Hervorholen, was in unseren Daten steckt! Mehrwerte durch Analysen großer Bibliotheksdatenbestände

Angela Vorndran, Deutsche Nationalbibliothek[1]

Zusammenfassung
Die Deutsche Nationalbibliothek (DNB) verfolgt das Ziel, den unter Culturegraph.org verfügbaren großen Datenbestand von mehr als 160 Millionen Titeldaten deutschsprachiger Bibliotheksverbünde sowie der Deutschen Nationalbibliothek und der British National Bibliography über Analysen, Verknüpfungen und Auswertungen in größerem Umfang nutzbar zu machen. Der Beitrag gibt einen Überblick, welche Themenstellungen und Methoden bislang im Zentrum stehen. Dies ist einerseits die Bündelung von Werken, die erlaubt, mehrere Ausgaben, Auflagen oder Übersetzungen eines Werks zusammenzuführen. Inhaltserschließende Informationen wie Klassifikation oder verbale Erschließung, ebenso wie Normdatenverknüpfungen, können so auf alle Mitglieder eines Bündels übertragen werden, so dass ein Gewinn an Standardisierung und Erschließungstiefe zu erreichen ist. Andererseits können über bibliothekarische Daten hinaus auch externe Datenquellen zur Anreicherung herangezogen werden. Dies wird anhand eines Abgleichs von Personen in der Gemeinsamen Normdatei (GND) und der Datenbank Open Researcher and Contributor ID (ORCID) dargestellt. Unter Verwendung der Culturegraph-Titeldaten werden Personen mittels der von ihnen verfassten Publikationen abgeglichen und zusammengeführt. Abschließend werden einige statistische Auswertungen des Datenbestandes vorgestellt.

Summary
The German National Library (DNB) strives to make use of the large numbers of bibliographic records in culturegraph.org. More than 160 millions of records originating from German and Austrian regional library networks, the British National Bibliography and DNB may be used for data analyses, evaluation of connections and statistical analyses. This paper gives an overview of the central topics: On the one hand, the clustering of works to comprise different editions and translations of a work. Indexing and classification information as well as links to authority data can then be shared among the members of each cluster to achieve a surplus in standardization and subject indexing. On the other hand, external data can serve as sources for enrichment of bibliographic records. This is exemplified by matching data from the Open Researcher and Contributor ID (ORCID) with the Integrated Authority File (GND). Using bibliographic records from Culturegraph, persons are matched on the basis of their publications´ titles. Finally, a few statistical analyses of the aggregated data are presented.

Zitierfähiger Link (DOI): https://doi.org/10.5282/o-bib/2018H4S166-180
Autorenidentifikation: Vorndran, Angela: GND: 1126308366
ORCID: https://orcid.org/0000-0001-7162-9875
Schlagwörter: Metadatenanalyse; Normdatenanreicherung; Datenabgleichsverfahren.

1 Die hier vorgestellten Verfahren wurden im Team Datenmanagement der Deutschen Nationalbibliothek unter maßgeblicher Mitarbeit von Jan Eberhardt und Stefan Grund entwickelt.

1. Hintergrund

Die deutsche Bibliothekslandschaft zeichnet sich durch ihre föderale Struktur aus. Insgesamt sechs regional organisierte Bibliotheksverbünde bilden jeweils Zusammenschlüsse vieler Bibliotheken und schaffen für diese unter anderem gemeinsame Datenbestände. Diese sind der Bibliotheksverbund Bayern (BVB), der Gemeinsame Bibliotheksverbund (GBV), der hbz-Verbund beim Hochschulbibliothekszentrum des Landes Nordrhein-Westfalen (HBZ), das Hessische Bibliotheks-Informationssystem (HeBIS), der Kooperative Bibliotheksverbund Berlin-Brandenburg (KOBV) und der Südwestdeutsche Bibliotheksverbund (SWB).

Mit Culturegraph[2] bietet die Deutsche Nationalbibliothek (DNB) eine Plattform, auf der die bibliographischen Metadaten aller deutschen Verbünde sowie des Österreichischen Bibliotheksverbundes (OBV), der British National Bibliography und der Deutschen Nationalbibliothek basierend auf den Datenlieferungen der Partner aggregiert werden und zur Analyse zur Verfügung stehen. Die aggregierten Datenbestände sollen unter anderem für Datenanalysen, Datenabgleiche und weitergehende Vernetzung der Bestände genutzt werden. Zurzeit handelt es sich um einen Datenbestand von über 160 Millionen Datensätzen (Stand: Juli 2018). In diesem Zusammenhang sind vielfältige Anwendungen denkbar, von denen hier drei Bereiche eingehender dargestellt werden sollen: das Bündeln von Werken, das Einbeziehen externer Datenquellen zur Datenanreicherung und die statistischen Auswertungen des Datenbestandes.

2. Bündelung von Werken

2.1. Frühere Ansätze zur Werkbündelung

Der Abgleich von Publikationen und die Zusammenfassung zu Werkbündeln kann anhand verschiedener Bestandteile der beschreibenden Metadaten erfolgen. Eine naheliegende und häufig verwendete Vorgehensweise ist der Abgleich von Autor/inn/en und Titeln. Dies wird beispielsweise in dem vom Online Computer Library Center (OCLC) erstellten Functional Requirements for Bibliographic Records (FRBR) Work-Set Algorithm[3] vorgenommen. Hier werden die Akteur/e/innen, die für die Schaffung eines Werks verantwortlich sind, in Autor/inn/en für die die MARC-Felder 100, 110 und 111 ausgewertet werden und Namen (MARC-Felder 700, 710, 711) unterteilt, verschiedene Titelformen (MARC-Felder 240, 242, 245, 246, 247, 740) nach Präferenzen gruppiert sowie eine weitere Nachverarbeitung vorgesehen. Titel werden um wenig aussagekräftige Begriffe bereinigt und, wie auch Personennamen, an Normdaten abgeglichen.

In Ergänzung zu diesem Algorithmus wurden von OCLC in dem Projekt GLIMIR (Global Library Manifestation Identifier)[4] weitere Schritte unternommen, auch Datensätze für gleiche Manifestationen zu aggregieren. Hieraus ergaben sich weitere relevante Ansatzpunkte auch für den Abgleich

2 culturegraph, <http://hub.culturegraph.org/relo>, Stand: 23.11.2018.
3 Hickey, Thomas B.; Toves, Jenny: FRBR Work-Set Algorithm. Version 2.0, 2009, <https://www.oclc.org/content/dam/research/activities/frbralgorithm/2009-08.pdf>, Stand: 23.11.2018.
4 Gatenby, Janifer; Greene, Richard O.; Oskins, W. Michael u.a.: GLIMIR: Manifestation and Content Clustering within WorldCat, in: code{4}lib Journal 17 (2012), <http://journal.code4lib.org/articles/6812>, Stand: 23.11.2018.

von Werken wie z.B. die Relevanz weiterer Felder für zu berücksichtigende Inhalte. Beispielsweise wird MARC-Feld 720[5] für Autor/inn/en herangezogen und eine Inspektion ergab 11 verschiedene Felder, die für originalsprachliche Titel verwendet werden. Außerdem wird die Notwendigkeit zur Normalisierung von Titelinformationen beispielsweise durch Weglassen irrelevanter Informationen und Normalisierung von Abkürzungen betont.

Für den deutschsprachigen Bestand wurde ebenfalls ein Abgleichverfahren entwickelt, das auf einem Vergleich von Titel, Titelzusatz und beteiligten Personen und Körperschaften basiert (vgl. Pfeffer,[6] Wiesenmüller/Pfeffer[7]). Hier wird ein Abgleich von Einheitstitel, bzw. wenn nicht vorhanden, Titel und Titelzusatz vorgenommen. Dabei werden die Normdateneinträge aller in Beziehung stehenden Personen und Körperschaften, z.B. Autor/inn/en und Mitwirkende, in den Abgleich mit einbezogen. Ein Bündel entsteht bei exakt gleichem Titel und Übereinstimmung einer in Beziehung stehenden Person.

Weitere ähnliche Ansätze sind auch in der Übersicht von Pfeifer/Polak-Bennemann[8] zu finden.

2.2. Werkdefinition

Das Zusammenfassen mehrerer Publikationen zum selben Werk bietet weitgehende Möglichkeiten, um inhaltserschließende Information zu übertragen und die Erschließung zu vereinheitlichen. In der Vergangenheit wurden, wie oben beschrieben, in verschiedenen Zusammenhängen Verfahren entwickelt, die automatisiert eine Zusammenführung von einzelnen Publikationen zu Werken erreichen sollten. Eine Herausforderung, der sich alle entsprechenden Verfahren stellen müssen, ist die eindeutige Definition eines Werkes. Die mit den Functional Requirements for Bibliographic Records (FRBR) eingeführten Entitäten Exemplar, Manifestation, Expression und Werk geben erste Richtlinien für die Definition eines Werks vor. In der ersten Fassung von 1998 ist zu lesen: „the *work* itself exists only in the commonality of content between and among the various *expressions* of the *work*".[9]

Genauer wird ausgeführt:

> "Similarly, abridgements or enlargements of an existing text, or the addition of parts or an accompaniment to a musical composition are considered to be different *expressions* of the same *work*. Translations from one language to another, musical transcriptions and arrangements, and dubbed or subtitled versions of a film are also considered simply as different *expressions* of the same original *work*. [...]

5 Nebeneintragung – nicht normierter Name, <http://www.loc.gov/marc/bibliographic/bd720.html>, Stand: 23.11.2018.

6 Pfeffer, Magnus: Using Clustering Across Union Catalogues to Enrich with Indexing Information, in: Spiliopoulou, Myra; Schmidt-Thieme, Lars; Janning, Ruth (Hg): Data Analysis, Machine Learning and Knowledge discovery, Cham 2014, S. 437-445.

7 Wiesenmüller, Heidrun; Pfeffer, Magnus: Abgleichen, anreichern, verknüpfen, in: BuB 35 (9), 2013, S. 625–629. Online: <http://www.b-u-b.de/pdfarchiv/Heft-BuB_09_2013.pdf>, Stand: 23.11.2018.

8 Pfeifer, Barbara; Polak-Bennemann, Renate: Zusammenführen was zusammengehört – Intellektuelle und automatische Erfassung von Werken nach RDA, in: o-bib. Das offene Bibliotheksjournal 3 (4), 2016, S. 144–155, <https://doi.org/10.5282/o-bib/2016h4s144-155>.

9 IFLA: Functional Requirements for Bibliographic Records, Final Report, 1998, <https://www.ifla.org/files/assets/cataloguing/frbr/frbr_2008.pdf>, Stand: 23.11.2018, S. 16 f.

> By contrast, when the modification of a *work* involves a significant degree of independent intellectual or artistic effort, the result is viewed, for the purpose of this study, as a new *work*. Thus paraphrases, rewritings, adaptions for children, parodies, musical variations on a theme and free transcriptions of a musical composition are considered to represent new *works*."[10]

Die in den nachfolgenden Jahren verwendeten Definitionen eines Werks insbesondere im Rahmen der Regeln zu Resource Description and Access (RDA)[11] und des Library Reference Models (LRM)[12] orientieren sich an dieser Definition. Die Unterscheidung bezüglich der Zugehörigkeit zu einem Werk hängt somit stark von dem Grad des intellektuellen und künstlerischen Schaffens bei der Überarbeitung eines Originalwerkes ab. Wird das Originalwerk nur in der Länge oder der Sprache verändert, handelt es sich um dasselbe Werk. Beziehen sich die Überarbeitungen stärker auf den Inhalt, erfährt die sprachliche Darstellung weitergehende Veränderungen (beispielsweise zur Anpassung an eine Zielgruppe oder ein Genre) oder wird die literarische Gattung oder Medienform verändert, entsteht ein neues Werk. Der Grad des intellektuellen Aufwandes bei der Überarbeitung ist, vor allem bei der Verwendung automatisierter Verfahren und nur auf der Basis von Metadaten, allerdings nicht immer eindeutig zu bestimmen. Mit den in RDA vorgesehenen Spezifikationen zu Adaptionen und Überarbeitungen kann dies allerdings erleichtert werden[13].

In den in Culturegraph zur Verfügung stehenden Metadaten lassen sich zur Klärung dieser Fragestellung Angaben zum Titel, Titelzusatz und den an der Schaffung des Werks beteiligten Akteure und Akteurinnen heranziehen. Eine ergänzende Analyse der Volltexte des Originalwerkes und der überarbeiteten Expression ist in diesem Rahmen nicht möglich.

2.3. Werkbündelung in Culturegraph

2.3.1. Ausgangslage

Die Ausgangslage der hier dargestellten Aktivitäten zur Werkbündelung bildet ein bereits seit 2013 in Culturegraph verwendeter Bündelungsalgorithmus, der unter anderem die Java-Bibliothek Metafacture verwendet.[14] Hier wird, vergleichbar zu dem von Pfeffer[15] vorgestellten Verfahren, über Matchkeys eine Identifikation von Publikationen vorgenommen, die in einem anschließenden Abgleich der Schlüssel zu Bündeln zusammengefasst werden.

Die in der Ursprungsversion verwendeten Schlüssel kombinieren folgende Angaben:

10 Ebd.

11 „Werk: ein individuelle intellektuelle oder künstlerische Schöpfung, d.h. der intellektuelle oder künstlerische Inhalt", RDA Kapitel 5.1.2, <https://access.rdatoolkit.org/index.php>.

12 „The essence of the work is the constellation of concepts and ideas that form the shared content of what we define to be expressions of the same work. A work is perceived through the identification of the commonality of content between and among various expressions.", Riva, Pat; Le Bœuf, Patrick; Žumer, Maja: IFLA Library Reference Model. A Conceptual Model for Bibliographic Information, 2017, <https://www.ifla.org/files/assets/cataloguing/frbr-lrm/ifla-lrm-august-2017_rev201712.pdf>, Stand: 23.11.2018, S. 21 f.

13 RDA Kapitel 6.27.1.5, <https://access.rdatoolkit.org/index.php>.

14 Geipel, Markus Michael; Böhme, Christoph; Hannemann, Jan: Metamorph: A Transformation Language for Semi-Structured Data, in: D-Lib Magazine 21 (5/6), 2015, <https://doi.org/10.1045/may2015-boehme>. Metafracture bei GitHub unter <https://github.com/metafacture/metafacture-core>, Stand: 23.11.2018.

15 Pfeffer: Using Clustering Across Union Catalogues, 2014.

- EKI – Zählung eines Teils/einer Abteilung eines Werkes
  - Bei der EKI handelt es sich um die Erstkatalogisierungs-ID, die bei Datenaustausch oder Datenübernahme beibehalten wird und alle Titelaufnahmen einer Publikation verbinden soll.[16]
- ISBN – Zählung eines Teils/einer Abteilung eines Werkes
- Titel – Zählung eines Teils/einer Abteilung eines Werkes – Person

Der Titel einer Publikation wird aus dem Haupttitel, dem Titelzusatz und, wenn vorhanden, dem Titel eines Teils zusammengesetzt. Dies entspricht den MARC-Feldern 245 Unterfelder a, b und p. Als für die Publikation relevante Personen werden Autorinnen, Autoren und Beteiligte aus den MARC-Feldern 100 und 700 verwendet, allerdings mit der Einschränkung auf Personen mit der Beziehungskennzeichnung „aut“ für „author“ im jeweiligen Unterfeld 4.

2.3.2. Anpassungen des Algorithmus

In der Analyse der verwendeten Schlüssel wurde festgestellt, dass das Einbeziehen des kompletten Titelzusatzes und des Titels eines Teils häufig sehr spezifisch eine bestimmte Manifestation eines Werkes beschrieb und somit ein enger Werksbegriff angewendet wurde. Die in Tabelle 1 beispielhaft zusammengestellten Publikationen würden somit nicht in einem Werkbündel zusammengefasst.

Tabelle 1: Beispiel verschiedener Expressionen eines Werks mit abweichenden Titelzusätzen

MARC-Feld 245	MARC-Feld 100	MARC-Feld 700
$a Pschyrembel Klinisches Wörterbuch $b für MS Windows ; ca. 35000 Stichworte, über 2000 Abbildungen, davon über 800 in 16 Mio. Farben (Echtfarben) $c Pschyrembel		Pschyrembel, Willibald
$a Pschyrembel Klinisches Wörterbuch $h [Elektronische Ressource] $b jetzt mit englisch-deutschem, deutsch-englischem Glossar, Abkürzungen, Akronymen, Terminologia anatomica, weiteren Stichwörtern		Pschyrembel, Willibald
$a Pschyrembel Klinisches Wörterbuch $b [mit CD-ROM]		Pschyrembel, Willibald
$a Klinisches Wörterbuch $b Mit 763 Abb. im Text u.e. neubearb. u. erw. Nachtr. $c Willibald Pschyrembel. Gegr. von Otto Dornblüth	Pschyrembel, Willibald	
$a Klinisches Wörterbuch $b Mit 768 Abb. im Text u.e. neubearb. u. erw. Nachtr. $c Willibald Pschyrembel. Gegr. von Otto Dornblüth	Pschyrembel, Willibald	

16 Jaritz, Marko: Erstkatalogisierungs-ID, GBV Verbund-Wiki, 08.03.2017, <https://verbundwiki.gbv.de/display/VZG/Erstkatalogisierungs-ID>, Stand: 23.11.2018.

$a Klinisches Wörterbuch $b mit klin. Syndromen u. nomina anatomica $c W. Pschyrembel. [Gegr. von Otto Dornblüth]	Pschyrembel, Willibald	Dornblüth, Otto
$a Klinisches Wörterbuch $b Die Kunstausdrücke d. Medizin $c Otto Dornblüth. Neu durchges. u. erg. von Wilibald Pschyrembel	Dornblüth, Otto	Bannwarth, Emil Pschyrembel, Willibald
$a Klinisches Wörterbuch $b Die Kunstausdrücke d. Medizin $c Otto Dornblüth. Bearb. von Wilibald Pschyrembel	Dornblüth, Otto	Pschyrembel, Willibald

In einer Evaluierung durch VertreterInnen deutschsprachiger Bibliotheksverbünde wurde allerdings die Präferenz geäußert, eher kleinere und im Inhalt homogenere Bündel zu erstellen, um für eine mögliche automatisierte Übernahme von sacherschließenden Informationen eine höhere Genauigkeit zu erreichen. So wurde in der aktualisierten Version des Vergleichsschlüssels weiterhin die Titelangabe sowie der Titelzusatz (MARC-Feld 245 $b) berücksichtigt. Nun wird allerdings neben der Haupteintragung der Titelangabe (MARC-Feld 245) auch der Einheitstitel (MARC-Felder 130, 240, 700 Unterfeld t und 730) für die Schlüsselerstellung verwendet (vgl. Abbildung 1).

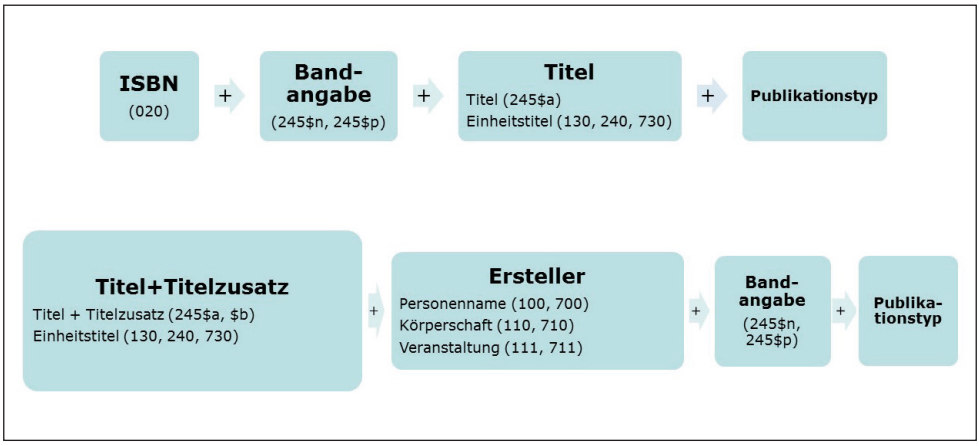

Abb.1: Überarbeitete Schlüssel zur Werkbündelung (mit Hinweis auf verwendete MARC-Felder)

Die im Datensatz genannten Personen und Institutionen werden in der überarbeiteten Version der Schlüsselerzeugung in größerem Ausmaß berücksichtigt. Bei genauerer Analyse zeigte sich, dass die Beschränkung auf Namen aus den MARC-Feldern 100 und 700 mit der Beziehungskennzeichnung „author“ eine große Anzahl an Namen unberücksichtigt ließ, da nicht in allen Fällen Beziehungskennzeichnungen erfasst waren. Außerdem existieren weitere Beziehungskennzeichnungen, die einen geistigen Schöpfer bezeichnen, wie beispielsweise „creator (cre)“, „compiler (com)“ oder „artist (art)“. Darüber hinaus sind weitere Beziehungskennzeichnungen durchaus relevant für die Identifikation von

Werken und den an ihrer Erstellung beteiligten Akteur/inn/en, vor allem wenn kein Eintrag mit der Beziehungskennzeichnung „author“ vorliegt. Hierzu zählen beispielsweise die Kategorien „contributor (ctb)“, „editor (edt)“, „translator (trl)“ oder auch „musician (mus)“ sowie „performer (prf)“ für nichttextuelle Medienformen.[17] Da nun auch Körperschaften und Veranstaltungen aus den MARC-Feldern 110, 111, 710 und 711 zur Schlüsselerstellung herangezogen werden, sind weitere Kennzeichnungen umso relevanter, da sie in diesen Feldern stark vertreten sind.

Weiterhin zeigte sich in der Überarbeitung der Schlüssel, dass Bündel mit gleichlautenden Erstkatalogisierungs-IDs einen geringen Zugewinn erzeugten. Im Fall von durch ISBN und Zählung eines Teils/einer Abteilung eines Werkes erstellten Bündeln fanden sich in exemplarischen Stichproben im Rahmen einer intellektuellen Evaluierung mehrere fehlerhafte Bündel, die durch fehlerhafte ISBNs oder Erfassung mehrerer ISBNs in einem Datensatz, z.B. bei Schriftenreihen, entstanden. Aus diesem Grund wurde in der Überarbeitung der Schlüssel zur Werkbündelung kein Schlüssel unter Verwendung der EKI erzeugt und der Schlüssel, der die ISBN enthält, um die Titelangabe und den Publikationstyp erweitert, um eine eindeutige Identifikation einer Publikation zu ermöglichen.

Für alle Kombinationen der verschiedenen Titelfelder und den an der Schaffung des Werks beteiligten Akteure und Akteurinnen werden Schlüssel erzeugt, mit der Einschränkung, dass maximal zwei Akteur/innen berücksichtigt werden. Die Zählung eines Teils/einer Abteilung eines Werkes ist ebenfalls Teil des Schlüssels. Ebenso wird die Publikationsform im Schlüssel angegeben, um den Vorgaben des Library Reference Models nachzukommen, dass die Überführung eines Werkes in eine andere Gattung, z.B. Verfilmung oder musikalische Komposition, ein neues Werk konstituiert.[18]

Folgende Maßnahmen zur Normalisierung der Personennamen und Titelangaben werden vorgenommen:

- Großbuchstaben werden durch Kleinbuchstaben ersetzt
- Umlaute und Sonderzeichen werden normalisiert
- Titel werden auf eine Länge von 80 Zeichen gekürzt
- Einleitende Artikel werden nicht berücksichtigt
- „und“, „and“ und „&“ werden auf „u“ gekürzt
- Abkürzungsliste für häufige Abkürzungen in Titelzusätzen
- Sehr kurze und unspezifische Titel werden nicht berücksichtigt (z.B. „Werke“, „Briefe“, „Sinfonien“)
- Bei Namen werden zweite und weitere Vornamen nicht berücksichtigt

Die für einen Datensatz erzeugten Schlüssel können somit beispielsweise folgendermaßen aussehen:

17 MARC Code List for Relators, LoC, <https://www.loc.gov/marc/relators/relaterm.html>, Stand: 23.11.2018.

18 In LRM wird, wie in Kap. 2.2 beschrieben, deutlich darauf verwiesen, dass Übertragungen in andere Gattungen, die häufig über verschiedene Publikationstypen deutlich werden, neue Werke darstellen. Deshalb wird die Medienform, spezifiziert in MARC Leader Position 06, ggf. ergänzt durch die Angaben in Leader Position 07, in zum Teil zusammengefassten Publikationstypen dem Schlüssel hinzugefügt. Dadurch wird eine gemeinsame Bündelung von textuellen und in andere Medienformen übertragenen Versionen eines Werks verhindert.


```
<record>
<isbnVolumeTitle>9781138026131-X-therootsoffootballhooliganism-book
</isbnVolumeTitle>
<titleCreator>rootsoffootballhooliganismhistoricalusociologicalstudy-
dunningeric-X-book</titleCreator>
<titleCreatorAddedEntry>therootsoffootballhooliganismhistoricalusocio
logicalstudy-murphypatrick-X-book</titleCreatorAddedEntry>
</record>
```

Der erste Schlüssel in diesem Beispiel wird aus der ISBN der Publikation, der Zählung eines Teils/einer Abteilung eines Werkes, die bei fehlendem Eintrag durch ein „X“ ersetzt wird, dem Haupttitel und dem Publikationstyp erstellt. Zwei weitere Schlüssel kombinieren den Haupttitel und Titelzusatz der Publikation mit zwei der drei in den MARC-Feldern 100 und 700 genannten Autoren und dem Publikationstyp. Bei nicht vorhandener Zählung eines Teils/einer Abteilung eines Werkes wird wiederum ein „X“ eingesetzt.

2.3.3. Ziel der Bündelung

Die Bündelung von Werken ermöglicht die Übertragung qualitativ hochwertiger und intellektuell erstellter inhaltserschließender Merkmale wie Klassifikation und Schlagwörtern von Mitgliedern eines Bündels auf alle anderen als inhaltsgleiche Titel erkannte Mitglieder. Das in Abbildung 2 dargestellte Beispiel illustriert deutlich, dass das Verfahren zur Bündelung von Werken robust gegenüber einer heterogenen Erfassungspraxis und unterschiedlichen Metadatenangaben der verschiedenen Ausgaben und Auflagen eines Werkes sein muss. Hier werden Datensätze trotz orthographischer Unterschiede im Titel, Abkürzungen im Titelzusatz und unterschiedlicher Erscheinungsjahre gebündelt.

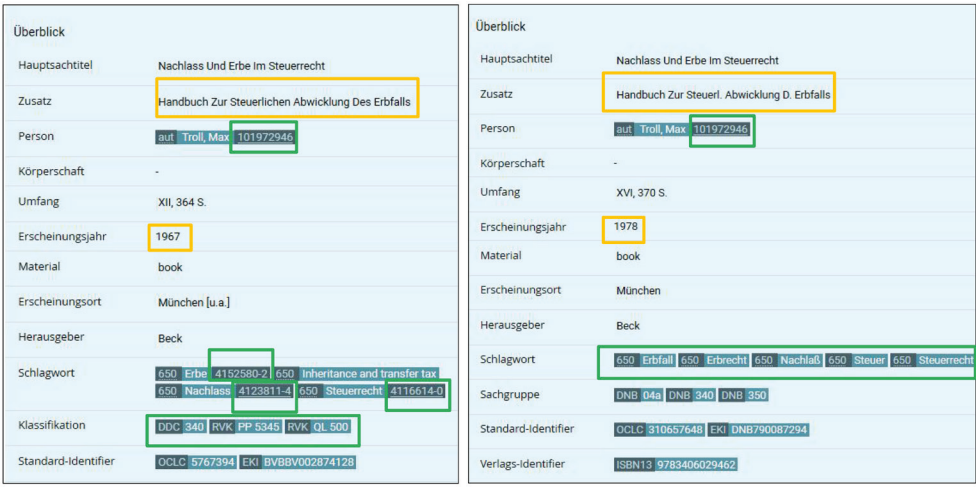

Abb. 2: Beispiele in einem Werkbündel zusammengefasster Datensätze

Eine Übertragung inhaltserschließender Merkmale bietet sich hier im Bereich der Schlagwörter und der Klassifikation(en), aber auch der Normdatenanreicherung in Form einer Ergänzung durch Identifier der Gemeinsamen Normdatei (GND) für Personen und Schlagwörter an.

Die Werkbündelung wird zurzeit durch Mitglieder der Bibliotheksverbünde und der Deutschen Nationalbibliothek evaluiert. Nach weiteren Überarbeitungen sollen Bündel mit inhaltserschließenden Merkmalen, den Notationen der Regensburger Verbundklassifikation (RVK), den Notationen der Dewey Decimal Classification (DDC) und Schlagwörtern in einem Austauschformat bereitgestellt werden.

3. Einbinden externer Informationsquellen

Neben der Nutzung von bibliothekarischen Titeldaten zur Übertragung inhaltserschließender Merkmale kann auch die Verwendung nicht-bibliothekarischer Informationsquellen ein Zugewinn für die Erschließungstiefe sein. Dies wurde im Rahmen des Projektes ORCID DE[19] mit den öffentlich zugänglichen Daten des Dienstes Open Researcher and Contributor ID (ORCID)[20] exemplarisch durchgeführt. ORCID-Records werden von Wissenschaftlerinnen und Wissenschaftlern selbst angelegt und enthalten eine eindeutige ORCID sowie den Namen der Person. Es können unter anderem weitere Namensvarianten, Affiliationen, externe Identifier anderer Organisationen und Publikationen eingetragen werden. Dieser Datenbestand wurde genutzt, um in ORCID eingetragene Personen mit solchen, die einen Personendatensatz in der Gemeinsamen Normdatei (GND) besitzen, abzugleichen. Dies wurde unter Zuhilfenahme der in Culturegraph verzeichneten Titeldatensätze in mehreren Schritten vorgenommen.

Zunächst werden die ORCID-Records, die in einem XML-Format vorliegen, und die verwendeten Culturegraph-Titeldaten im MARC-Format in ein einheitliches Zwischenformat überführt. Um die in den Titeldatensätzen verknüpften GND-Personendatensätze zu extrahieren, wird eine Liste der ID-Nummern dieser Sätze erstellt. Von beiden Datenbeständen werden in einem ähnlichen Verfahren wie zur Werkbündelung Matchkeys erstellt, die den Nachnamen und ersten Vornamen der an der Schaffung eines Werks beteiligten Akteure sowie Titel und Titelzusatz der Publikationen enthalten. Die erstellten Schlüssel werden abgeglichen und Personen mit gleichem Namen und Publikationstiteln in einer Liste mit korrespondierenden ORCIDs und GND-IDs zusammengeführt.

19 ORCID DE, Förderung der Open Researcher and Contributor ID in Deutschland, <https://www.orcid-de.org/>, Stand: 23.11.2018

20 ORCID Public Data File 2017, <https://orcid.org/content/download-file>, Stand: 23.11.2018. Anzahl verarbeiteter Records: 3.919.340.

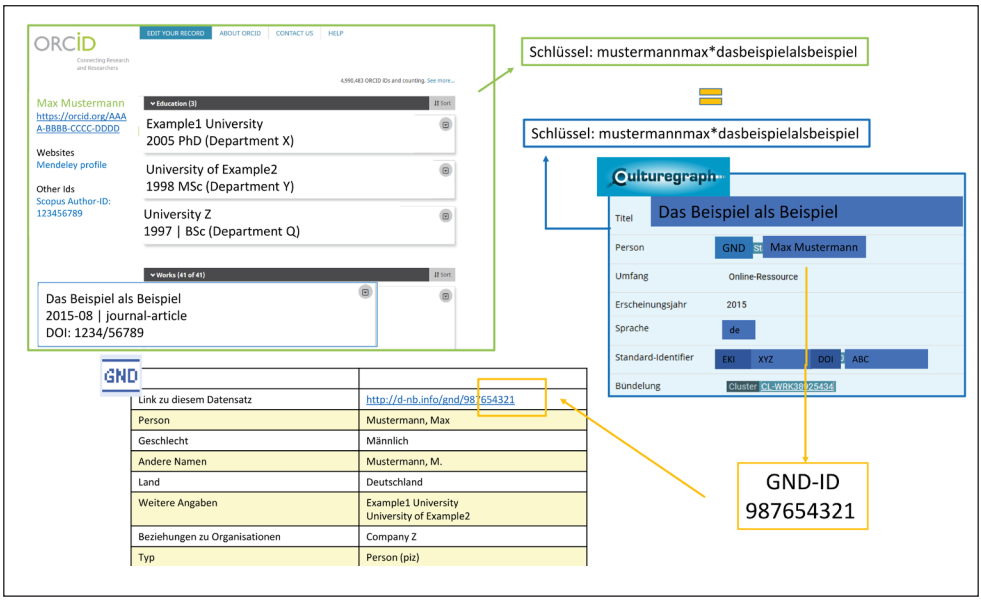

Abb. 3: Beispiel des Abgleichs ORCID - GND

Abbildung 3 verdeutlicht den Abgleichprozess anhand eines nachgestellten Beispiels. Aus einem ORCID-Record wird ein Schlüssel erzeugt, der den Namen oder angegebene Namensvarianten des/r Record-Inhaber/s/in mit den dort eingetragenen Publikationstiteln verbindet. Auch von Culturegraph-Datensätzen werden aus den Namen der an der Schaffung des Werks beteiligten Akteur/e/innen und den Titelangaben Schlüssel erzeugt. Stimmen zwei Schlüssel überein und der Culturegraph- Datensatz enthält eine Referenz auf einen GND-Personendatensatz, handelt es sich um ein relevantes Match. Die GND-ID (im Beispiel 987654321) und die ORCID (im Beispiel AAAA-BBBB-CCCC-DDDD) werden so als Bündel identifiziert und in der Treffermenge ausgewiesen.

Der Abgleich hat bislang ca. 377.000 Schlüsselpaare identifiziert, die jeweils mindestens einen Eintrag aus ORCID mit einem Eintrag aus Culturegraph verbinden. Diese beziehen sich auf ca. 131.000 ORCID-Records, da pro Record mehrere Schlüssel, einer für jede Publikation im Record, gebildet werden können. In den Culturegraph-Titeldatensätzen werden nur solche berücksichtigt, die eine Referenz der/s an der Schaffung des Werks beteiligten Akteur/in/s auf einen GND-Personendatensatz enthalten. Unter dieser Voraussetzung können ca. 120.000 Schlüsselpaare identifiziert werden. Jeder Treffer bezieht sich auf eine Publikation einer bestimmten Person, so dass mehrere Treffer pro GND-Eintrag bzw. ORCID Record auftreten können, wenn mehrere Publikationen für ein Match sorgen. Letztlich wurde eine Menge von mehr als 25.000 korrespondierenden ORCID-Records und GND-Personendatensätzen gefunden.

Im Rahmen des Abgleichs können auch weitere relevante Erkenntnisse erlangt werden. Beispielsweise wenn der Abgleich ein Aufeinandertreffen von mehreren GND-Einträgen pro ORCID-Record

oder mehrere ORCID-Records pro GND-Eintrag ergibt. Dies kann einerseits auf eine Verknüpfung einer Publikation mit einem falschen GND-Eintrag hindeuten. Andererseits werden so auch doppelte GND-Einträge oder ORCID-Records, die mit gleichen Publikationen verknüpft sind, entdeckt.

Ein direkter Mehrwert dieses Abgleichs ist die Möglichkeit, den GND-Personendatensatz um die ORCID anzureichern. Weitere noch zu prüfende Potentiale liegen in einer häufig umfangreicheren Beschreibung des professionellen Hintergrundes im ORCID-Record. Beispielsweise sind hier oft mehrere Affiliationen mit Wirkungsdaten und weitere externe Identifier wie verlagsspezifische Identifikationsnummern von Autor/inn/en enthalten, die auch in der GND Verwendung finden könnten.

4. Verbundübergreifende statistische Auswertungen

Die in Culturegraph versammelten Metadaten der Bibliotheksverbünde und der DNB bilden über die bereits dargestellten Anwendungsfälle hinaus die Möglichkeit, einen umfassenden Blick über verschiedene Fragestellungen der Erschließung und des Bibliotheksbestands im deutschsprachigen Raum zu gewinnen. Statistiken zu verschiedenen in den Metadaten dokumentierten Sachverhalten können interessante Erkenntnisse liefern. Exemplarisch sollen hier drei Bereiche beleuchtet werden, hauptsächlich mit dem Ziel, darzustellen, welche Art von Abfragen in diesem Datenbestand lohnenswert sein können. Eine tiefergehende Analyse der Hintergründe der einzelnen dargestellten Auswertungen ist leider in diesem Rahmen nicht möglich.

4.1. Nutzung von Klassifikationssystemen

Um die Verwendung zweier Klassifikationssysteme, der Regensburger Verbundklassifikation (RVK) und der Dewey Decimal Classification (DDC) gegenüberzustellen, werden Statistiken der am häufigsten auftretenden Klassen erstellt.

Für die Statistik zu DDC-Notationen werden die MARC-Felder 082 $a, 083 $a und 085 $a sowie 084 $a, wenn im Unterfeld $2 DDC angegeben ist, verwendet. RVK-Notationen werden aus MARC-Feld 084 $a ausgelesen, wenn in $2 RVK angegeben ist. Der Bestand umfasst insgesamt 34.648.819 DDC Notationen und 30.079.854 RVK-Notationen. Bei der Erstellung der Statistik werden lediglich die ersten drei Ziffern der DDC-Klassifikation bzw. die ersten beiden Buchstaben der RVK-Klassifikation verwendet. Damit kann nicht zwischen den DDC-Sachgruppen, die häufig auf der 2. Ebene der DDC liegen, und vollständigen DDC-Notationen unterschieden werden.

Es zeigt sich, dass die 10 am häufigsten auftretenden ersten drei Stellen der vergebenen DDC-Notationen mehrere Fachgebiete abdecken (vgl. Tabelle 2). Während weit verbreitete und umfangreiche Fachgebiete wie Wirtschaft und Medizin auch in den DDC-Notationen über alle Datenquellen stark vertreten sind, lassen sich die großen Anzahlen der deutschen und englischen Literatur auf in den jeweiligen Nationalbibliotheken häufig vergebene Notationen zurückführen. Beispielsweise ist ein Grund für das häufige Auftreten der Notation 830 „Deutsche Literatur“ die von der DNB generell für deutschsprachige Belletristik vergebene Sachgruppe 830.

Tabelle 2: Top 10 der vergebenen DDC-Notationen

DDC	Bezeichnung	Anzahl
330	Wirtschaft	979.659
610	Medizin und Gesundheit	837.101
830	Deutsche Literatur	699.680
370	Bildung und Erziehung	654.768
823	Englische, altenglische Literaturen: Englische Erzählprosa	639.908
340	Recht	611.185
320	Politikwissenschaft	587.923
658	Management, Öffentlichkeitsarbeit: Allgemeines Management	558.310
050	Zeitschriften, fortlaufende Sammelwerke	546.025
616	Medizin und Gesundheit: Krankheiten	474.862

Tabelle 3 zeigt analog die am häufigsten in den Metadaten verzeichneten RVK-Klassen. Hier sind einige weitere Fachgebiete zu finden, die in den häufigen DDC-Klassen eine untergeordnete Rolle spielen. An erster Stelle stehen Spezielle Soziologien. Weiterhin spielen die Fächer Geschichte, Kunstgeschichte, Mathematik und Informatik eine größere Rolle, während Recht, Medizin und Bildung und Erziehung, die in den DDC-Klassen prominent vertreten sind, hier nicht in den häufigsten 10 Klassen zu finden sind. Es ist auch zu bemerken, dass die genannten RVK-Notationen gleichmäßig stark in allen Verbünden vertreten sind, während bei der Verteilung von DDC-Klassen häufig verbundspezifische Schwerpunkte existieren.

Tabelle 3: Top 10 der vergebenen RVK-Notationen

RVK	Bezeichnung	Anzahl
MS	Spezielle Soziologien	779.022
QP	Allgemeine Betriebswirtschaftslehre	717.255
AP	Medien- und Kommunikationswissenschaften, Kommunikationsdesign	581.574
ST	Informatik: Monographien	572.601
LH	Allgemeine Kunstgeschichte	509.913
NQ	Geschichte seit 1918	463.316
LI	Kunstgeschichte: Künstler-Monographien	408.128

CC	Systematische Philosophie	391.153
SK	Mathematik: Monographien	361.065
MG	Politische Systeme einzelner Länder: Europa, Nordamerika	357.581

Gründe für die zum Teil unterschiedlichen Schwerpunkte in der klassifikatorischen Inhaltserschließung der untersuchten Daten können sowohl in der Struktur des jeweiligen Klassifikationssystems als auch der inhaltlichen Ausrichtung der Bibliotheken, die ein Klassifikationssystem nutzen, liegen.

4.2. Zeitreihenauswertungen

Ein weiter Anwendungsfall, für den der Datenbestand in Culturegraph genutzt werden kann, ist eine Übersicht über die zeitliche Entwicklung der Publikationskultur oder der bibliothekarischen Praxis. Ein Beispiel hierfür ist die in Abbildung 4 verdeutlichte Entwicklung der Anzahl von Monographien in gedruckter Form und online seit 1950.

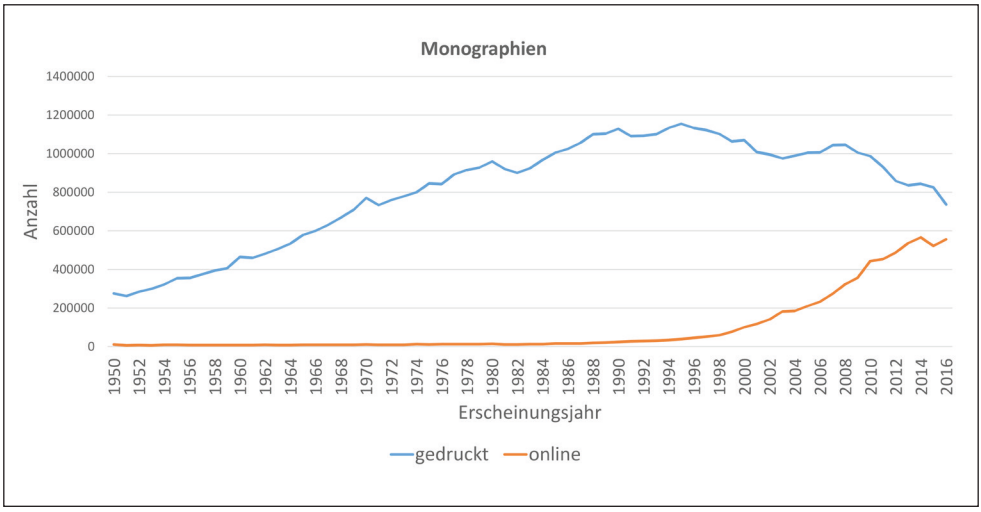

Abb. 4: Anzahl Monographien gedruckt und online im Zeitraum 1950-2016

Es zeigt sich, dass im untersuchten Bestand die Anzahl der gedruckten und onlinebasierten Monographien insgesamt stetig steigt. Seit Beginn des Jahrtausends nimmt die Zahl an onlinebasierten Monographien deutlich zu und geht auch mit einem leichten Rückgang gedruckter Monographien einher.

4.3. Nutzung von Normdatenverknüpfungen

Die Culturegraph-Daten bieten außerdem die Möglichkeit, die in den Titeldaten gespeicherten Normdatenverknüpfungen auszuwerten. Dies wurde hier exemplarisch für die Verknüpfungen zu Personennormdaten durchgeführt. Wie Tabelle 4 zeigt, ist unter den häufig verknüpften Personen eine große Anzahl an Komponisten, die vermutlich aufgrund einer Vielzahl von Werken und verschiedenen Reproduktionen dieser Werke häufig in bibliothekarischen Metadaten auftauchen.

Tabelle 4: Top 10 der Verknüpfungen von Personennormdaten

Name	GND-ID	Anzahl der Verknüpfungen
Johann Sebastian Bach	11850553X	369.890
Wolfgang Amadeus Mozart	118584596	357.772
Johann Wolfgang von Goethe	118540238	271.176
Ludwig van Beethoven	118508288	235.494
Franz Schubert	118610961	189.720
William Shakespeare	118613723	178.556
Martin Luther	118575449	170.352
Joseph Haydn	118547356	167.128
Johannes Brahms	118514253	135.648
Marcus Tullius Cicero	118520814	134.330

Weitere Autoren, die ebenfalls eine hohe Reproduktionszahl ihrer Werke über die Jahre aufweisen, wie Johann Wolfgang von Goethe, William Shakespeare und Martin Luther, sind außerdem unter den 10 am häufigsten verknüpften Personendatensätzen zu finden.

5. Fazit

Die Aggregation von Titeldaten aus verschiedenen Bibliotheksdatenbeständen eröffnet vielfältige Möglichkeiten der Analyse und Auswertung. Im Vordergrund steht hierbei die Nachnutzung intellektuell erstellter Bestandteile der Metadaten, die sowohl eine Rationalisierung der Arbeitsprozesse als auch eine umfassendere Erfassung und Erschließung einer größeren Zahl von Titeldaten ermöglicht. Parallel zu Verfahren der automatischen Erschließung können so automatisierte Verfahren verwendet werden, um intellektuell erstellte Information weitergehend zu nutzen. Die Zahl und der Umfang inhaltserschließender Information können so verbessert werden. Darüber hinaus kann eine größere Standardisierung durch eine Übertragung von Normdatenverknüpfungen erreicht werden.

Auch externe Datenquellen bieten in wachsendem Ausmaß Schnittstellen und standardisierte Informationen, die als Anreicherung bibliographischer Daten dienen und durch den Abgleich mit aggregierten Bibliotheksdatenbeständen nutzbar gemacht werden können.

Die vorgestellten Verfahren werden in der nächsten Zeit noch präzisiert und an verschiedene Bedarfe angepasst. Weitere Vorhaben, die unter Verwendung der in Culturegraph gespeicherten Daten geplant sind, umfassen Bestrebungen zur Deduplizierung von Personennormdaten und die Konsolidierung der Werkbündelung durch Entwicklung einer Metrik zur Beurteilung der Bündel.

Literaturverzeichnis

- Gatenby, Janifer; Greene, Richard O.; Oskins, W. Michael u.a.: GLIMIR: Manifestation and Content Clustering within WorldCat, in: code{4}lib Journal 17 (2012), <http://journal.code4lib.org/articles/6812>, Stand: 23.11.2018.

- Geipel, Markus Michael; Böhme, Christophe; Hannemann, Jan: Metamorph: A Transformation Language for Semi-Structured Data, in: D-Lib Magazine 21 (5/6), 2015, <https://doi.org/10.1045/may2015-boehme>.

- Hickey, Thomas B.; Toves, Jenny: FRBR Work-Set Algorithm. Version 2.0, 2009, <https://www.oclc.org/content/dam/research/activities/frbralgorithm/2009-08.pdf>, Stand: 23.11.2018.

- IFLA: Functional Requirements for Bibliographic Records, Final Report, 1998, <https://www.ifla.org/files/assets/cataloguing/frbr/frbr_2008.pdf>, Stand: 23.11.2018.

- Pfeffer, Magnus: Using Clustering Across Union Catalogues to Enrich with Indexing Information, in: Spiliopoulou, Myra; Schmidt-Thieme, Lars; Janning, Ruth (Hg): Data Analysis, Machine Learning and Knowledge discovery, Cham 2014, S. 437–445.

- Pfeifer, Barbara; Polak-Bennemann, Renate: Zusammenführen was zusammengehört – Intellektuelle und automatische Erfassung von Werken nach RDA, in: o-bib. Das offene Bibliotheksjournal 3 (4), 2016, S. 144–155, <https://doi.org/10.5282/o-bib/2016h4s144-155>.

- Riva, Pat; Le Bœuf, Patrick; Žumer, Maja: IFLA Library Reference Model. A Conceptual Model for Bibliographic Information, 2017, <https://www.ifla.org/files/assets/cataloguing/frbr-lrm/ifla-lrm-august-2017_rev201712.pdf>, Stand: 23.11.2018.

- Wiesenmüller, Heidrun; Pfeffer, Magnus: Abgleichen, anreichern, verknüpfen, in: BuB 35 (9), 2013, S. 625–629. Online:<http://www.b-u-b.de/pdfarchiv/Heft-BuB_09_2013.pdf>, Stand: 23.11.2018.

