

# Personen in GND und Titeldaten

## Vorschläge, Anreicherungen und Verknüpfungen

Stephanie Glagla-Dietz, Deutsche Nationalbibliothek, Frankfurt am Main

Stefan Grund-Davidov, Deutsche Nationalbibliothek, Frankfurt am Main

Nicole Habermann, Deutsche Nationalbibliothek, Frankfurt am Main

Angela Vorndran, Deutsche Nationalbibliothek, Frankfurt am Main

### Zusammenfassung

Seit 2011 werden in der Deutschen Nationalbibliothek (DNB) maschinelle Verfahren zur Anreicherung bibliografischer Daten eingesetzt. Speziell für den Bereich Personen wird diese Entwicklung seit 2019 besonders vorangetrieben. Der Artikel berichtet über die Verfahren zur Anreicherung von Daten mit Standardnummern und zur Verknüpfung von Titeldaten mit Normdaten der Gemeinsamen Normdatei (GND) mithilfe der Standardnummern. In einem weiteren Abschnitt wird die Übernahme von GND-Verknüpfungen aus Culturegraph-Werk- und Manifestationsbündeln besprochen. Der Beitrag wirft auch einen Blick auf zukünftige Entwicklungen, wie beispielsweise die Auswertung von Ko-Autorenschaften.

### Summary

Since 2011 the German National Library (DNB) has applied machine-based techniques to enrich bibliographic metadata records. From 2019 on, DNB has extended these activities focusing particularly on persons. This paper describes the enrichment of metadata records with controlled identifiers and with links to authority control records from the Integrated Authority File (GND) derived from these identifiers. Furthermore, we discuss how to transfer links to GND records for persons derived from Culturegraph work and manifestation clusters. The article also takes a look at future developments, such as the evaluation of co-author relationships.

**Zitierfähiger Link:** <https://doi.org/10.5282/o-bib/5998>

### Autorenidentifikation:

Stephanie Glagla-Dietz: GND: <https://d-nb.info/gnd/115526935>, ORCID: <https://orcid.org/0000-0001-8762-3005>

Stefan Grund-Davidov: GND: <https://d-nb.info/gnd/1300858966>, ORCID: <https://orcid.org/0009-0008-1572-3654>

Nicole Habermann: GND: <https://d-nb.info/gnd/1218808381>, ORCID: <https://orcid.org/0000-0002-4103-101X>

Angela Vorndran: GND: <https://d-nb.info/gnd/1301084379>, ORCID: <https://orcid.org/0000-0001-7162-9875>

**Schlagwörter:** Gemeinsame Normdatei (GND), Deutsche Nationalbibliothek, Normdaten, Disambiguierung, Maschinelle Erschließung, ORCID, Persistent Identifier

## 1. Einleitung

Im Jahr 2022 importierte die Deutsche Nationalbibliothek (DNB) mehr als 1,6 Millionen Netzpublikationen aus verschiedenen Quellen, beispielsweise von Verlagen und institutionellen Repositorien.<sup>1</sup> Im Gegensatz zu physisch ins Haus kommenden Publikationen können die mitgelieferten Metadaten für Netzpublikationen aufgrund der großen Menge nicht intellektuell bearbeitet werden, sondern werden durch maschinelle Verfahren angereichert.

Die Disambiguierung von Personen in Bibliothekskatalogen mittels Normdaten der Gemeinsamen Normdatei (GND)<sup>2</sup> erfolgt in der Regel intellektuell durch Erschließende. Wenn jedoch immer knapper werdende Ressourcen einer stetig wachsenden Zahl von Veröffentlichungen gegenüberstehen, ist es sinnvoll, bereits andernorts bestehende Zuordnungen maschinell in die eigene Datenbank zu übernehmen. Die DNB hat bereits seit 2011 maschinelle Verfahren im Einsatz, die Titeldaten miteinander verknüpfen und Erschließungsdaten austauschen.<sup>3</sup> Seit 2019 arbeitet die DNB nun verstärkt an weiteren maschinellen Verfahren speziell für Personen.<sup>4</sup>

Dieser Artikel berichtet über die Anreicherung von Daten mit Standardnummern und die daraus erzeugten GND-Verknüpfungen sowie die Übernahme von GND-Nummern aus Culturegraph-Bündeln. Außerdem bietet er einen Ausblick auf die Verfahren, die noch in der Entwicklung sind.

## 2. Verknüpfung mittels Standardnummern

### 2.1 Standardnummern

Zusammen mit Personennamen eignen sich Standardnummern sehr gut für maschinell unterstützte Verknüpfungen von Normdaten in Titeldaten. In der GND werden Standardnummern gespeichert, die als eindeutige Identifikationsmerkmale dienen und die jeweilige Entität in einem bestimmten System repräsentieren.<sup>5</sup> So, wie die GND-Nummern in Normdaten anderer Systeme referenziert werden und damit über Linked Data vernetzt sind, werden in Personendatensätzen der GND zum

---

1 Für weitere Informationen und aktuelle Zahlen siehe den Jahresbericht der DNB unter <https://jahresbericht.dnb.de>, Stand: 31.01.2024.

2 <https://gnd.network>, Stand: 31.01.2024.

3 Beyer, Christian; Trunk, Daniela: Automatische Verfahren für die Formalerschließung im Projekt PETRUS, in: Dialog mit Bibliotheken 23 (2), 2011, S. 5-10. Online: <https://nbn-resolving.org/urn:nbn:de:101-2012030831>. Diebel, Cornelia: Netzpublikationen. Sammlung, Archivierung und Bereitstellung in der Deutschen Nationalbibliothek, in: Dialog mit Bibliotheken 27 (1), 2015, S. 24-30. Online: <https://nbn-resolving.org/urn:nbn:de:101-2015100136>.

4 Bisher konzentrieren sich die Verfahren auf Personennormdaten. Im Rahmen des Projekts PID Network Deutschland wird über Standardnummern und Verknüpfungen weiterer Entitätentypen diskutiert, <https://www.pid-network.de>, Stand: 31.01.2024.

5 Voraussetzung für die Erfassung ist der Eintrag des Standardnummernsystems in die Liste „Standard Identifier Source Codes“ der Library of Congress, <https://www.loc.gov/standards/sourcelist/standard-identifier.html>, Stand: 31.01.2024.

Beispiel die Identifikatoren von ISNI<sup>6</sup>, ORCID<sup>7</sup>, Wikidata<sup>8</sup>, GEPRI Historisch<sup>9</sup> und LCCN<sup>10</sup> gespeichert. Derzeit kommen ORCID iDs (siehe 2.2) und wenige ISNI-IDs<sup>11</sup> auch mittels maschinell unterstützter Verfahren in die GND-Datensätze.

Normdatensysteme werden nicht nur untereinander verknüpft. In der DNB können personenbezogene Standardnummern seit 2017 auch in den Personenfeldern der Titeldaten gespeichert werden. Dort werden sie nicht manuell erfasst, sondern Verlage und Abliefernde von Netzpublikationen speichern die Standardnummern (derzeit ORCID iDs oder ISNI-IDs) in den Personenfeldern und geben diese bei Meldung beziehungsweise Ablieferung an die DNB weiter. Seit 2019 werden Titeldaten auch durch weitere maschinell unterstützte Verfahren mit ORCID iDs angereichert (siehe 2.2 und 2.3). In den Titeldatensätzen der Deutschen Nationalbibliografie waren Ende 2023 etwa 2 Millionen Standardnummern in den Personenfeldern enthalten.

Standardnummern in den Titeldaten sind die Voraussetzung für automatisierte Verknüpfungen zu GND-Personendatensätzen, die in einer täglichen Routine abgeglichen und gegebenenfalls verknüpft werden (siehe 2.3).

## 2.2 ORCID

Im weltweiten wissenschaftlichen Kontext hat sich ORCID zum Standard entwickelt. Seit 2012 können wissenschaftlich Arbeitende einen ORCID-Record erstellen, der Affiliationen, Ausbildungen und vor allem eigene Publikationen zeigt. Die Daten im ORCID-Record können beispielsweise mit einer Hochschulbibliografie und anderen Publikationsservern synchronisiert werden. Das spart nicht nur den Forschenden Zeit, weil sie neue Publikationen immer nur in einem System ergänzen müssen, sondern ihre Angaben geben standardisiert und maschinenlesbar Auskunft über Laufbahn, Themen und Publikationen. Auch wenn Metadaten aus Datenbanken in den ORCID-Record übernommen werden oder sogar von diesen Organisationen verändert werden können, liegt die Kontrolle über die eigenen Daten und alle Vorgänge bei den Forschenden selbst; sie können jederzeit ändern, welche Information unter welchen Bedingungen sichtbar, abruf- und änderbar ist.<sup>12</sup> Mit Stand August 2023 wurden weltweit bereits mehr als 18 Millionen ORCID iDs vergeben, mehr als die Hälfte werden regelmäßig aktualisiert.

6 International Standard Name Identifier, <https://isni.org/page/what-is-isni>, Stand: 31.01.2024.

7 <https://info.orcid.org/what-is-orcid>, Stand: 31.01.2024.

8 <https://www.wikidata.org>, Stand: 31.01.2024.

9 Wissenschaftler und Wissenschaftlerinnen, die zwischen 1920 und 1945 Projektanträge bei den Vorgängerorganisationen der Deutschen Forschungsgemeinschaft (DFG) eingereicht haben, <https://gepris-historisch.dfg.de/editorial/about>, Stand: 31.01.2024.

10 Library of Congress Control Number, <https://id.loc.gov/authorities/names.html>, Stand: 31.01.2024.

11 Im Projekt GND4P (GND für Verlage) sind Einspielungen von ISNI-IDs über die MVB geplant, <https://mvb-online.de/ueber-uns/unternehmen> und [https://gnd.network/Webs/gnd/DE/Projekte/projekte\\_node.html](https://gnd.network/Webs/gnd/DE/Projekte/projekte_node.html), Stand: 31.01.2024.

12 Dreyer, Britta; Hagemann-Wilholt, Stephanie; Vierkant, Paul u. a.: Die Rolle der ORCID iD in der Wissenschaftskommunikation. Der Beitrag des ORCID-Deutschland-Konsortiums und das ORCID-DE-Projekt, in: *ABI Technik* 39 (2), 2019, S. 112-121. Online: <https://doi.org/10.1515/abitech-2019-2004>. Schrader, Antonia C.; Pampel, Heinz; Vierkant, Paul u. a.: Die ORCID iD. Der persönliche Identifier in der Wissenschaft, *Handbuch Qualität in Studium, Lehre und Forschung*, 77: C3.32, 2021. Online: <https://doi.org/10.48440/os.helmholtz.032>.

In zwei DFG-geförderten Projekten<sup>13</sup> wurde 2016 bis 2022 die ORCID-Etablierung an deutschen wissenschaftlichen Organisationen gefördert. Seit 2016 können ORCID iDs in das für Standardnummern vorgesehene Feld in GND-Datensätzen eingetragen werden. Für die GND und die Deutsche Nationalbibliografie wurden mehrere Verfahren entwickelt, die die intellektuelle Erfassung von ORCID iDs und Verknüpfungen zwischen GND-Datensätzen und Titeldaten maschinell unterstützen. Durch die im Folgenden beschriebenen Verfahren war es möglich, bereits mehr als 237.000 GND-Datensätze mit ORCID-Records zu vernetzen. Einmal jährlich wird der ORCID-Dump<sup>14</sup> mit Datensätzen der GND und von Culturegraph (siehe 3.1) abgeglichen, um über Affiliationen und Titeldatenschlüssel weitere ORCID iDs zu speichern (siehe Abbildung 2).<sup>15</sup> Bisher konnten auf diesem Weg mehr als 110.000 ORCID iDs ermittelt werden. Einspielungen nach Abgleichen führen zu sprunghaften Anstiegen der ORCID iDs in der GND, während intellektuelle Erfassungen und tägliche Auswertungen der maschinellen Verfahren als stetiges Wachstum zu erkennen sind (siehe Abbildung 1).<sup>16</sup>

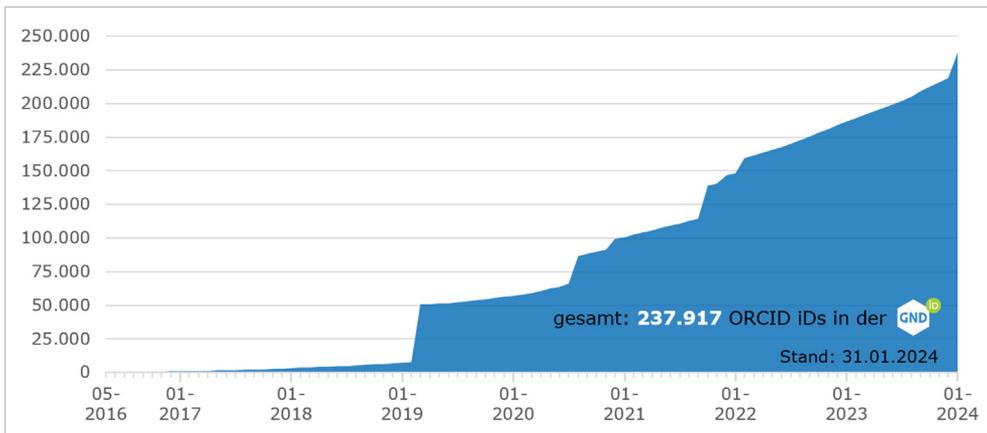


Abb. 1: Wachstum der ORCID iDs in der GND

Claiming<sup>17</sup>-Services erlauben es ORCID-Nutzenden, die Metadaten einer Publikation mit wenigen Klicks aus einer Datenbank zu übernehmen. Diese durch Urheber\*innen verifizierte Zuordnung

- 13 ORCID DE und ORCID DE 2: Förderung von ORCID in Deutschland, <https://www.orcid-de.org/ueber-orcid-de/orcid-de-projekt>, Stand: 31.01.2024. Pampel, Heinz; Schrader, Antonia C.; Vierkant, Paul u. a.: Stand und Perspektive von ORCID in Deutschland, in *Bibliothek Forschung und Praxis*, 2024, im Erscheinen.
- 14 ORCID stellt im Oktober jeden Jahres einen Abzug der Metadaten aller ORCID-Records zur Verfügung, die von den ORCID-Nutzer\*innen als „öffentlich sichtbar“ gekennzeichnet sind, <https://support.orcid.org/hc/en-us/articles/360006897394>, Stand: 31.01.2024.
- 15 Glagla-Dietz, Stephanie; Habermann, Nicole: Standardnummern für Personen. Qualitätsverbesserung durch das Zusammenspiel intellektueller und maschineller Formalerschließung, in: *Dialog mit Bibliotheken* 32 (2), 2020, S. 20-25. Online: <https://nbn-resolving.org/urn:nbn:de:101-2020062250>.
- 16 Das monatliche Wachstum der ORCID iDs in der GND und andere Kennzahlen mit Bezug zu ORCID können im ORCID DE Monitor <https://orcid-monitor.ub.uni-bielefeld.de/auswertung5.php> verfolgt werden. Die einzelnen GND-Personendatensätze können im GND Explorer <https://explore.gnd.network> nach diesem Identifikator gefiltert werden, Stand: 31.01.2024.
- 17 Claimen bedeutet, dass eine Person ihre Urheberschaft der Publikation selbst mitteilt, die Zuordnung also nicht durch eine andere Person oder einen anderen maschinellen Prozess erfolgt. Im ORCID-Record werden die

ist auch für Titeldaten nützlich: In den beiden DFG-Projekten wurden drei Claiming-Services eingerichtet, die zur Anreicherung der DNB-Daten mit ORCID iDs ausgewertet werden. Forschende ergänzen über ihren ORCID-Record damit täglich Verknüpfungen und ORCID iDs in der Deutschen Nationalbibliografie und der GND:

- Bereits seit Mai 2017 können ORCID-Nutzende die Titel-Metadaten von BASE, der größten Datenbank für Open-Access-Publikationen, claimen.<sup>18</sup> Den ersten Claiming-Service für die GND hat der Projektpartner UB Bielefeld im November 2018 im Nachgang an das Claiming in BASE eingerichtet. Ein GND-Datensatz kann direkt recherchiert und ausgewählt werden. Mehr als 28.000 ORCID-Nutzende haben seither ihren GND-Datensatz geclaimt.
- Im Juli 2019 folgte der Claiming-Service für etwa 20 Millionen Titeldaten der Deutschen Nationalbibliografie<sup>19</sup>. Geclaimt wird immer eine Personenangabe in einem Titeldatensatz, über die jeweils die GND-Nummer ermittelt wird.<sup>20</sup> Mehr als 19.600 unterschiedliche ORCID-Nutzende haben den Service in den ersten viereinhalb Jahren erfolgreich genutzt, dabei in mehr als 116.000 Titeldaten ihre Urheberschaft sowie sich zu mehr als 10.000 GND-Datensätzen zugeordnet.
- Seit Dezember 2022 ist der Claiming-Service für die Culturegraph-Daten produktiv.<sup>21</sup> Unter den mehr als 100 Millionen claimbaren Titeldaten stehen mehr Zeitschriftenartikel und Veröffentlichungen aus dem nicht-deutschsprachigen Raum zur Verfügung als im ersten Claiming-Service, so dass ihn Forschende ohne Deutschlandbezug stärker nutzen können. Knapp 4.000 unterschiedliche ORCID-Nutzende haben den Service in den ersten dreizehn Monaten erfolgreich genutzt und dabei mehr als 34.000 Titel geclaimt. Die Claiming-Ergebnisse werden 2024 in die GND und in die DNB-Titeldaten sowie in die abendliche Routine übernommen.

Derzeit sind in den 6,1 Millionen Personennormdatensätzen der GND 237.000 ORCID iDs enthalten. Abbildung 2 zeigt die Verteilung der Herkunft durch Claimings, Dump-Abgleiche und manuelle Erfassung.<sup>22</sup> Über die Claiming-Services BASE und Deutsche Nationalbibliografie (DNB) wurden zusammen 14 % aller ORCID iDs hinzugefügt oder bestätigt. Mithilfe maschineller Verfahren wird der jährlich angebotene Dump aller öffentlich zugänglichen ORCID-Records analysiert. Hierzu zählt einerseits der Abgleich der ORCID-Daten mit den Inhalten von GND-Datensätzen über Namen

Claiming-Services „Search&Link“ genannt.

18 ORCID-Claiming in BASE möglich, <https://www.orcid-de.org/support/blogbeitraege/orcid-claiming-in-base-moeglich>, Stand: 31.01.2024.

19 ORCID-Claiming in der Deutschen Nationalbibliografie und der GND möglich, <https://www.orcid-de.org/support/blogbeitraege/orcid-claiming-in-der-deutschen-nationalbibliografie-und-der-gnd-moeglich>, Stand: 31.01.2024. Die Screenshots zeigen den Zugang zu den im ORCID-Record „Search&Link“ genannten Claiming-Services.

20 Ist das Personenfeld bereits mit einem GND-Datensatz verknüpft, kann die GND-Nummer sofort in den ORCID-Record übernommen werden; ist im GND-Datensatz noch keine ORCID iD vorhanden, wird sie dort – bei einer übereinstimmenden Namensform – am Abend maschinell ergänzt. Ist das Personenfeld noch nicht verknüpft, kann der ORCID-Record nicht um die GND-Nummer angereichert werden; am Abend wird die ORCID iD im Personenfeld des Titeldatensatzes ergänzt und mit dem GND-Datensatz verknüpft, sofern dort bereits die ORCID iD eingetragen ist (siehe 2.3). Gibt es noch keinen GND-Datensatz für die claimende Person, wird ein Vorschlagsdatensatz angelegt, der zeitsparend von einer GND-Redaktion in einen GND-Datensatz umgewandelt werden kann (siehe 6.1). Viele Claimende nutzen den Service bereits wiederholt für neue Publikationen.

21 Dritter Claiming-Service unterstützt ab sofort Publizierende und die GND-Community, <https://www.orcid-de.org/support/blogbeitraege/dritter-claiming-service-produktiv>, Stand: 31.01.2024.

22 Ausführlich werden die Verfahren beschrieben in Glagla-Dietz; Habermann: Standardnummern für Personen, 2020.

und Affiliationen (24 %) und andererseits der Abgleich von ORCID-Daten mit den Publikationen in Culturegraph (17 %). Fast die Hälfte der ORCID iDs (45 %) wird intellektuell erfasst oder es werden intellektuell die vorher mittels Dump-Abgleich eingespielten ORCID iDs überprüft.

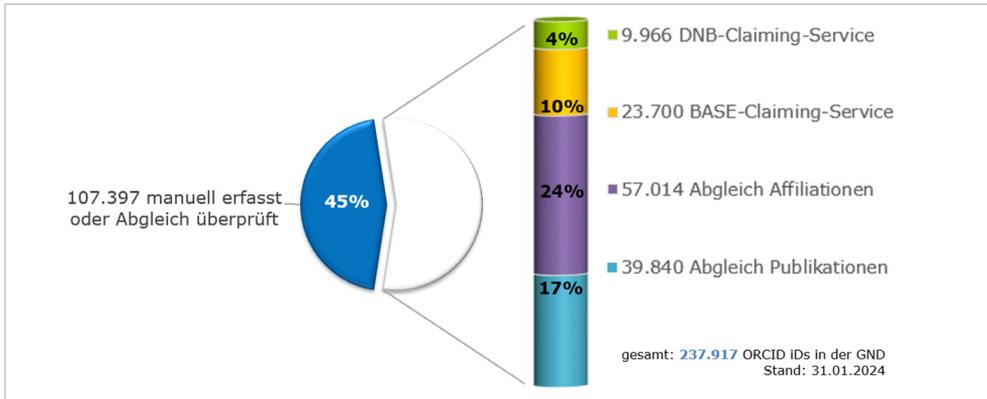


Abb. 2: Herkünfte der ORCID iDs in GND-Personendatensätzen

### 2.3 Verknüpfungslauf über Standardnummern

Mithilfe der in den Titel- und Normdatensätzen vorhandenen Standardnummern können diese maschinell miteinander verknüpft werden. Wenn eine Standardnummer gleichzeitig in einem Titeldatensatz und einem Personendatensatz vorhanden ist und der Name übereinstimmt, wird der Namensstring durch die GND-Nummer ersetzt. Im April 2020 wurde zunächst ein Retro-Lauf über die bestehenden Daten durchgeführt und anschließend das tägliche Verfahren gestartet. Dieses verknüpft täglich im Durchschnitt 210 Personennamen. Seit Start des Verfahrens wurden mehr als 267.000 Verknüpfungen erstellt (siehe Abbildung 3).

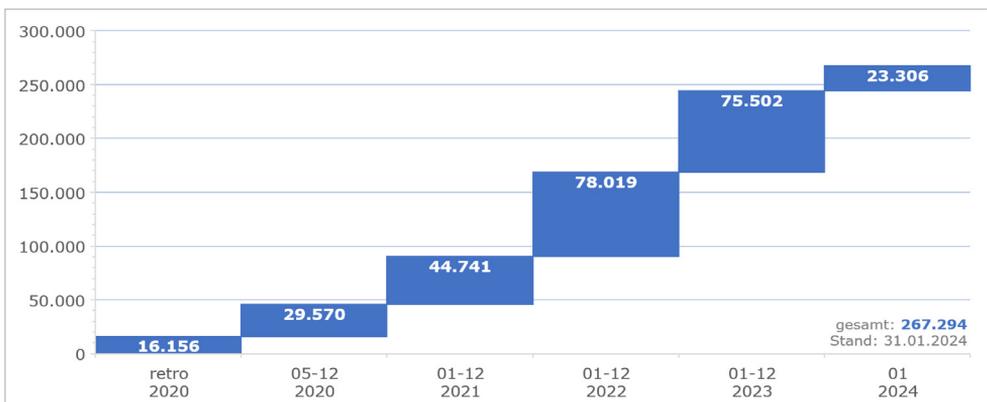


Abb. 3: Anzahl GND-Titel-Verknüpfungen via Standardnummer

Wenn die Namen nicht übereinstimmen oder es mehrere GND-Datensätze mit der gleichen Standardnummer gibt, werden GND-Redakteur\*innen benachrichtigt und können so Fehler bereinigen oder beispielsweise abweichende Namen eintragen. Die maschinellen Prozesse weisen also auch auf Aktualisierungsbedarf in den Daten hin.

### 3. Auswertung der Culturegraph-Daten

#### 3.1 Culturegraph als Aggregationsplattform

Mit Culturegraph bietet die DNB eine Plattform, auf der die bibliographischen Metadaten aller deutschen Bibliotheksverbände sowie der Österreichischen Bibliothekenverbund und Service GmbH (OBVSG) und der DNB aggregiert werden. Über eine Kooperation mit dem Gemeinsamen Verbändeindex (GVI)<sup>23</sup> werden die Daten täglich aktualisiert. Dieser aggregierte Datenbestand umfasst im Januar 2024 mehr als 230 Millionen Datensätze (siehe Abbildung 4).



Abb. 4: Aggregierte Daten in Culturegraph

In diesem Abschnitt werden mehrere Verfahren vorgestellt, die das Ziel haben, mit Hilfe dieses Datenbestandes Personeninformationen in Norm- und Titeldaten anzureichern. Dazu gehört u.a. die Übertragung von personenbezogenen Informationen aus Werk- oder Manifestationsbündeln. Hier werden die Datensätze in Culturegraph zu Bündeln zusammengefasst, die jeweils ein Werk, beziehungsweise eine Manifestation eines Werkes repräsentieren. Dies passiert über Schlüssel, die aus verschiedenen Metadatenelementen zusammengesetzt und dann über eine Breitensuche gebündelt werden. Für jeden Datensatz entstehen meist mehrere Schlüssel.<sup>24</sup>

<sup>23</sup> <https://www.agv-gvi.de/>, Stand: 31.01.2024.

<sup>24</sup> Vorndran, Angela: Hervorholen, was in unseren Daten steckt! Mehrwerte durch Analysen großer Bibliotheksdatenbestände, in: o-bib. Das offene Bibliotheksjournal, 5 (4), 2018, S. 166-180. Online: <https://doi.org/10.5282/o-bib/2018h4s166-180>. Vorndran, Angela; Grund, Stefan: Metadata Sharing – How to Transfer Metadata Information among Work Cluster Members, in: Cataloging & Classification Quarterly, 59 (8), 2021, S. 757-774. Online: <https://doi.org/10.1080/01639374.2021.1989101>.

Die Ergebnisse der Werkbündelung werden auch zur allgemeinen Verwendung öffentlich unter <https://data.dnb.de/culturegraph/> zum Download bereitgestellt und monatlich aktualisiert. Diese Datei enthält alle Werkbündel in einem modifizierten MARCXML-Format, das pro Bündel die Identifikationsnummern (IDN) aller Mitglieder, alle im Bündel enthaltenen beteiligten Akteure gegebenenfalls mit Normdatenverknüpfungen und weiteren Informationen sowie inhaltserschließende Informationen, beispielsweise Klassifikationsnotationen und Schlagwörter, enthält.

### 3.2 Übertragung von Personenverknüpfungen aus Culturegraph-Werkbündeln

#### Abgleich von Personenangaben

In einem Werkbündel<sup>25</sup> sind mehrere Manifestationen eines Werkes enthalten. Die darin verzeichneten namensgleichen Personen werden durch ihre Verbindung zu der Publikation als gleiche Person bestätigt. Ziel des Abgleichs der Personenangaben in einem Werkbündel ist eine Übernahme von GND-Verknüpfungen im Bereich der Personen aus einem Titeldatensatz des Bündels in weitere Titeldatensätze, die diese Informationen noch nicht enthalten. Hierfür erfolgt ein Abgleich der aus den verschiedenen Bündelmitgliedern stammenden Personen, die mit einem Werk verbunden sind.

Aus der im vorherigen Kapitel beschriebenen Bündeldatei werden zu diesem Zweck alle Personennamen aus einem Bündel mit gegebenenfalls vorhandener Verknüpfung zur GND und der IDN des jeweiligen Titeldatensatzes in einer Liste gespeichert. Personen mit GND-Verknüpfung werden gesondert von Personen ohne Verknüpfung behandelt. Zunächst werden die Richtigkeit der GND-Nummer sichergestellt und die Namensformen im Titeldatensatz und GND-Datensatz verglichen. Enthält der GND-Datensatz weitere alternative Namensformen, werden diese ebenfalls in den Abgleich aufgenommen. Nach einer Normalisierung aller Namensformen wird ein Abgleich vorgenommen und für namensgleiche Personen in demselben Werkbündel eine möglicherweise vorhandene GND-Verknüpfung als Vorschlag für einen bestimmten Datensatz ausgegeben.

Werden für namensgleiche Personen verschiedene GND-Nummern in einem Bündel angetroffen, wird keine GND-Verknüpfung vorgeschlagen, sondern dieser Fall in eine Liste für intellektuell zu überprüfende Verknüpfungen eingetragen. Es kann sich hier möglicherweise um einen Verknüpfungsfehler handeln oder auch um einen dubletten GND-Datensatz. Auch bei häufigen Namen ist das Verfahren robust, da die Zugehörigkeit zu einem Werk namensgleiche Personen stark differenziert und Irrtümer bei der Zuordnung zu einem von mehreren namensgleichen GND-Datensätzen häufig durch sich widersprechende GND-Nummern in einem Bündel zutage treten.

In der Entwicklung des Verfahrens, die iterativ in einer Zusammenarbeit von Kolleg\*innen verschiedener Abteilungen der DNB organisiert war, wurde eine umfangreiche intellektuelle Evaluation durchgeführt (n=1199), die eine Quote von 90 % korrekten, 6 % schwierig zu verifizierenden und 4 %

---

25 Die hier verwendete Definition eines Werkes orientiert sich an dem IFLA Library Reference Model, vgl. IFLA Library Reference Model. A Conceptual Model for Bibliographic Information, August 2017, S. 21 ff. Online: <https://www.ifla.org/files/assets/cataloguing/frbr-lrm/ifla-lrm-august-2017.pdf>, Stand: 31.01.2024.

falschen Anreicherungen ergab. Letztere führen sich in der Regel auf fehlerhafte Verknüpfungen in den Ausgangsdaten zurück.

### Eintragen in Titeldatensätze

Sind in dem oben beschriebenen Verfahren geeignete Anreicherungen von Normdatenverknüpfungen ermittelt worden, werden sie für Personen in DNB-Titeldatensätzen eingetragen, die noch keine Verknüpfung aufweisen (siehe Abbildung 5). Vorschläge, die sich auf Titeldatensätze aus den Bibliotheksverbänden beziehen, wurden an diese zunächst sporadisch, ab September 2023 monatlich zur eigenen Nutzung weitergeleitet.

Bei der Anreicherung von DNB-Titeldatensätzen wird zunächst eine weitere Prüfung vorangestellt. Ein gegebenenfalls im Personennormdatensatz vorhandenes Geburtsjahr wird mit dem Publikationsjahr des Titeldatensatzes abgeglichen. Liegt das Geburtsdatum mindestens 15 Jahre vor Publikationsdatum, ist die Prüfung erfolgreich. Dann wird in dem durch die IDN identifizierten Datensatz im Katalogisierungssystem die Namensnennung durch eine Verknüpfung zur Normdatei ersetzt.

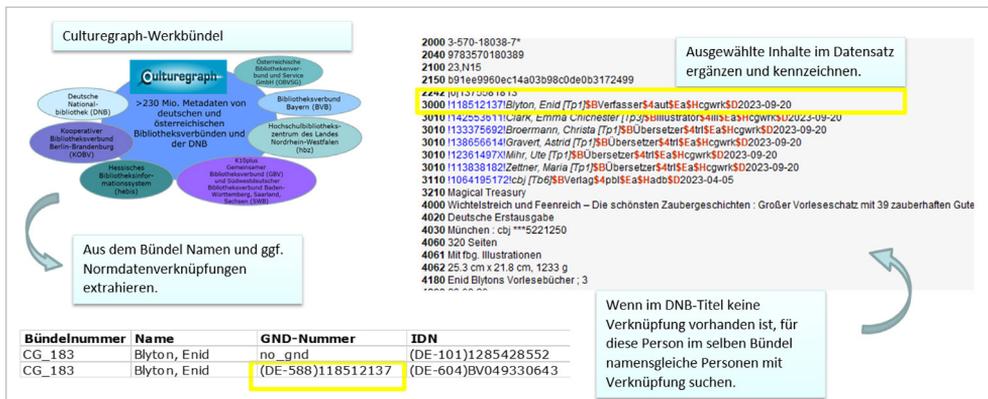


Abb. 5: Workflow des Abgleichs an einem Beispiel illustriert

### Täglicher Ablauf und Retroläufe

Nachdem das Verfahren der Übertragung von Personenverknüpfungen aus Werkbündeln seit 2021 in mehreren Läufen retrospektiv durchgeführt wurde, findet jetzt auch ein tagesaktueller Abgleich statt. Dieser tägliche Lauf beginnt mit dem Update der Culturegraph-Datenbasis (siehe 3.1). Es werden anschließend für alle neu hinzugekommenen Datensätze Schlüssel erzeugt und mit den bereits vorhandenen Schlüsseln der vorherigen Läufe neu zu Bündeln zusammengefasst. Für diese Bündel werden die neu hinzugekommenen zu übertragenden Verknüpfungen ermittelt und in einer Datei ausgegeben. Diese Datei wird im zentralen Bibliothekssystem weiterverarbeitet und die Datensätze werden um Verknüpfungen ergänzt.

Durch die blockweise Retrobearbeitung konnten in einem ersten Lauf 2021 mehr als 1,3 Millionen Verknüpfungen in mehr als 1,2 Millionen Titeldatensätzen ergänzt werden. Ein weiterer Lauf im Jahr 2022 ergab etwa 300.000 weitere zu ergänzende Verknüpfungen.

Das tägliche Ermitteln von neuen Verknüpfungen startete im Dezember 2022 und erbringt seitdem gleichmäßige Zuwächse an Normdatenverknüpfungen mit durchschnittlich knapp 600 neuen Verknüpfungen pro Tag (siehe Abbildung 6).

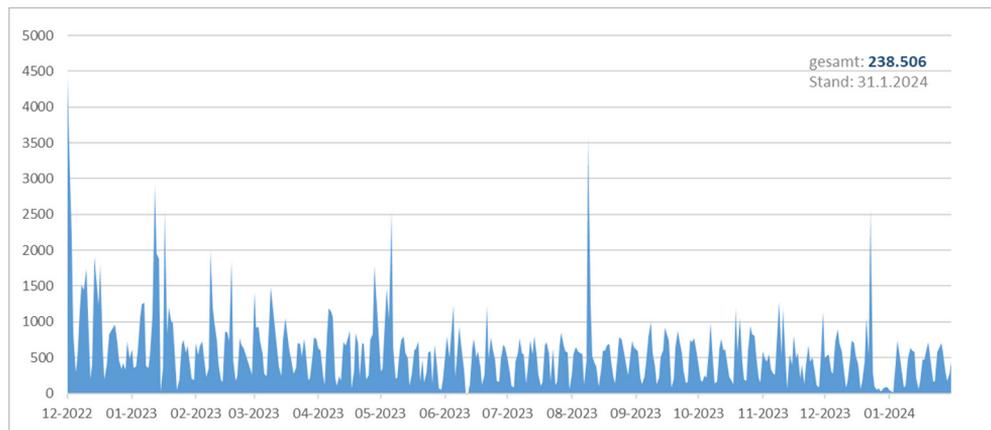


Abb. 6: Täglich automatisiert eingespielte Verknüpfungen via Culturegraph seit 01.12.2022

### 3.3 Übertragung von Personenangaben aus Culturegraph-Manifestationsbündeln

Die Werkbündelung bündelt per Definition verschiedene Auflagen, Ausgaben und auch Übersetzungen eines Werkes. Für manche Einsatzzwecke ist allerdings eine engere Bündelung von in sich homogeneren Publikationen hilfreicher. Aus diesem Grund wurde eine Manifestationsbündelung entwickelt, die nur gleiche Auflagen oder Ausgaben in jeweils eigenen Bündeln zusammenfassen soll. Die Manifestationsbündelung erfolgt analog zum Verfahren der Werkbündelung, allerdings werden die für jede Publikation erstellten Schlüssel um manifestationsbezogene Elemente wie beispielsweise das Erscheinungsjahr ergänzt.

Die Manifestationsbündelung erlaubt, auch Personenangaben zu übernehmen, die nicht zusätzlich durch das gemeinsame Vorkommen eines gleichlautenden Namens verifiziert werden, weil hier nur gleiche Manifestationen desselben Verlags im selben Erscheinungsjahr gebündelt werden und Autor\*innen und beteiligte Personen übereinstimmen. Wie in Abbildung 7 ersichtlich, fehlen in manchen Datensätzen Personenangaben vollständig, so dass die Möglichkeit, solche aus anderen Bündelmitgliedern zu übernehmen, eine deutliche Verbesserung des Datensatzes bedeuten würde. In einem ersten Schritt wird zurzeit die Übernahme von Personennamen wie auch Personenverknüpfungen und Beziehungskennzeichnungen für Titel ohne Personenangaben evaluiert.

Durch das Fehlen eines Personennamens im zu ergänzenden Titeldatensatz ergeben sich zusätzliche Fragen, die vor einer Anreicherung entschieden werden müssen. Wenn in dem Manifestationsbündel mehrere Vorkommen eines zu ergänzenden Namens auftreten, kommen zum Teil Unterschiede in der Erfassung bezüglich des gewählten Feldes und der Beziehungskennzeichnung zum Tragen. Beispielsweise ist ein Name in einem Datensatz im Feld des ersten geistigen Schöpfers (PICA3 3000/MARC 100) eingetragen und in einem anderen Datensatz im Feld für weitere geistige Schöpfer und sonstige Mitwirkende (PICA3 3010/MARC 700). Auch die Beziehungskennzeichnungen, die die Beziehung der Person zu der Publikation beschreibt (beispielsweise „aut“ für Autor, „edt“ für Herausgeber), weicht in verschiedenen Titeldatensätzen häufig ab. Für einen Vorschlag der Anreicherung wird in solchen Fällen der jeweils häufiger auftretende Fall übernommen.

(DE-101)450337596 EINZEL-ANSICHT ▶		Überblick	
<b>Überblick</b>		Hauptsachtitel	Beiträge zur Entwicklung der Bevölkerung in der Sowjetunion und in Mitteldeutschland
Hauptsachtitel	Beiträge zur Entwicklung der Bevölkerung in der Sowjetunion und in Mitteldeutschland	Person	<a href="#">GND: Härmisen, Hans</a> <a href="#">11854604X</a>
Person	-	Körperschaft	-
Körperschaft	-	Umfang	VI, 132 S.
Umfang	VI, 132 S., 4 Bl., 1 Faltkt.	Erscheinungsjahr	1957
Erscheinungsjahr	1957	Material	text
Ländercode	XA-DE	Erscheinungsort	Hamburg
Material	text	Verlag	Akad. f. Staatsmedizin
Erscheinungsort	Hamburg	Standard-Identifizier	<a href="#">DCLC: 1072025960</a>
Verlag	Akademie f. Staatsmedizin in Hamburg	Verlags-Identifizier	-
		Klassifikation	<a href="#">RVK: RQ 10585</a>

Abb. 7: Möglichkeit zur Übernahme von Personenangaben aus Manifestationsbündeln

## 4. Übersicht der automatisiert erstellten Anreicherungen in Titeldaten

Wie in den vorangehenden Kapiteln deutlich wurde, werden über verschiedene Quellen und Verfahren automatisiert Anreicherungen vorgenommen, die die intellektuelle Bearbeitung von Datensätzen ergänzen. Hier soll ein Überblick über die Normdatenverknüpfungen von Personen in Titeldaten gegeben und ein Augenmerk auf die verschiedenen Herkünfte dieser Information gelegt werden.

In den DNB-Titeldaten finden sich insgesamt mehr als 11 Millionen Verknüpfungen zu Personen-normdaten. Alle maschinell erstellten oder angereicherten sowie seit 2020 auch alle intellektuell bearbeiteten Personenangaben werden gekennzeichnet. Anhand der Kennzeichnung ist ersichtlich, durch welchen Prozess und zu welchem Zeitpunkt die letzte Änderung vorgenommen wurde. Der größte Teil der Verknüpfungen stammt aus dem historisch gewachsenen Bestand und verfügt über keine Kennzeichnung (72 %). Dieser Teil wurde überwiegend intellektuell erstellt oder durch einen Abgleich paralleler Print- und Onlineausgaben einer Publikation übernommen. Die gekennzeichneten Verknüpfungen enthalten einen Anteil intellektuell erstellter Verknüpfungen von sechs Prozent.

Weitere Kennzeichnungen beziehen sich auf verschiedene automatisierte Verfahren, von denen der Culturegraph-Werkabgleich den größten Teil stellt (17 %).

Betrachtet man nur die in Abbildung 8 dargestellten, aus automatischen Verfahren eingespielten Verknüpfungen, zeigt sich, dass die Übertragung aus Culturegraph-Werkbündeln (siehe 3.2) mit 77 % den größten Teil einnimmt. Häufig sind außerdem Verknüpfungen aus dem Abgleich von Standardnummern (siehe 2.3) mit acht Prozent und über den Import von Netzpublikationen erfasste Verknüpfungen (6 %).

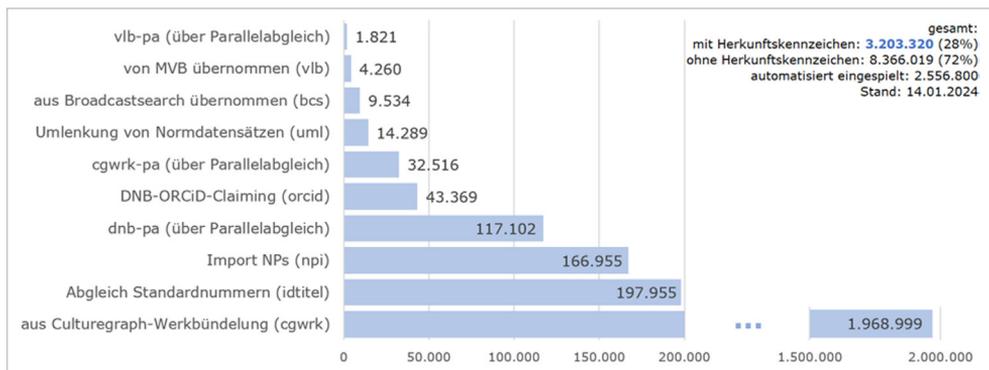


Abb. 8: Automatisiert eingespielte Verknüpfungen zu Personennormdatensätzen in Titeldatensätzen (ab einem Vorkommen von 1.000)

## 5. Vorschläge für neue GND-Personendatensätze

Um die maschinellen Prozesse möglichst gewinnbringend einsetzen zu können, werden Normdatensätze für möglichst viele Personen benötigt. Deshalb wurde in der DNB 2022 die maschinelle Generierung von Vorschlägen für neue GND-Datensätze gestartet. Dafür wurde das bereits bestehende Verfahren der Auswertung der DNB-ORCID-Claimings ausgeweitet.

Wenn im Titeldatensatz keine GND-Verknüpfung vorhanden ist und es keinen Normdatensatz mit diesem Personennamen gibt<sup>26</sup>, wird ein Vorschlagsdatensatz angelegt. Der Vorschlagsdatensatz enthält Informationen aus dem ORCID-Record der claimenden Person und dem geclaimten Titeldatensatz. Hierzu zählen der bevorzugte und gegebenenfalls abweichende Name, die ORCID iD, Affiliationen, Wirkungsdaten und Titelangaben. Claimt eine Person mehrere ihrer Publikationen, werden diese im bereits bestehenden Vorschlagsdatensatz ergänzt. Die Anzahl der in einem Vorschlagsdatensatz enthaltenen Titelangaben ist ein wichtiger Indikator für die Priorisierung bei der Aufarbeitung eines Vorschlagsdatensatzes zu einem GND-Datensatz.

<sup>26</sup> Für die Hälfte der rund 20.000 ORCID-Nutzer\*innen, die bisher den DNB-Claming-Service genutzt haben, gibt es noch keinen GND-Datensatz, beim GND Network Claimer ist es weniger als ein Viertel (siehe 2.2).

In unregelmäßigen Abständen läuft ein weiteres Verfahren: Für alle ORCID iDs, die in mindestens fünf Titeldatensätzen der Deutschen Nationalbibliografie vorkommen, werden neue Vorschlagsdatensätze angelegt, sofern es noch keinen GND-Datensatz mit diesem Namen gibt. Ist bereits ein Vorschlagsdatensatz vorhanden, werden die Titel hinzugefügt.

Insgesamt wurden bisher knapp 40.000 Vorschlagsdatensätze generiert. Sie sind nicht Bestandteil der GND, sondern werden intellektuell zu GND-Datensätzen aufgearbeitet. Dies geschieht durch die gezielte Bearbeitung von nach Anzahl der Quelltitel sortierten Listen und wenn ein GND-Datensatz für die intellektuelle Erschließung benötigt wird. Dafür steht maschinelle Unterstützung zur Verfügung, die möglichst viele Felder des Datensatzes vorab bearbeitet oder ergänzt.

Ein auf diesem Wege neu erstellter GND-Datensatz löst in der folgenden Nacht die Suche nach Titeldatensätzen mit der gespeicherten ORCID iD aus (siehe 2.3). Alle entsprechenden Personenangaben werden automatisch mit dem neuen GND-Datensatz verknüpft.

## 6. Verfahren in Entwicklung

### 6.1 Hintergrund

Ergänzend zu den in den vorangegangenen Kapiteln beschriebenen Verfahren werden weitere Möglichkeiten evaluiert, die Identität einer Person in einem Metadatensatz zu ermitteln und möglichst mit einer eindeutigen Verknüpfung zur GND zu versehen. Hierzu sollen Titeldatensätze im DNB- und zukünftig auch Culturegraph-Bestand nach Zugehörigkeit zu Personen gebündelt werden, indem verschiedene Elemente der Metadatensätze zum Abgleich herangezogen werden. Gibt es dann innerhalb eines solchen Bündels einen Titeldatensatz mit einer GND-Verknüpfung für diese Person, kann diese für alle Vorkommen der Person auch in die anderen Titeldatensätze übernommen werden. Verschiedene Bündelungsmöglichkeiten werden in den folgenden Unterkapiteln vorgestellt.

### 6.2 Auswertung der Beziehungen von Ko-Autor\*innen

Unter den Netzpublikationen der DNB finden sich zahlreiche wissenschaftliche Aufsätze. Gerade in dieser Literaturgruppe ist kollaboratives Arbeiten weit verbreitet. Artikel werden von mehreren Personen verfasst, und diese kooperieren häufig für mehrere Veröffentlichungen miteinander. Daher kann die identische Kombination von zwei oder mehr Autorennamen in verschiedenen Artikeln darauf hinweisen, dass es sich um dieselben Personen handelt. Auf diese Weise können Bündel von Titeldaten dieser Personen gebildet und vorhandene GND-Verknüpfungen für alle Mitglieder des Bündels übernommen werden (siehe Abbildung 9).

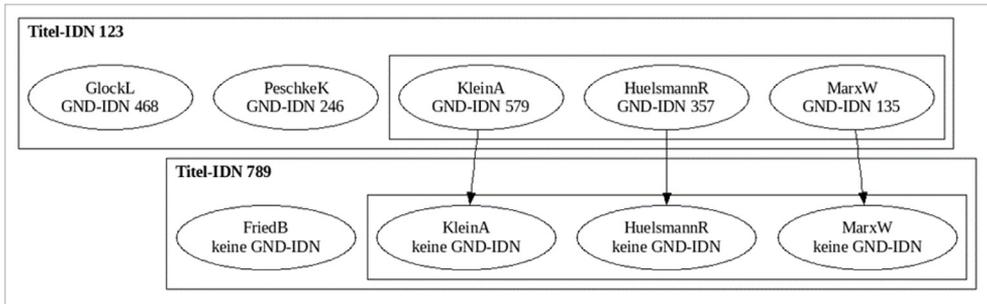


Abb. 9: Beispiel für Ko-Autor\*innen-Beziehungen in zwei Publikationen

Erste Tests bezogen auf den DNB-Bestand ergaben, dass durch die Auswertung von Ko-Autorenschaften circa 140.000 Titeldatensätze um etwa 280.000 Personenverknüpfungen zur GND angereichert werden könnten. Bei Einbezug sämtlicher Culturegraph-Daten wäre die Zahl entsprechend höher.

Weitere Tests und Evaluierungen<sup>27</sup> sind noch notwendig, um beispielsweise entscheiden zu können

- Wie viele übereinstimmende Namen müssen in der Kombination enthalten sein (zwei, drei oder mehr), um Fehlverknüpfungen möglichst auszuschließen? (Je mehr übereinstimmende Autor\*innen, umso zuverlässiger)
- Kann das Verfahren auch für häufig vorkommende Namen (Müller, Schmidt, ...) angewandt werden? Muss für sie die Zahl der übereinstimmenden Namen in der Kombination höher sein als für nicht so häufig vorkommende Namen?

### 6.3 Auswertung von weiteren Titeldatenelementen zur Bündelung von Titeln einer Person

Neben Ko-Autor\*innen können auch weitere Elemente eines Titeldatensatzes herangezogen werden, um Titel mit identischen Personenangaben zu bündeln, beispielsweise:

- DDC-Sachgruppe
- Zeitschrift, in der ein Artikel enthalten ist
- Verlag, in dem die Publikation erschienen ist
- Basis-URL, die auf die Publikation verweist
- Kontrolliertes Vokabular, wie Schlagwörter

Je nach Testergebnissen muss dann festgelegt werden, wie viele dieser Elemente neben dem Personennamen übereinstimmen müssen, um Titel einer Person gemeinsam bündeln zu können.

<sup>27</sup> Ausführlich werden die zu beachtenden Herausforderungen und Einschränkungen der Nutzung von Ko-Autorenschaften in einer B.A.-Arbeit von Swantje Wiechmann beschrieben: Wiechmann, Swantje: Examining concepts of author disambiguation – co-authorship as a disambiguation feature in EconBiz, Bachelorarbeit Hochschule für Technik, Wirtschaft und Kultur Leipzig, 2022. Online: <https://nbn-resolving.org/urn:nbn:de:bsz:l189-qucosa2-788413>.

	title_idn_ziel	name	title_idn_quelle	gnd_idn	sg_ziel	sg_quelle	vlg_ziel	vlg_quelle	sg_equal	vlg_equal
6693	103192924X	Wenzel, Gerd	1007865237 (DE-588)134674944	360	360	beck	beck	True	True	
6694	103192924X	Wenzel, Gerd	106711971X (DE-588)134674944	360	360	beck	beck	True	True	
6708	1044000090	Müller, Wolfgang H.	1043995161 (DE-588)13320040X	620	620	hansercarl	hansercarl	True	True	
6710	1044000090	Müller, Wolfgang H.	1078822727 (DE-588)13320040X	620	620	hansercarl	hansercarl	True	True	
7042	1118889711	Murray, Paul	1098154983 (DE-588)130399086	820	820	antjekunstmann	antjekunstmann	True	True	

Abb. 10: Ausgewählte Metadatenelemente und Abgleichsergebnisse

Im abgebildeten Beispiel (Abbildung 10) ist im Titel mit der IDN 103192924X eine Person namens „Wenzel, Gerd“ genannt, es gibt dort aber keine GND-Verknüpfung für diese Person. Durch den beschriebenen Abgleich weiterer Metadatenelemente kann dieser Titel jetzt mit zwei anderen Titeln gebündelt werden, den Titel-IDNs 1007865237 und 106711971X, da für sie derselbe Verlag und dieselbe DDC-Sachgruppe erfasst wurden. In den beiden anderen Titeln ist für die Person eine GND-Nummer vorhanden, die übernommen werden kann.

#### 6.4 Auswertung von Literaturreferenzen einer Publikation

Als weitere Bündelungsmöglichkeit können die Referenzen eines Artikels herangezogen werden. Es ist davon auszugehen, dass ein zitierter Artikel, der den gleichen Namen als Autor\*in angibt wie der Artikel, in dem die Referenz enthalten ist, von derselben Person verfasst wurde und es sich um ein Selbstzitat handelt. Durch eine Netzwerkanalyse von Literaturverzeichnissen können so sämtliche zitierten Publikationen einer Person in einem Bündel zusammengebracht werden. Für diese Analyse bietet sich der Datenbestand von Crossref<sup>28</sup> an, da er neben den zitierten Publikationen der erfassten Artikel auch persistente Publikations-Identifikatoren wie Digital Object Identifier (DOI) enthält, die die Netzwerkanalyse stark vereinfachen. Anhand der einer Person zugeordneten Publikationen und deren DOIs können dann die im DNB-Datenbestand vorliegenden Publikationen gefunden und, wenn möglich, Personenverknüpfungen von anderen Bündelmitgliedern übernommen werden.

### 7. Fazit

Die Arbeit mit Personenangaben in Titeldaten hat sich in der DNB in den letzten Jahren stark verändert. Durch die maschinelle Anreicherung mit Standardnummern und GND-Verknüpfungen werden deutlich mehr Personenangaben in Titeldaten mit GND-Datensätzen verknüpft als intellektuell möglich wäre. Die maschinellen Prozesse, die vorrangig für die Versorgung von Online-Publikationen eingerichtet wurden, reichern aber auch die Titeldaten von körperlichen Medienwerken mit GND-Verknüpfungen an. Daher finden Katalogisierer\*innen oft schon Verknüpfungen vor, die sie nur noch überprüfen müssen.

Es hat sich gezeigt, dass Investitionen in die Entwicklung von neuen maschinellen Verfahren zur Disambiguierung von Personen lohnenswert und wegen der großen Anzahl von Netzpublikationen auch unerlässlich sind. Neben der Übernahme von bereits in anderen Datenbanken bestehenden Zuordnungen wird die Entwicklung von Verfahren, die Titeldaten aufgrund verschiedener Metadatenelemente

28 <https://www.crossref.org/about>, Stand: 31.01.2024.

nach ihrer Zugehörigkeit zu Personen bündeln, derzeit vorrangig betrachtet. Daraus sollen weitere neue Verknüpfungen, Hilfestellungen bei der Bereinigung von Fehlern und Erkenntnisse entstehen, die zu Ideen für weitere Verfahren führen.

## Literaturverzeichnis

- Beyer, Christian; Trunk, Daniela: Automatische Verfahren für die Formalerschließung im Projekt PETRUS, in: Dialog mit Bibliotheken 23 (2), 2011, S. 5-10. Online: <https://nbn-resolving.org/urn:nbn:de:101-2012030831>.
- Diebel, Cornelia: Netzpublikationen. Sammlung, Archivierung und Bereitstellung in der Deutschen Nationalbibliothek, in: Dialog mit Bibliotheken 27 (1), 2015, S. 24-30. Online: <https://nbn-resolving.org/urn:nbn:de:101-2015100136>.
- Dreyer, Britta; Hagemann-Wilholt, Stephanie; Vierkant, Paul u. a.: Die Rolle der ORCID iD in der Wissenschaftskommunikation. Der Beitrag des ORCID-Deutschland-Konsortiums und das ORCID-DE-Projekt, in: ABI Technik 39 (2), 2019, S. 112-121. Online: <https://doi.org/10.1515/abitech-2019-2004>.
- Glagla-Dietz, Stephanie; Habermann, Nicole: Standardnummern für Personen. Qualitätsverbesserung durch das Zusammenspiel intellektueller und maschineller Formalerschließung, in: Dialog mit Bibliotheken 32 (2), 2020, S. 20-25. Online: <https://nbn-resolving.org/urn:nbn:de:101-2020062250>.
- IFLA Library Reference Model. A Conceptual Model for Bibliographic Information, August 2017, S. 21 ff. Online: <https://www.ifla.org/files/assets/cataloguing/frbr-irm/ifla-irm-august-2017.pdf>, Stand: 31.01.2024.
- Pampel, Heinz; Schrader, Antonia C.; Vierkant, Paul u. a.: Stand und Perspektive von ORCID in Deutschland, in Bibliothek Forschung und Praxis, 2024, im Erscheinen.
- Schrader, Antonia C.; Pampel, Heinz; Vierkant, Paul u. a.: Die ORCID iD. Der persönliche Identifier in der Wissenschaft, Handbuch Qualität in Studium, Lehre und Forschung, 77: C3.32, 2021. Online: <https://doi.org/10.48440/os.helmholtz.032>.
- Vorndran, Angela: Hervorholen, was in unseren Daten steckt! Mehrwerte durch Analysen großer Bibliotheksdatenbestände, in: o-bib. Das offene Bibliotheksjournal, 5 (4), 2018, S. 166-180. Online: <https://doi.org/10.5282/o-bib/2018h4s166-180>.
- Vorndran, Angela; Grund, Stefan: Metadata Sharing – How to Transfer Metadata Information among Work Cluster Members, in: Cataloging & Classification Quarterly, 59 (8), 2021, S. 757-774. Online: <https://doi.org/10.1080/01639374.2021.1989101>.

- Wiechmann, Swantje: Examining concepts of author disambiguation – co-authorship as a disambiguation feature in EconBiz, Bachelorarbeit Hochschule für Technik, Wirtschaft und Kultur Leipzig, 2022. Online: <https://nbn-resolving.org/urn:nbn:de:bsz:l189-qucosa2-788413>.