Stephanie Glagla-Dietz, Nicole Habermann

Standardnummern für Personen

Qualitätsverbesserung durch das Zusammenspiel intellektueller und maschineller Formalerschließung

Tag für Tag sorgen Erschließer*innen mittels Normdaten für eine hohe Qualität der Katalogisate. Für die große Menge Netzpublikationen[1], deren Metadaten aus Fremddaten in den Katalog der Deutschen Nationalbibliothek (DNB) übernommen werden, ist dies nicht möglich – für sie gilt ein eigenes Erschließungskonzept.[2] Doch auch zwischen Feierabend und Arbeitsbeginn werden die Metadaten bearbeitet. Bereits seit 2011 werden in der DNB maschinelle Verfahren[3] eingesetzt, um Nacht für Nacht Titeldaten untereinander und Titel- mit Normdaten zu verknüpfen.

Seit April 2020 werden für diese Verknüpfungen nun auch andere Standardnummern genutzt, angefangen mit ORCID und ISNI. Immer häufiger werden sie in den Personenangaben der Netzpublikationen mitgeliefert.[4] In die Gemeinsame Normdatei (GND) kommen sie auf unterschiedlichen Wegen. In der maschinellen Formalerschließung sorgen diese Standardnummern für eine höhere Qualität, in der intellektuellen Erschließung helfen sie zeitsparend bei der Individualisierung.

Erster Schritt mit ORCID iDs

Im ersten Schritt liegt der Fokus auf wissenschaftlichen Publikationen unter Einbindung einer externen Informationsquelle. Seit Mai 2016 können in GND-Datensätze ORCID iDs in das Feld für Standardnummern eingetragen werden.[5] Mit mittlerweile 9,2 Millionen ORCID iDs (Stand: 31. Juli 2020) hat sich das Identifikationssystem orcid.org zu einem internationalen Standard in der Wissenschaft etabliert. Die in Deutschland registrierten ORCID iDs sind von etwa 44.000 im April 2016 auf rund 200.000 im Juli 2020 gewachsen. Deutschlandbezug haben weitere Wissenschaftler*innen, die beispielsweise an deutschen Hochschulen promoviert haben, jetzt jedoch an Hochschulen in der ganzen Welt arbeiten. Für die Angaben in den ORCID-Records werden Datenbanken angefragt und verlinkt, beispielsweise zu affiliierten Organisationen, Projekten und Publikationen. Wissenschaftler*innen sparen Zeit, wenn sie diese Angaben mit nur wenigen Klicks in ihren ORCID-Record übernehmen und die Metadaten dadurch vereinheitlichen. Dabei können sie selbst bestimmen, welche der Angaben für jeden sichtbar und welche nur für bestimmte Organisationen abfragbar sind.

Die Eignung der ORCID iDs zur Qualitätsverbesserung maschineller Formalerschließung wurde im Rahmen des von der Deutschen Forschungsgemeinschaft (DFG) geförderten Projekts ORCID DE[6] geprüft. Das Ergebnis war so vielversprechend, dass nach Abgleichverfahren mit den 5,5 Millionen GND-Personendatensätzen in mehreren Etappen insgesamt 63.547 ORCID iDs in die GND eingespielt werden konnten[7] (siehe Abbildung 1).

Die Abgleichverfahren beschränken sich auf die besten Übereinstimmungen. Eingespielte ORCID iDs werden mit einem Herkunftskennzeichen versehen. Zum einen werden die Affiliationen in ORCID-Records und GND-Datensätzen verglichen

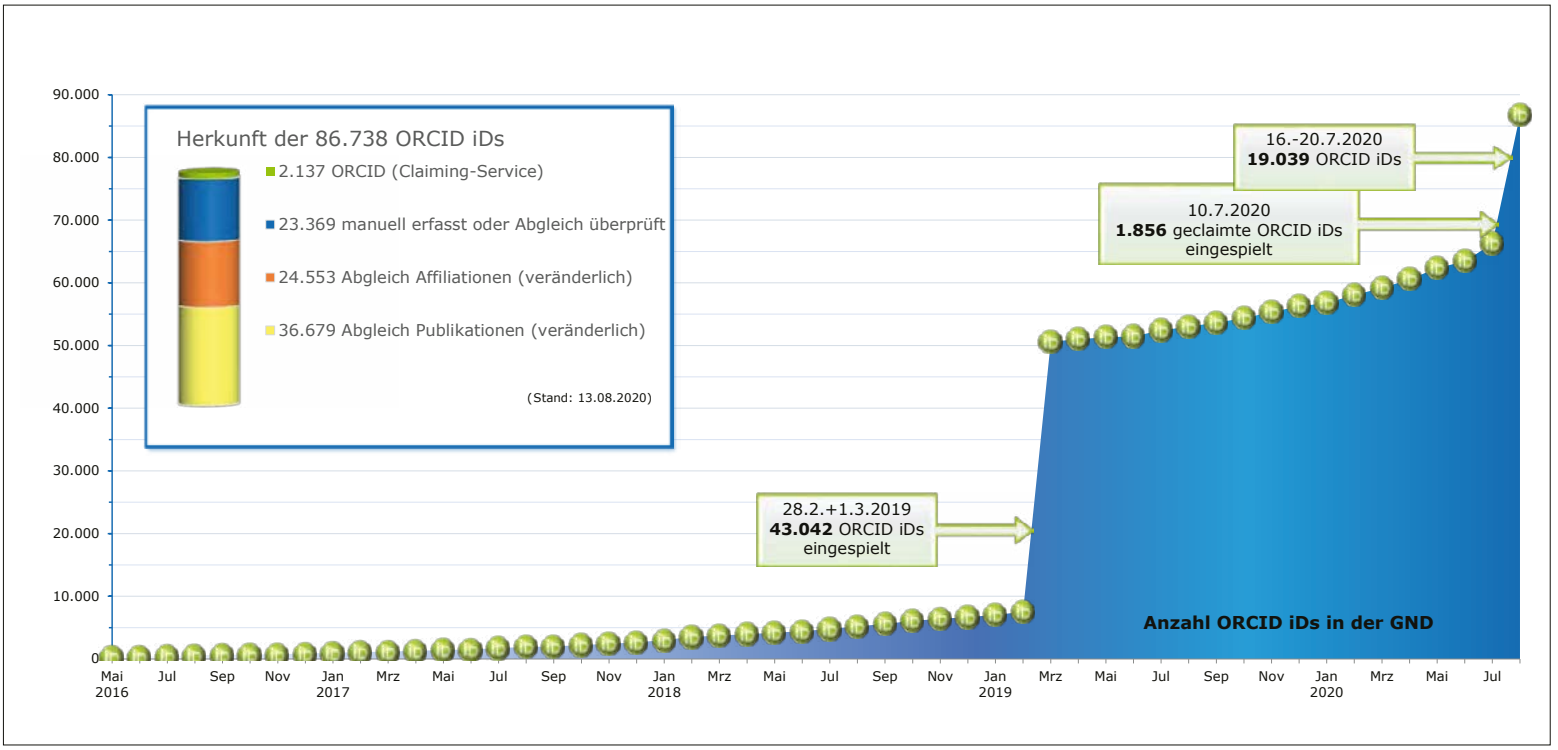

Abbildung 1: Herkunft der ORCID iDs in der GND

(gegebenenfalls bis auf Institutsebene, kleinere Abweichungen matchen nicht); dabei werden nur vollständig übereinstimmende Personennamen berücksichtigt und Personen mit Lebensdaten vor dem 20. Jahrhundert ausgeschlossen.

In einem zweiten Verfahren entspricht mindestens eine Publikation im ORCID-Record (works section) einem Titeldatensatz im deutschsprachigen Raum (Culturegraph[8] mit 170 Millionen Titeldaten), der mit einem GND-Datensatz verknüpft ist; dafür war die Übereinstimmung des normalisierten Textstrings (Matchschlüssel) aus Name und Titel ausschlaggebend.[9]

Schließlich werden die in BASE[10] geclaimten GND-IDs ausgewertet. In der weltweit größten Suchmaschine für Open-Access-Publikationen können ORCID-Nutzer*innen ihre Publikationen bereits seit Juni 2017 claimen. Seit Dezember 2018 wird der zusätzliche Service angeboten, direkt den eigenen GND-Datensatz zu identifizieren und im BASE-Profil zu verlinken. 9.216 ORCID-Nutzer*innen haben bisher ihren GND-Datensatz in BASE geclaimt (Stand: 31. Juli 2020). Im September 2020 werden anhand dieser Claimings die ORCID iDs in die GND-Datensätze eingespielt und mit einem eigenen Herkunftskennzeichen versehen.

Die mit den oben genannten Verfahren durch widersprüchliche Matches aufgespürten Dubletten bleiben von der Einspielung ausgenommen. Sie werden jedoch aufgezeichnet und anschließend in der GND-Zentrale bereinigt.

Ein Jahr Claiming-Service in der Deutschen Nationalbibliografie

Seit Juli 2019 ist es allen ORCID-Nutzer*innen möglich, ihre in der Deutschen Nationalbibliografie gelisteten Publikationen aus ihren ORCID-Records heraus zu claimen.[11]

Abbildung 2: Mit dem Claiming-Service der Deutschen Nationalbibliothek ergänzen ORCID-Nutzer*innen ihre Publikationsliste aus der Deutschen Nationalbibliografie und ihre GND-ID

Vor allem Forscher*innen im deutschen Sprachraum erleichtert der Claiming-Service, ihre Publikationsliste zu ergänzen und sich an das weltweite Normdatennetzwerk anzuschließen. Dabei wird die Verzahnung mit der GND »ganz nebenbei« erreicht, indem man sich selbst als beteiligte Person einer eigenen, im Katalog der DNB nachgewiesenen Publikation auswählt. Die Metadaten der Publikation werden in die Publikationsliste des ORCID-Records übernommen und mit dem DNB-Katalog verlinkt. Wenn der eigene GND-Datensatz in dieser Publikation verlinkt ist, wird die GND-ID gleichzeitig im ORCID-Record unter »Other IDs« verlinkt. Mit der Deutschen Nationalbibliografie stehen den ORCID-Nutzer*innen die Metadaten sämtlicher deutschsprachiger und in Deutschland erschienener Publikationen zur Verfügung.

Diese einfache »Search&link«-Funktion haben 5.173 ORCID-Nutzer*innen im ersten Jahr für 27.968 Titeldatensätze der Deutschen Nationalbibliografie genutzt (siehe Abbildung 3). Durchschnittlich wurden 5,4 Publikationen geclaimt. Ein Drittel der Personenangaben in diesen Publikationen waren mit einem GND-Datensatz verknüpft. So konnten durch das Claiming inzwischen 2.089 ORCID iDs in den GND-Datensätzen bestätigt und beidseitig verknüpft werden (Stand: 31. Juli 2020). Das entspricht 40 Prozent der »Claimer«.

Die geclaimten Publikationen werden ausgewertet und passende ORCID iDs in die Personenfelder der Titeldaten im DNB-Katalog geschrieben. Das ermöglicht Verknüpfungen zum GND-Personennormdatensatz. Damit keine versehentlich geclaimten Publikationen verknüpft werden, werden

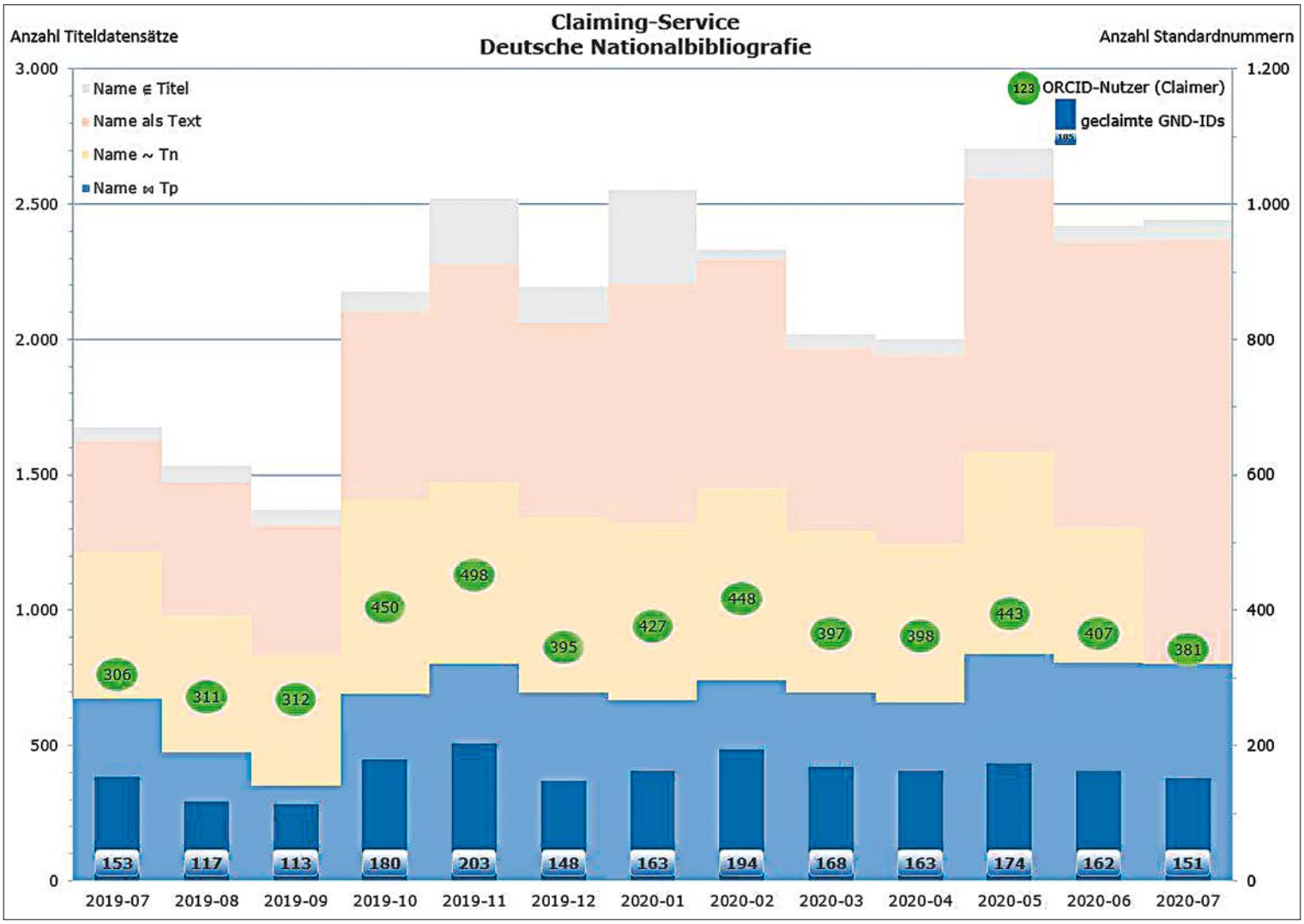

Abbildung 3: Statistik des ersten Claiming-Service-Jahres

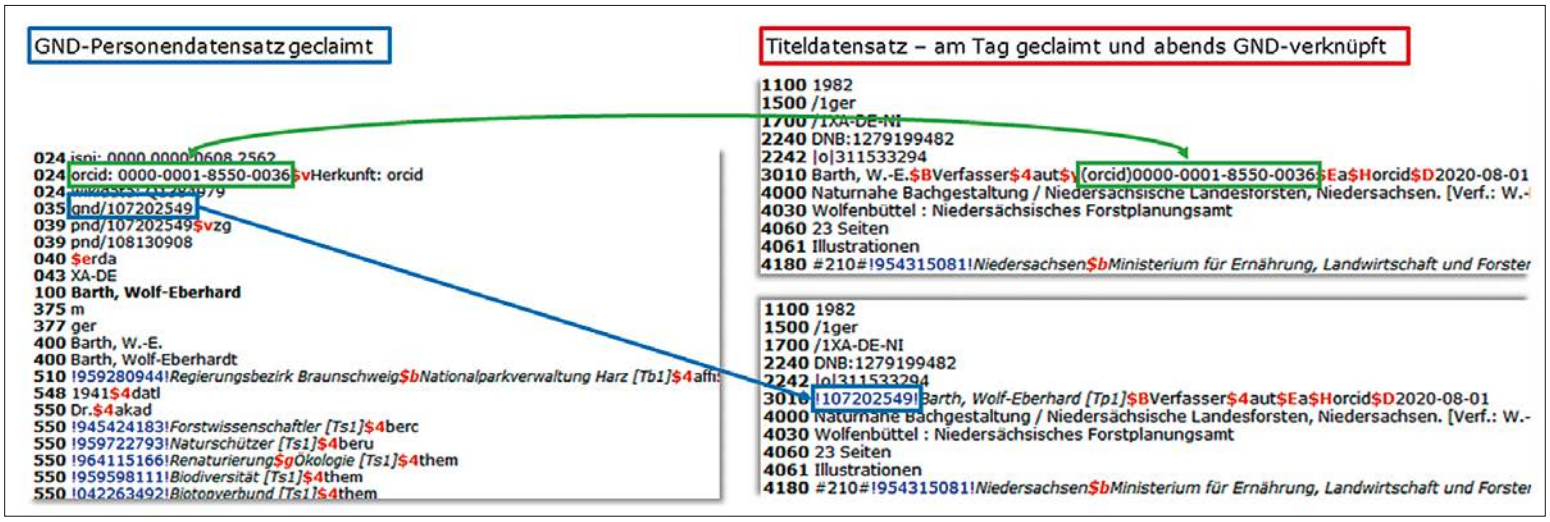

Abbildung 4: Import einer ORCID iD nach dem Claimen einer Publikation und die nachfolgende GND-Verknüpfung

die Personennamen zusätzlich abgeglichen. Diese Auswertung läuft jeden Abend für die tagsüber geclaimten Publikationen. Retrospektiv wurden am 10. Juli 2020 in alle geclaimten GND-Datensätze die ORCID iDs ergänzt. Vom 19. bis 21. August 2020 wurden in alle vor Juli 2020 geclaimten Titeldatensätze die ORCID iDs eingetragen und, sofern möglich, mit GND-Datensätzen verknüpft.
Die Herkunft einer ORCID iD wird im Feld für die Standardnummer gekennzeichnet. Inzwischen ist in 86.738 GND-Personendatensätzen die jeweilige ORCID iD enthalten (Stand: 31. Juli 2020).

Personen-Standardnummern in Titeldaten

Befindet sich die ORCID iD im GND-Datensatz, werden alle abgelieferten Publikationen, zu denen der Verlag die ORCID iD mitliefert, in der folgenden Nacht verknüpft (siehe Beispiel in Abbildung 4). Nachfolgendes Claiming verstetigt eine bereits bestehende Verknüpfung.
Täglich kommen bereits rund 500 Publikationen mit mindestens einer Name-ORCID-iD-Kombination in den DNB-Katalog. Aber auch das Hinzufügen einer ORCID iD in einem GND-Datensatz, beispielsweise in einer Hochschulbibliothek, löst in der folgenden Nacht die Suche nach Publikationen mit einer solchen Name-ORCID-iD-Kombination aus und verknüpft bereits im DNB-Katalog vorhandene Publikationen. Die Metadaten dieser Publikationen stehen fortan weltweit mit der GND-Verknüpfung zur Verfügung.
Das Unterfeld, das in den Titeldaten die Personen-Standardnummer aufnimmt, gibt es seit Juli 2017. Seitdem werden in jedem Monat über 10.000 Publikationen mit mindestens einer Personen-Standardnummer abgeliefert, insgesamt sind es knapp 400.000 (Stand: 31. Juli 2020).

Mitarbeit erwünscht

Mit den ORCID-Records hat ORCID die Möglichkeit geschaffen, als Wissenschaftler*in selbst dazu beizutragen, dass Titel-Zuordnungen in Normdateien, Bibliothekskatalogen und anderen Datenbanken korrekt sind. Voraussetzung ist das gewissenhafte Claiming in den Datenbeständen, die die ORCID-API implementiert haben. Hochschulbibliotheksmitarbeiter*innen können in Gesprächen mit Bibliotheksnutzer*innen auf die Vorteile dieser Claiming-Möglichkeiten hinweisen.
Damit die maschinellen Verfahren und nächtlichen Verknüpfungen funktionieren können, ist ein Mindestmaß an Qualität der GND-Datensätze erforderlich. Hat eine an einer Publikation beteiligte Person einen ORCID-Record, sollte ihre ORCID iD als Standardnummer in den GND-Datensatz eingetragen werden[12].

Noch nicht jede*r Wissenschaftler*in hat einen ORCID-Record angelegt. Darüber hinaus ist es jederzeit möglich, jede im ORCID-Record hinterlegte Information zu verdecken oder ganz zu löschen – einschließlich des Namens. Aus diesen Gründen sollten zur Identifizierung geeignete Daten auch in den GND-Datensatz eingetragen werden. Hierzu eignen sich Verknüpfungen zu Arbeitgebern (Affiliationen), aussagekräftigen Berufen (also nicht nur allgemein »Wissenschaftler«), gegebenenfalls Forschungsthemen und Angaben, wie die wesentlichen Publikationen, Wirkungsdaten, Ländercodes und der akademische Grad. Je rudimentärer ein GND-Datensatz erfasst wird, umso größer ist die Gefahr, dass namensgleiche Personen verwechselt werden. Wird einem GND-Datensatz in einem Katalog ein falscher Titeldatensatz zugeordnet, ist diese Unstimmigkeit auch in Culturegraph enthalten und bleibt in einem Abgleichverfahren unberücksichtigt. Zur Vermeidung der Fehlerpotenzierung ist es hilfreich, auf fehlerhafte Verknüpfungen hinzuweisen.

Ausblick

Fehlerhafte Zuordnungen aus den Abgleichverfahren aufzuräumen ist eine Aufgabe der GND-Zentrale. Bei künftigen Abgleichen sollen häufige Namen anders behandelt werden. Zu diesem Zweck werden für den Abgleich mit dem ORCID-Dump 2020 Listen häufiger Namen erstellt und erprobt. Ziel ist es, bei gleichbleibender Qualität unter Ausschluss dieser häufigen Namensformen mehr Treffer für den Abgleich berücksichtigen zu können.

Fehlerhafte Titel-GND-Verknüpfungen müssen von der jeweiligen Bibliothek bereinigt werden. Aber auch Aufräumarbeiten an GND-Datensätzen können langfristig nicht von der GND-Zentrale alleine durchgeführt werden. Die Zusammenführung von Dubletten, das Splitting einzelner Datensätze und die Anreicherung von GND-Datensätzen mit Berufen, Themen und Affiliationen kann langfristig nur durch viele GND-Anwender*innen bewältigt werden. Fehlerhafte Datensätze müssen an geeigneter Stelle dokumentiert, ihre Bereinigung als Aufgabe der GND-Anwendergemeinschaft gesehen und möglichst viele GND-Anwender*innen in geeigneter Weise beteiligt werden. Je größer die Beteiligung, desto wirksamer sind die maschinellen Verfahren. Das Engagement zahlt sich also am Ende für alle aus.

Im Projekt ORCID DE 2 ist der Ausbau der Abgleichverfahren durch Einbeziehung weiterer Metadaten, wie Forschungsthemen, Berufe, Ländercodes, Wirkungsdaten und im GND-Datensatz gelisteter Publikationen, vorgesehen. Ab Herbst 2020 werden die Möglichkeiten DNB-intern diskutiert und in den nächsten Abgleichverfahren mit dem ORCID-Dump 2020 getestet.

Ebenfalls im Projekt ORCID DE 2 ist die Nutzung einer ORCID-API vorgesehen, um den ORCID-Nutzer*innen, deren ORCID iD bereits im GND-Datensatz eingetragen wurde, eine Claiming-Aufforderung zu senden. Dies wird die Qualität der Verknüpfungen weiter erhöhen.

ORCID-Nutzer*innen sollen ihre Publikationen in allen Bibliotheksbeständen des deutschen Sprachraums claimen können. Um diesem Ziel näher zu kommen, ist im Projekt ORCID DE 2 die Ausweitung des Claiming-Service auf Culturegraph-Daten geplant. Nach dem Launch der neuen GND-Recherche-Plattform (GND-Explorer) 2021, wird darin eine Claiming-Möglichkeit für alle in Culturegraph enthaltenen Publikationen entwickelt.

Darüber hinaus wird an einem Vorschlagssystem für GND-Datensätze gearbeitet, bei dem auch Inhalte, wie Länder, Affiliationen, akademische Grade, Publikationen und mögliche Themen und Berufe für neue GND-Datensätze berücksichtigt werden. Voraussetzung ist die strukturierte Erfassung dieser Informationen im ORCID-Record, in ISNI oder in Wikidata. Diese Vorschläge sollen es GND-Anwender*innen erlauben, auf einfache und schnelle Art einen neuen GND-Datensatz zu erstellen, in dem dann auch die ORCID iD vorhanden ist. Zuerst sollen Vorschläge für Personennormdatensätze für die ORCID-Nutzer*innen erstellt werden, die Publikationen in der Deutschen Nationalbibliografie geclaimt haben, für die es jedoch noch keinen GND-Datensatz gibt.

Parallel wird im Projekt GND4P (GND für Verlage) mit der MVB[13] daran gearbeitet, ähnliche Verfahren mit Hilfe der ISNI aufzubauen. Die MVB ist dafür 2019 ISNI-Mitglied geworden und wird geeignete Autorenseiten im Verzeichnis lieferbarer Bücher

(VLB) aufbauen[14]. Auch ISNI-IDs sind als Standardnummern bereits in der GND vorhanden. Der Abgleich wird eingerichtet, sobald in den Personenangaben der Publikations-Metadaten vermehrt ISNI-IDs geliefert werden.

In die GND über verschiedene Quellen manuell erfasste Standardnummern sorgen mit den etablierten und noch zu ergänzenden maschinellen Verfahren für eine stetig wachsende Zahl von Verknüpfungen im DNB-Katalog. Diese sind in Culturegraph und Entity Facts enthalten und stehen somit weltweit zur Verfügung. Ein Mindestmaß an Qualität kann jedoch nur durch intellektuelle Erschließung gewährleistet werden. Durch den Claiming-Service haben Erschließer*innen nun die indirekte Unterstützung seitens der wissenschaftlich Publizierenden – andere Autor*innen und Verlage werden folgen. So können sich maschinelle und intellektuelle Formalerschließung bei schnell wachsenden Beständen sinnvoll ergänzen.

Anmerkungen

1 8,64 Millionen (Stand: 31. Juli 2020)

2 siehe Gömpel, Renate; Junger, Ulrike; Niggemann, Elisabeth: Veränderungen im Erschließungskonzept der Deutschen Nationalbibliothek. In: Dialog mit Bibliotheken, 22 (2010) 1, S. 20–22. <https://nbn-resolving.org/urn:nbn:de:101-2011012858> und Schöning-Walter, Christa: Automatische Erschließungsverfahren für Netzpublikationen. Zum Stand der Arbeiten im Projekt PETRUS. In: Dialog mit Bibliotheken, 23 (2011) 1, S. 31–36. <https://nbn-resolving.org/urn:nbn:de:101-2011101170>

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

4 Inzwischen sind es täglich mehr als 500 Netzpublikationen mit mindestens einer beteiligten Person, für die eine ORCID iD mitgeliefert wird (zum Beispiel SpringerLink, Universitäten). Insgesamt wurden knapp 400.000 Netzpublikationen mit mindestens einer ORCID iD abgeliefert, von denen mehr als 24.000 mindestens mit einem GND-Personendatensatz verknüpft werden konnten (Stand: 31. Juli 2020).

5 Voraussetzung für die Erfassung in Feld 024 <https://wiki.dnb.de/x/vYYGAw> ist der Eintrag des Standardnummernsystems in die Liste Standard Identifier Sources der Library of Congress <http://www.loc.gov/standards/sourcelist/standard-identifier.html> (Stand: 31. Juli 2020). Siehe auch Abbildung 5.

6 siehe <https://www.orcid-de.org> (Stand: 31. Juli 2020).

7 siehe Blogpost <https://www.orcid-de.org/mehr-als-50-000-personendatensaetze-der-gemeinsamen-normdatei-gnd-mit-orcid-recordsverknuepft> (Stand: 31. Juli 2020). Zwischen dem 16. und 20. Juli 2020 fanden die Einspielungen nach Datenabgleich mit dem ORCIDDump 2019 statt.

8 siehe Vorndran, Angela: Hervorholen, was in unseren Daten steckt! Mehrwerte durch Analysen großer Bibliotheksdatenbestände. In: O-Bib. Das Offene Bibliotheksjournal, 5 (2018) 4, S. 166–180. <https://doi.org/10.5282/o-bib/2018H4S166-180> (Stand: 31. Juli 2020

9 siehe Glagla-Dietz, Stephanie; Vorndran, Angela: Nutzung von Matching-Verfahren zur GND-Anreicherung aus externen Quellen – Vorteile für Metadatenqualität und Erschließung. Kommentierte Folien eines Vortrags auf dem 7. Bibliothekskongress 2019 in Leipzig. <http://nbn-resolving.de/urn:nbn:de:0290-opus4-163741> (Stand: 31. Juli 2020)

10 Bielefeld Academic Search Engine, siehe <https://www.base-search.net/about/de/legal_notice.php> (Stand: 31. Juli 2020) und zur Zusammenarbeit mit BASE siehe Walger, Nadine: Sammlung von Netzpublikationen erreicht nächste Stufe. In: Dialog mit Bibliotheken, 29 (2017) 2, S. 14–16. <https://nbn-resolving.org/urn:nbn:de:101-20170929458>

11 siehe Blogpost <https://www.orcid-de.org/orcid-claiming-gnd> (Stand: 31. Juli 2020)

12 Zur Unterstützung der WinIBW-Nutzer steht ein Script zur Verfügung, mit dem zeitsparend nach einem ORCID-Record gesucht werden kann, siehe <https://wiki.k10plus.de/x/CgCUC> (Stand: 31. Juli 2020).

13 siehe <https://mvb-online.de> (Stand: 31. Juli 2020)

14 siehe <https://www.boersenblatt.net/archiv/1745103.html> und <https://german-isbn.de/isni/die-isni> (beide Stand: 31. Juli 2020)