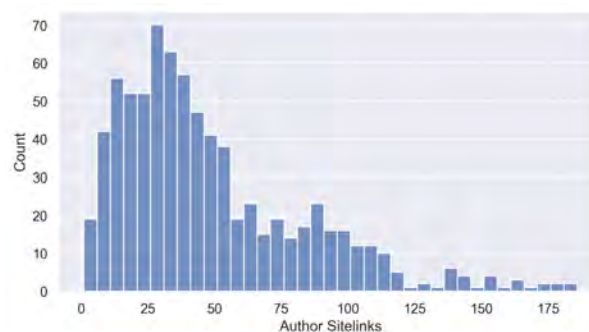


Abbildung 4. Histogramm der Anzahl an Wikipedia-Sitelinks für Autor*innen (n = 768, h = 5).

Im Gegensatz zu den Werken tun sich bei den Autor*innen keine Wikipedia-Lücken auf, alle haben mindestens einen zugehörigen Wikipedia-Artikel (Abb. 4). Die Mittelwerte liegen hier deutlich höher: Autor*innen haben im Durchschnitt 58,6 Sitelinks mit einem Medianwert von 48.

Tabelle 1 zeigt für die Autor*innen die Korrelation zwischen QRanks und Sitelinks, konkret die zehn Autor*innen mit den höchsten QRanks.

Tabelle 1. QRanks und Sitelinks der Autor*innen, gerankt nach Author QRank.

Author	Author QRank	Author Sitelinks
Poe, Edgar Allan	6.770.125	145
Kafka, Franz	5.827.580	166
King, Stephen	5.752.199	110
Hemingway, Ernest	5.732.463	148
Dostoevsky, Fyodor	5.714.591	173
Christie, Agatha	5.620.277	139
Tolkien, J.R.R.	5.533.813	154
Tolstoy, Leo	5.046.400	182
Wilde, Oscar	4.522.706	127
Dickens, Charles	4.158.626	162

Rankt man die Top-10 nach Anzahl der Sitelinks, ergeben sich interessante Änderungen (Tab. 2). Hier befindet sich nur ein englischsprachiger Autor unter den ersten zehn.

Tabelle 2. QRanks und Sitelinks der Autor*innen, gerankt nach Author Sitelinks.

Author	Author QRank	Author Sitelinks
Goethe, Johann Wolfgang von	3.380.565	186
Tolstoy, Leo	5.046.400	182
Hugo, Victor	3.917.537	176
Cervantes Saavedra, Miguel de	2.272.986	176
Dostoevsky, Fyodor	5.714.591	173
Voltaire	3.293.244	171
Kafka, Franz	5.827.580	166
Pushkin, Alexander	4.027.476	164
Dickens, Charles	4.158.626	162
Tagore, Rabindranath	4.068.155	161

Für die Werke haben wir die QRank-Sitelink-Korrelation visualisiert (Abb. 5). Erwartungsgemäß zeigt sich ein positiver Zusammenhang zwischen der Anzahl an Sitelinks und dem QRank der Werke. Werke mit einer höheren Anzahl an Sitelinks weisen im Durchschnitt einen höheren QRank auf. Es wurde eine quadratische Regression durchgeführt, um den Zusammenhang zwischen der Anzahl an Werk-Sitelinks (x) und dem Werk-QRank (y) zu untersuchen. Diese konnte einen Wert von 0,72 für das Bestimmtheitsmaß R^2 erreichen, was darauf hinweist, dass 72 % der Varianz der QRank-Variable durch die Anzahl der Sitelinks und die quadratische Komponente erklärt werden.

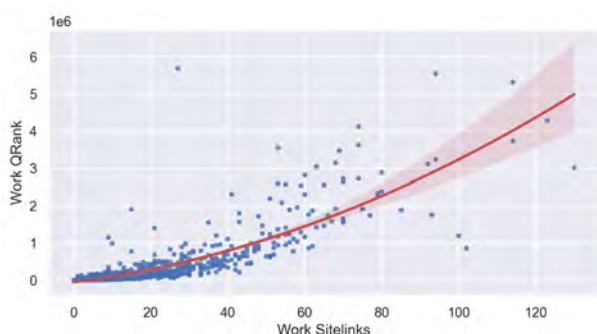


Abbildung 5. Zusammenhang zwischen Anzahl an Werk-Sitelinks und Werk-QRank mit quadratischer Regressionslinie (n = 1230, $R^2 = 0,72$).

Nennenswerte Ausreißer sind *The Fall of the House of Usher* von Edgar Allan Poe, welches mit 5.547.853 den höchsten QRank unter allen Werken in unserem Korpus aufweist, allerdings nur 27 Sitelinks hat. Ein ähnlicher Effekt, wenn auch auf niedrigerem Niveau, ist bei Cormac McCarthys *Blood Meridian* zu beobachten, welches einen QRank von 1.912.318 jedoch nur 15 Sitelinks aufweist. Eventuell lässt sich diese Diskrepanz dadurch erklären, dass beide Texte im englischsprachigen Bildungskontexten oft als Lektüre verwendet werden und daher Schüler*innen und Studierende öfter darauf zugreifen.

Eine weitere interessante Diskrepanz: Orwells *Nineteen Eighty-Four* hat einen höheren QRank als der Autor (5.547.853 Werk gegenüber 3.995.278 Author); dasselbe

gilt für *The Handmaid's Tale* (1.799.096) von Margaret Atwood (1.156.848). Hier stehen die Werke prominenter als ihre Autor*innen da.

Zusammenfassung und Ausblick

Insgesamt hat sich gezeigt, dass sich Wikipedia und besonders die Wikipedia-Sitelinks gut eignen als »Hallraum für Kanonizität«. Zu ausnahmslos allen Autor*innen sowie über 90% der Werke des Beispielkorpus gibt es mindestens einen Wikipedia-Artikel. Die Korrelation mit QRank hat gezeigt, dass es andere hilfreiche Ranking-Varianten gibt, die sich ebenso leicht heranziehen lassen.

Nun sagt die schiere Existenz eines Artikels in einer Sprachversion noch nichts über deren Qualität aus. Für diese weiterführenden Zwecke haben wir mit WikiMetrix bereits ein anderes Tool zur Verfügung gestellt (Illmer et al. 2024). Auch den vorliegenden Beitrag veröffentlichen wir zusammen mit dem zugehörigen Programmcode, der prinzipiell auch auf andere Kanonzusammenstellungen anwendbar ist.

Bibliographie

Boxall, Peter. 2006. *1001 Books You Must Read Before You Die*. London: Cassell Illustrated.

Boxall, Peter. 2008. *1001 Books You Must Read Before You Die*. Zweite Ausgabe. London: Cassell Illustrated.

Brawer, Sascha. 2024. »Wikidata QRank: Technical Design.« <https://github.com/brawer/wikidata-qrank/blob/main/doc/design.md> (zugegriffen: 26. November 2024).

»Data:Wikipedia statistics/meta.tab.« Stand vom 19. Juli 2024. Permalink: https://commons.wikimedia.org/w/index.php?title=Data:Wikipedia_statistics/meta.tab&oldid=900618439 (zugegriffen: 26. November 2024).

Deutsche Nationalbibliothek. 2023. »GND Geographic Area Codes.« <https://d-nb.info/standards/vocab/gnd/geographic-area-code.html> (zugegriffen: 26. November 2024).

Deutsche Nationalbibliothek. 2024. »Der Ländercode (LC) in der Gemeinsamen Normdatei (GND) – Leitfaden zu seiner Vergabe.« <https://wiki.dnb.de/download/attachments/90411323/Laendercodeleitfaden.pdf> (zugegriffen: 26. November 2024).

Delpuch, Antonin et al. 2024. OpenRefine/OpenRefine: Zenodo. Version 3.8.2. <https://doi.org/10.5281/zenodo.12689605>

Fischer, Frank, Jacob Blakesley, Paula Wojcik, Robert Jäschke. 2023. »Preface: World Literature in an Expanding Digital Space.« In: *Journal of Cultural Analytics*. Vol. 8, Nr. 2. <https://doi.org/10.22148/001c.74598>

GeoNames. 2024. GeoNames Web Services Documentation. <https://www.geonames.org/export/web-services.html#get> (zugegriffen: 26. November 2024).

GeoPandas Development Team. 2024. GeoPandas Version 1.0.1. <https://zenodo.org/records/12625316>

Illmer, Viktor J., Bart Soethaert, Lilly Welz, Frank Fischer, Robert Jäschke. 2024. »Literatur im Wikiversum – Eine praktische Annäherung über API-Abfragen und Wikipedia-Metriken.« In: *DHd2024: »DH Quo Vadis«*. Book of Abstracts. Universität Passau. <https://doi.org/10.5281/zenodo.10698426>

Kukkonen, Karin. 2020. »Does Cognition Translate? Predictions, Plot, and World Literature.« In: *Poetics Today*. Vol. 41, Nr. 2, S. 243–259. <https://doi.org/10.1215/03335372-8172556>

lobid. 2024. lobid-gnd API. <https://lobid.org/gnd/api> (zugegriffen: 26. November 2024).

Matplotlib Development Team. 2024. Matplotlib: Visualization with Python. Zenodo. Version 3.9.1. <https://doi.org/10.5281/zenodo.12652732>

N. N. 2024. »Listopia: 1001 Books You Must Read Before You Die.« In: *goodreads.com*. https://www.goodreads.com/list/show/952.1001_Books_You_Must_Read_Before_You_Die (zugegriffen: 16. Juli 2024).

Robinson, Douglas. 2017. »Metrics of World Literature.« In: (ders.): *Aleksis Kivi and/as World Literature*. Leiden und Boston: Brill, S. 43–83. https://doi.org/10.1163/9789004340268_003

Rohe, Jonas, Viktor J. Illmer, Lisa Poggel, Frank Fischer. 2024. »1001 Books You Must Read Before You Die.« GitHub-Repositorium. <https://github.com/temporal-communities/1001-books>

Vink, Ritchie. 2024. Python Polars 1.2.0. Zenodo. <https://doi.org/10.5281/zenodo.12751859>

Waskom, Michael. 2021. »seaborn: statistical data visualization.« In: *Journal of Open Source Software* 6 (60), 3021. Version 0.13.2. <https://doi.org/10.21105/joss.03021>

Wikidata. 2024a. Charles Dickens (Q5686). Permalink: <https://www.wikidata.org/w/index.php?title=Q5686&oldid=2199384142>. A Tale of Two Cities (Q308918). Permalink: <https://www.wikidata.org/w/index.php?title=Q308918&oldid=2183733159>. (zugegriffen: 19. Juli 2024).

Wikidata. 2024b. Property P495 country of origin. <https://www.wikidata.org/wiki/Property:P495> (zugegriffen: 26. November 2024).

Winko, Simone. 2007. »Textbewertung.« In: Thomas Anz (Hg.): *Handbuch Literaturwissenschaft*. Band 2. Stuttgart/Weimar: Metzler, S. 233–266. https://doi.org/10.1007/978-3-476-01271-5_13