Zusammenführen was zusammengehört

Intellektuelle und automatische Erfassung von Werken nach RDA

Barbara Pfeifer, Deutsche Nationalbibliothek
Renate Polak-Bennemann, Deutsche Nationalbibliothek

Zusammenfassung:
Der Beitrag will die Erschließungspraxis der Deutschen Nationalbibliothek (DNB) im ersten Implementierungsschritt der RDA zur Angabe der Werkebene darstellen, Erfahrungen zur grundsätzlichen Vorgehensweise und zu Sonderregelungen vermitteln und einen Ausblick in die Zukunft bieten. Die Angabe der Werkebene als Kernelement des neuen Standards wird bei der Erschließung von Ressourcen in der DNB immer berücksichtigt. Bei der intellektuellen Erschließung von Monografien wird das Element als Normdatensatz oder als textuelle Angabe im Katalogisat erfasst. Allerdings setzt die derzeitige Praxis auch auf zukünftige automatische Clusterverfahren. Der erreichte Stand zu Algorithmen und Testläufen für das Werkclustering soll ebenso aufgezeigt werden wie die daraus resultierenden Entscheidungen und die weitere Perspektive. Von zentraler Bedeutung ist die Frage, ob und unter welchen Bedingungen auf der Basis der Werkcluster Normdatensätze für die Gemeinsame Normdatei (GND) generiert werden können, um sie im deutschen Sprachraum kooperativ zu nutzen und sie in der Linked-Data-Cloud zu vernetzen.

Summary:
This article explains how works are catalogued at the Deutsche Nationalbibliothek (German National Library) in the first implementation step of RDA. It discusses experiences with the general approach as well as some special rules and points out future prospects. When cataloguing resources at the Deutsche Nationalbibliothek, the work entity is always included as the work elements are core elements of the new standard. If the cataloguing is done intellectually, the work entity is described either as a link to an authority record or as a textual string within the description of the resource. However, the current practice builds also on automatic clustering of works in the future. The state of the art of automatic cluster algorithms and their evaluation and testing is described as well as the next steps for the future. An essential question is whether and under which conditions authority records for works can be generated on the basis of automatic clustering in order to use them cooperatively and to position them as crosspoints in the linked data cloud.

Zitierfähiger Link (DOI): http://dx.doi.org/10.5282/o-bib/2016H4S144-155
Autorenidentifikation: Pfeifer, Barbara: GND 139338349; ORCID: http://orcid.org/0000-0002-1751-9808; Polak-Bennemann, Renate: GND 1115381415; ORCID: http://orcid.org/0000-0001-9963-9397
Schlagwörter: RDA; Resource Description and Access; Werk; Cluster; Clusterverfahren

1. Grundlagen der Katalogisierung von Werken

1.1. Werkebene nach RDA

Mit der Implementierung des neuen Regelwerks Resource Description and Access (RDA) im deutschsprachigen Raum (D-A-CH) im Herbst 2015 werden Ressourcen gemäß der dem neuen Regelwerk zu Grunde liegenden Functional Requirements for Bibliographic Records (FRBR) erschlossen. Entitäten des FRBR-Modells in Gruppe 1 sind Werk, Expression, Manifestation und Exemplar. In diesem Beitrag soll auf die Bedeutung und den Umgang mit der Entität Werk und der Behandlung der Werkebene bei der Katalogisierung sowie auf die Möglichkeiten automatischer Verfahren zur Verbesserung von Katalogrecherchen eingegangen werden. Das Werk ist gemäß RDA eine individuelle, intellektuelle oder künstlerische Schöpfung und bildet somit die Grundlage für die Entstehung aller Ressourcen. Die Werkebene wird im Katalogisierungsprozess der Formalerschließung mit RDA immer betrachtet. Diese Herangehensweise stellt eine wesentliche Neuerung bei der Katalogisierung dar und fordert in vielen Fällen eine andere Denkweise der Katalogisierenden.

Der bevorzugte Titel des Werks und der für das Werk hauptverantwortliche geistige Schöpfer sind Kernelemente in RDA und damit Bestandteile des im D-A-CH-Raum festgelegten Standardelemente-Sets – die Angabe der Werkebene ist damit immer obligatorisch. Aus Sicht des FRBR-Modells ist dies logisch und konsequent. Ziel der RDA ist es, die Anforderungen des Nutzers „Ressourcen zu finden", „sie zu identifizieren und auszuwählen" und „Zugang zu ihnen zu erhalten" zu unterstützen. Dabei ist die Angabe der Werkebene als intellektuelle Grundlage einer jeden Ressource ein wichtiger Aspekt bei der Recherche in nationalen und internationalen Bibliothekskatalogen. Ist die Werkebene in jedem Katalogisat verzeichnet, können über diese Angabe verschiedene Expressionen, d. h. verschiedene Sprachausgaben oder Ausgaben in einer anderen Form, z. B. als gesprochenes Wort, aber auch unterschiedliche Manifestationen, d. h. Ausgaben in verschiedenen Verlagen oder Ausgaben unterschiedlicher Ausgestaltung, im Katalog zusammengeführt werden.

RDA unterscheidet Werke eines oder mehrerer geistiger Schöpfer sowie Werke von ungesicherter oder unbekannter Herkunft; außerdem gibt es Zusammenstellungen von mehreren Werken. Die Angabe des geistigen Schöpfers – dies kann eine Person, Familie oder Körperschaft sein (Entitäten der FRBR-Gruppe 2) – erfolgt als Beziehung zum Werk. Mit dem erstgenannten oder hauptverantwortlichen geistigen Schöpfer und dem bevorzugten Titel des Werks (dies ist meist der Titel der Originalausgabe) wird der normierte Sucheinstieg für das Werk gebildet. Für Werke von ungesicherter oder unbekannter Herkunft sowie für Zusammenstellungen von Werken verschiedener Personen, Familien oder Körperschaften wird der normierte Sucheinstieg nur mit dem bevorzugten Titel des Werks gebildet. Normierte Sucheinstiege, die ansonsten identisch wären, werden durch weitere identifizierende Merkmale wie „Form des Werks", „Datum des Werks", „Ursprungsort des Werks" oder auch die Angabe einer „Sonstigen unterscheidenden Eigenschaft des Werks" ergänzt.

Das folgende Beispiel 1 zeigt einen einfachen Fall – eine Ressource, die die tschechische Übersetzung eines Werks von Giulia Enders beinhaltet. In der ersten Spalte steht jeweils die Nummer des RDA-Elements, in der zweiten dessen Name und in der dritten der erfasste Inhalt. Wie man sieht, ist der auf dem Titelblatt stehende Haupttitel auf Tschechisch. Der bevorzugte Titel des Werks ist hingegen

der deutsche Originaltitel. Im Element 17.8 wird das in der Manifestation verkörperte Werk in der Form des normierten Sucheinstiegs für das Werk ausgedrückt. Dieser besteht aus dem normierten Sucheinstieg für die Verfasserin und dem bevorzugten Titel des Werks.

Beispiel 1

RDA	Element	Erfassung
2.3.2	Haupttitel	Střevo není tabu
6.2.2	Bevorzugter Titel des Werks	Darm mit Charme
17.8	In der Manifestation verkörpertes Werk	Enders, Giulia, 1970-. Darm mit Charme
19.2	Geistiger Schöpfer	Enders, Giulia, 1970-
18.5	Beziehungskennzeichnung	Verfasser

1.2. Normdatensätze für Werke

Die Gemeinsame Normdatei (GND) enthielt von Beginn an Werknormdatensätze (Satzart „Tu“). Diese wurden vor dem Umstieg auf RDA nur für Werke der Musik in der Formalerschließung und für Werke, die als Schlagwort in der Inhaltserschließung benötigt wurden, genutzt. Seit der im Herbst 2015 erfolgten RDA-Implementierung ist die GND für alle Werke in der Formalerschließung nutzbar, d. h. wenn gewünscht, kann der Titeldatensatz mit dem jeweiligen Normdatensatz verknüpft werden. Die Entscheidung, ob die GND in dieser Weise in der Formalerschließung genutzt werden soll, erfolgt in der jeweiligen Institution. Neu ist außerdem die einheitliche Erfassung von Werken in der Formal- und Inhaltserschließung gemäß den RDA-Regeln. Die GND enthält momentan 252.564 Werknormsätze (Stand Februar 2016). Der monatliche Zuwachs beträgt seit der RDA-Einführung monatlich ca. 2500 Datensätze, so dass in den kommenden Jahren ein enormer Zuwachs an Tu-Sätzen zu erwarten ist.

1.3. Wie wird die Werkebene im D-A-CH-Raum angegeben?

Gemäß den Festlegungen der AG RDA in den Anwendungsrichtlinien für den deutschen Sprachraum (D-A-CH) wird der bevorzugte Titel des Werks nicht gesondert im Titeldatensatz angegeben, wenn die Angabe identisch mit der Angabe des Haupttitels der Manifestation wäre (D-A-CH-AWR zu RDA 6.2.2.8). Dies wäre z. B. bei einer deutschen Ausgabe des Werks aus dem obigen Beispiel der Fall: Der Haupttitel der Manifestation wäre „Darm mit Charme“ und damit identisch zum bevorzugten Titel des Werks. Die Angabe erfolgt in einem solchen Fall nur implizit: Aus der Tatsache, dass im entsprechenden Feld kein bevorzugter Werktitel erfasst wurde, kann man schließen, dass dieser mit dem (in einem anderen Feld erfassten) Titel der Manifestation übereinstimmt. Wenn hingegen der bevorzugte Werktitel nicht identisch mit dem Titel der Manifestation ist, muss er explizit im Katalogisat angegeben werden. Dafür gibt es zwei Möglichkeiten: Es kann entweder – falls die Institution mit der GND arbeitet – ein GND-Normdatensatz verknüpft werden oder alternativ die Erfassung des bevorzugten Werktitels (und ggf. eines oder mehrerer unterscheidender Merkmale) als Text erfolgen.

Die DNB folgt den D-A-CH-Empfehlungen. Allerdings werden in jedem Fall vorhandene GND-Normdatensätze für Werke in der Formalerschließung mit dem Katalogisat verknüpft, wenn die Suche in der GND erfolgreich ist. Ist noch kein Normdatensatz für das Werk vorhanden, so ist es den Katalogisierenden freigestellt, ob sie einen neuen Werknormdatensatz anlegen oder die Angaben zum Werk als Volltext im Katalogisat angeben. Eine Angabe nur als Text im Pica-Feld 3210 zeigt Beispiel 2; eine Verknüpfung mit einem Werknormdatensatz der GND (im selben Feld) zeigt Beispiel 3.

Beispiel 2

3210 Steuertipps für Künstler 4000 Steuertipps für Künstlerinnen und Künstler / Bayerisches Staatsministerium der Finanzen, für Landesentwicklung und Heimat

Anmerkung: Datensatzauszug aus dem DNB-Datenbestand (im PICA-Iltis-Format), Angabe des bevorzugten Titel des Werks als Text

Beispiel 3

3000 !1020814381!Estep, Jennifer$BVerfasser$4aut 3210 !1084585189!Estep, Jennifer$aKiller frost 4000 Frostkiller / Jennifer Estep ; aus dem Amerikanischen von Vanessa Lamatsch

Anmerkung: Datensatzauszug aus dem DNB-Datenbestand (im PICA-Iltis-Format), Verknüpfung zum Werknormdatensatz der GND in Feld 3210

Diese Festlegungen zum Umgang mit der Werkebene wurden im Hinblick auf die bisher entwickelten Verfahren zum maschinellen Clustern von Werken getroffen, die im zweiten Teil dieses Beitrags ausgeführt werden.

1.4. Datenanalyse der seit der RDA-Einführung angelegten Titeldaten

Eine Datenanalyse im Titelbestand der DNB (Stand Februar 2016) zeigt, dass bei ca. 15 % (6.528 Titeldatensätze) der nach RDA erschlossenen Monografien (ohne Musikalien und Tonträger) der bevorzugte Titel des Werks explizit besetzt wurde. Die Analyse zeigt, dass nur bei einem Anteil von einem Drittel ein Werknormdatensatz aus der GND verknüpft wurde; bei einem Anteil von zwei Dritteln wurde die Werkebene als textuelle Angabe im Katalogisierungsformat erfasst. Die Ergebnisse der Analyse sind in Abb. 1 dargestellt.

Die Stichprobe zeigt, dass bei ca. 85 % der Ressourcen der Haupttitel der Manifestation identisch mit dem bevorzugten Titel des Werks ist und somit keine explizite Angabe des Werktitels nötig ist (d.h. der Haupttitel der Manifestation fungiert gleichzeitig als bevorzugter Titel des Werks). Um diese Ergebnisse zu verifizieren, sind zukünftig weitere Stichproben notwendig, da die vorgestellte Datenanalyse nach nur 5-monatiger RDA-Praxis in der DNB erfolgte.

Interessant ist auch die Analyse der im Rahmen der Formalerschließung neu angelegten oder nachgenutzten GND-Werknormsätze. Von 2.332 genutzten Normdatensätzen waren 1.937 neu angelegte

Normdatensätze aus der Formalerschließung; die übrigen Verknüpfungen erfolgten mit schon vorhandenen Werknormsätzen in der GND, die aus den Teilbeständen Sacherschließung und Musik stammen. Die Datenanalyse lässt also erwarten, dass die GND in Zukunft eine starke Anreicherung um Werknormsätze aus dem Bereich der Formalerschließung erfahren wird.

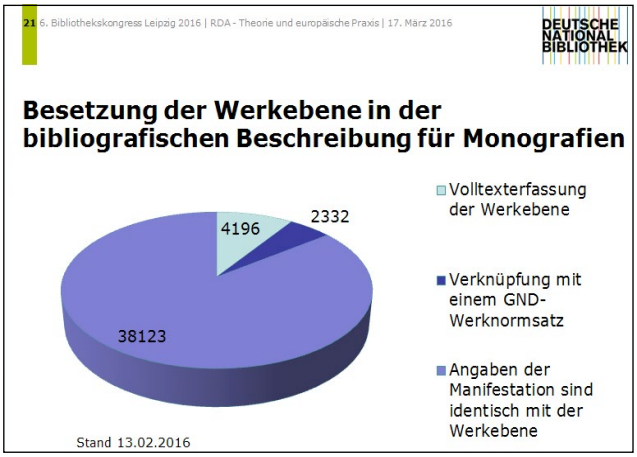

Abb. 1: Ergebnisse der Datenanalyse

1.5. Bestimmung der Werkangaben im Katalogisierungsprozess

In der DNB erfolgt die Bestimmung der benötigten Angaben für das Werk durch die Katalogisierenden schon nach kurzer Schulungsphase in den meisten Fällen korrekt und ohne besonderen Aufwand. Bestehende GND-Normdatensätze können in der Formalerschließung sehr einfach verknüpft werden. Auch das Anlegen von neuen GND-Normdatensätzen erfolgt mit Hilfe eines Erfassungsskripts schnell und unkompliziert.

Bei einer kleinen Zahl der Ressourcen ist die Bestimmung der Werkangaben allerdings nicht einfach. Zu Beginn der RDA-Implementierung gab es in einigen Fällen Unsicherheit bei der Unterscheidung von Werken eines Verfassers.

Ein Beispiel für einen schwierigen Fall sind zwei Reiseführer desselben Autors über dieselbe Stadt oder dasselbe Gebiet, die auch denselben Titel haben, aber in unterschiedlichen Verlagen oder in verschiedenen Reihen desselben Verlags erschienen sind. Inhaltlich unterscheiden sich diese in der Regel stark. Ein weiteres Beispiel sind Werke über Software-Produkte, bei denen sich jede Neuauflage auf eine andere Version der Software bezieht. In beiden Fällen wurde von den nationalen RDA-Expertinnen und -Experten entschieden, dass jeweils neue Werke vorliegen.

Neben diesen und anderen grundsätzlichen Fragen zur Bestimmung des Werks gibt es auch bei der Bestimmung von identifizierenden Merkmalen zur Unterscheidung von ansonsten gleichnamigen Werken noch Klärungsbedarf. Die im RDA-Kapitel 6 vorgesehenen identifizierenden Merkmale „Form

des Werks", „Datum des Werks" und „Ursprungsort des Werks" sind für aktuelle Publikationen oft schwer bestimmbar. Die Katalogisierenden können hier nur aus den Angaben zur Manifestation – beispielsweise aus der Veröffentlichungsangabe – die Angaben zum Werk herleiten. Einen Ausweg bietet das RDA-Attribut „Sonstige unterscheidende Eigenschaft des Werks", welches frei gewählt werden kann. Aber auch hier kann es schwierig werden, eine geeignete identifizierende Angabe auf Werkebene zu finden. Möglich ist z. B. die Angabe der monografischen Reihe, wenn die Ressource in einer solchen erschienen ist. Hingegen sind Angaben zum Verlag oder auch zum Erscheinungsjahr eigentlich nicht regelwerksgerecht, da sie der Ebene der Manifestation zuzuordnen sind. In der angloamerikanischen Praxis wird damit jedoch sehr liberal umgegangen. An dieser Stelle wird die AG RDA – auch nach Rücksprache im internationalen RDA-Anwenderkreis – nach Lösungen suchen, um die praktische Anwendung möglichst zu vereinfachen. Darüber hinaus soll eine frei zugängliche Beispielsammlung im RDA-Info-Wiki[1] der DNB die Entscheidungsfindung bei schwierigen Fällen für alle Anwender erleichtern und Orientierung bieten.

Das Hinzuziehen und die Ausgestaltung der Besetzung von identifizierenden Merkmalen zum Sucheinstieg liegen jeweils im Ermessen der Katalogisierenden. Die Entscheidung erfolgt momentan auf Grund der eigenen bzw. nationalen Katalogrecherche. Damit sollte zumindest weitgehend sichergestellt sein, dass in den Katalogen des D-A-CH-Raums für alle Manifestationen desselben Werks auch dieselben identifizierenden Merkmale verwendet werden. International können die Ergebnisse an dieser Stelle jedoch differieren und es können unterschiedliche differenzierende Merkmale für ein Werk verwendet werden. Das Ziel, dem Nutzer über die Werkebene Katalogdaten weltweit zusammenzuführen, kann dadurch erschwert werden.

Die gewonnenen Praxiserfahrungen sollten daher national und international diskutiert werden und bei der Weiterentwicklung des Regelwerks einfließen. Bei der gleichzeitigen Anwendung von intellektuellen und maschinellen Erschließungsverfahren ist es wichtig, bei der intellektuellen Arbeit strukturierte und nachnutzbare Ausgangs- und Vergleichsdaten als Grundlage für maschinelle Verfahren zu erzeugen.

2. Einsatz maschineller Verfahren

Das maschinelle Zusammenführen von Daten (Clustern) ist seit dem Beginn der RDA-Entwicklung in der Diskussion und verschiedene Verfahren – schon auf Basis von mit AACR (Anglo-American Cataloguing Rules) und RAK erstellten Daten[2] - wurden entwickelt. Die DNB beschäftigt sich seit 2013 mit der Thematik.

Im Folgenden wird auf die bisher entwickelten maschinellen Clusterverfahren eingegangen.

1 „RDA-Info," DNB, zuletzt geprüft am 26.07.2016, https://wiki.dnb.de/display/RDAINFO/RDA-Info.
2 Aus diesem Grund wird im Text bei der Beschreibung die RAK-Terminologie gebraucht.

2.1. Clusterverfahren

Clusterverfahren allgemein

Automatisierte Clusterverfahren zielen immer darauf ab, Objekte mit ähnlichen Eigenschaften zusammenzufassen. Bezogen auf Werke gemäß RDA bedeutet es, dass die unterschiedlichen Expressionen und Manifestationen eines Werks in einem Cluster zusammengeführt werden. Im RDA-Projekt wurden gleich zu Beginn des Projekts unterschiedliche Verfahren zur automatisierten Erstellung von Werkclustern untersucht. Bei allen Verfahren wird in einem ersten Schritt ein Schlüssel aus einzelnen Elementen der Datensätze, ein sogenannter Matchkey, gebildet, und in einem zweiten Schritt werden die Datensätze mit gleichem Matchkey zu einem Werkcluster zusammengefasst (Abb. 2).

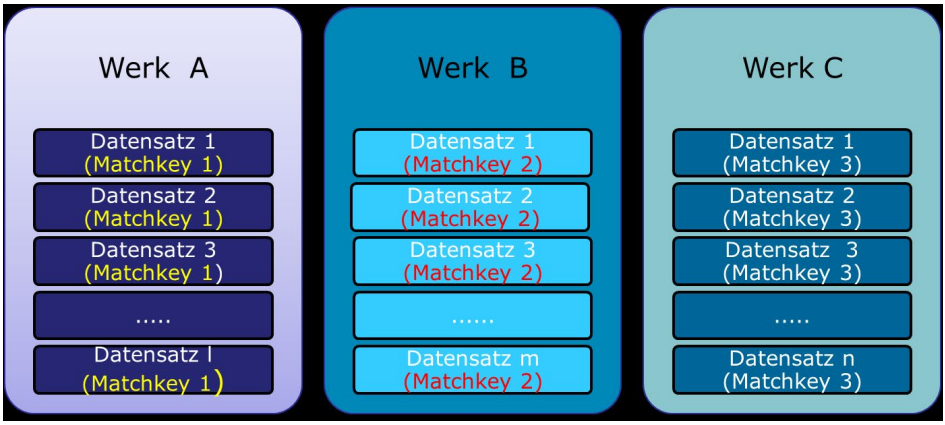

Abb. 2: Grundprinzip der Clustering-Verfahren über Matchkey

Der wesentliche Unterschied der untersuchten Verfahren, die im Folgenden kurz beschrieben werden, liegt in der Bildung des Matchkeys, wobei in der Regel die verfügbaren Datenelemente für den geistigen Schöpfer und den Titel verwendet werden.

OCLC-FRBR-Algorithmus[3]

Bei diesem Algorithmus wird ein Matchkey aus den Verfasser- und Titel-Feldern eines AACR-Datensatzes gebildet. Die einzelnen Elemente dafür werden nach genau definierten Regeln ausgesucht, kombiniert und bearbeitet. So gibt es genaue Anweisungen für Bedingungen und Abfolge der Feldauswahl, das Auslesen und die Bearbeitung des Feldinhalts zur Normalisierung, den Abgleich an Normdaten und den Ausschluss von Titeln, die nicht spezifisch genug sind, wie z. B. „volume“, oder „annual report“.

3 Thomas B. Hickey und Jenny Toves, *FRBR Work-Set Algorithm*, Version 2.0 (Dublin, Ohio: OCLC Online Computer Library Center, Inc. (Research disivion), 2009), zuletzt geprüft am 26.07.2016, http://www.oclc.org/research/activities/past/orprojects/frbralgorithm/2009-08.pdf.

Wenn im Datensatz ein Feld für den Einheitssachtitel vorhanden ist (MARC 130), so gilt der dort vorhandene Titel alleine als ausreichender Schlüssel. Für Datensätze, die nicht genug individualisierende Inhalte haben, um eine eindeutige Zuordnung zu erreichen, wird ebenfalls ein Matchkey erzeugt, allerdings beinhaltet dieser die OCLC-Nummer des Datensatzes. Diese macht ihn eindeutig (unique) und verhindert, dass er falsch zugeordnet wird.

GLIMIR

GLIMIR[4] zielt auf die Anzeige in Portalen und weitet den OCLC-FRBR-Algorithmus aus, um umfangreichere Cluster von Datensätzen erzeugen zu können, die ein Werk repräsentieren. Hierzu werden als Quelle für den Matchkey nicht nur die spezifischen MARC-21-Felder für geistige Schöpfer und Titel genutzt, sondern beispielweise auch Anmerkungen, in denen sich Werktitel sozusagen „verstecken" können. Zudem wird die Normalisierung der Feldinhalte so ausgeweitet, dass typische regelwerkspezifische Begriffe aus verschiedenen Sprachen gleichgesetzt und Abkürzungen aufgelöst werden. Für das Matching sind nicht nur die Ergebnisse „Treffer" und „Nichttreffer" möglich, sondern durch Ähnlichkeitsvergleiche (Trigram) können auch abgestufte Matchwerte gebildet werden. Ferner wird die sogenannte „Friends-of-a-friend-Regel" (siehe unten) angewendet. GLIMIR hält sich allerdings nicht exakt an die FRBR-Definitionen von Werk und Manifestation, so dass es für ein Werkclustering nur bedingt geeignet ist.

Primo

Das Werkclustering des Österreichischen Bibliothekenverbundes in Primo nutzt einen Algorithmus, der dem OCLC-FRBR-Algorithmus ähnelt. Es werden Matchkeys aus verschiedenen Elementen des Katalogisats gebildet, vor allem aus Angaben zu Personen als geistige Schöpfer und Titelangaben, inkl. Unterreihen zu monografischen Reihen und Bandangaben bei mehrbändigen Werken.

Das Clustering der Werke in Primo dient vor allem der Anzeige im Portal und kann von jeder Bibliothek für ihr eigenes Portal konfiguriert werden. Um beispielsweise eine bessere Präsentation von Lehrbüchern zu erreichen, die oft in zahlreichen Auflagen erscheinen, deren Autor oder Herausgeber aber wechselt, wird bei der Generierung des Matchkeys hier statt auf Personen als geistige Schöpfer auf den Verlag über die ISBN-Verlagsnummer zurückgegriffen, womit es sich nicht mehr um ein Werkcluster im FRBR-Sinne handelt.

Pfeffer-Algorithmus

Magnus Pfeffer gleicht für sein Werkclustering[5] auf Grundlage von RAK-Katalogdaten ab:

- Einheitssachtitel, Hauptsachtitel und Zusatz zum Hauptsachtitel
- Personen, Körperschaften als Verfasser, Urheber, sonstige beteiligte Personen bzw. Urheber

4 Janifer Gatenby *et al.* „GLIMIR: Manifestation and Content Clustering within WorldCat," *Code4Lib journal* 17 (2012), zuletzt geprüft am 26.07.2016, http://journal.code4lib.org/articles/6812

5 Magnus Pfeffer, „Using Clustering Across Union Catalogues to Enrich Entries with Indexing Information," in *Data Analysis, Machine Learning and Knowledge Discovery*, hrsg. Myra Spiliopoulou, Lars Schmidt-Thieme und Ruth Janning (Cham: Springer International Publishing, 2014), 437–445, https://dx.doi.org/10.1007/978-3-319-01595-8 und Heidrun Wiesenmüller und Magnus Pfeffer, „Abgleichen, anreichern, verknüpfen: das Clustering-Verfahren – eine neue Möglichkeit für die Analyse und Verbesserung von Katalogdaten," *BuB Forum Bibliothek und Information*, Nr. 9 (2013): 625–629.

Die Datensätze werden bei identischem Einheitssachtitel bzw. Hauptsachtitel und mindestens einer Übereinstimmung bei den Personen/Körperschaften zu einem Werkcluster zusammengeführt.

2.2. Testreihe im RDA-Projekt

Ausgehend von den Analysen der vorhandenen Clusterverfahren wurden schon zu Beginn des RDA-Projekts von der DNB, dem Österreichischen Bibliothekenverbund und dem Literaturarchiv Marbach mehrere Testreihen mit RAK-Daten durchgeführt, bei denen die Ergebnisse iterativ verbessert wurden. Als bestes Verfahren hat sich dabei die Generierung zweier Matchkeys nach unterschiedlichen Algorithmen und deren Kombination über die „Friends-of-a-friend-Regel“ erwiesen. Beim ersten Algorithmus wird der Matchkey aus den Elementen für den Einheitssachtitel und Personen bzw. Urheber gebildet. Beim zweiten Algorithmus wird statt des Einheitssachtitels der Hauptsachtitel verwendet. Die „Friends-of-a-friend-Regel“ sorgt dafür, dass auch Datensätze zusammengeführt werden, die nicht über einen identischen Matchkey verfügen, sofern ein weiterer Datensatz vorhanden ist, bei dem ein Matchkey identisch ist zu einem Matchkey des ersten Datensatzes und der zweite Matchkey zu einem Matchkey des zweiten Datensatzes. Ein Beispiel dafür zeigt Abb. 3.

Da der Datensatz C einen Matchkey besitzt, der identisch zu einem Matchkey von Datensatz A ist, sowie einen Matchkey, der identisch zu einem Matchkey von Datensatz B ist, werden über die „Friends-of-a-friend-Regel“ die Datensätze A und B dem Cluster des Werkes „Dumas, Alexandre: Vingt ans après“ zugeordnet, obwohl sie über keinen identischen Matchkey verfügen.

Abb. 3: Zusammenführen von Datensätzen über die „Friends-of-a-friend-Regel“

2.3. Persistenz der Cluster

Da die Algorithmen zur Matchkey-Bildung und zur Zusammenführung der Datensätze zu einem Cluster in der Regel aufwendig sind, kommt ein Verfahren „on the fly“ beim Suchen, Navigieren oder der Anzeige von Datensätzen nicht in Frage. Stattdessen ist eine Vorberechnung notwendig, deren Ergebnisse auf unterschiedliche Art und Weise aufbewahrt werden können.

Eine Möglichkeit besteht darin, den Matchkey als Feldinhalt im Datensatz abzulegen. Er kann dann in einem weiteren Verfahrensschritt indexiert werden. Dadurch wird ihm eine Liste von Identifikationsnummern von Datensätzen zugeordnet, die denselben Matchkey besitzen. Mit Hilfe der Liste der Identifikationsnummern lassen sich die Datensätze bei Bedarf leicht zu einem Cluster zusammenführen. Da der Matchkey Bestandteil des Datensatzes ist, lässt er sich auch über Datendienste austauschen und kann prinzipiell von anderen Systemen nachgenutzt werden. Die Nachnutzung setzt allerdings voraus, dass im fremden System das gleiche Verfahren implementiert ist.

Eine ähnliche Variante der Persistenz besteht darin, den Matchkey nicht im Datensatz selbst, sondern nur im dazugehörigen Index zu hinterlegen. Ein Austausch über Datendienste ist bei dieser Variante nicht möglich. Auch muss der aufwendige Prozess der Erstellung des Matchkeys zusammen mit der Indexierung erfolgen.

Da die Matchkeys bei komplexen Verfahren umfangreich werden und zum Teil mit mehreren Matchkeys pro Datensatz gearbeitet wird, zudem auch die Clusterverfahren (z.B. durch Ähnlichkeitsvergleiche oder „Friends-of-a-friend-Regel“) aufwendig werden, ist es sinnvoll, nicht die Matchkeys an sich, sondern schon das Ergebnis des Clustering im Datensatz festzuhalten und alle Datensätze, die zu einem Cluster gehören, mit einer Cluster-Identifikationsnummer zu versehen. Dieses Verfahren wird bei GLIMIR (siehe oben) verwendet.

2.4. Generierung von Werknormdatensätzen

Der logisch nächste Schritt ist die Generierung von Werksätzen und die Hinterlegung der Werk-Identifikationsnummer in den Datensätzen, die das Werkcluster bilden. Der Inhalt des Werksatzes ergibt sich aus den nach RDA vorgesehenen Datenelementen für das Werk, die aus den geclusterten Datensätzen abgeleitet werden.

Es liegt nahe, die automatisch generierten Werksätze, sowie die intellektuell erstellten Werknormdatensätze in der GND abzulegen. Das Werk bildet die zentrale Entität innerhalb des FRBR-Modells. Darüber können Expressionen und Manifestationen strukturiert und vernetzt werden. An ihm hängen weitere Entitäten, wie z. B. Personen oder Themen, und es stellt auch über den Bibliotheksbereich hinaus einen Knotenpunkt im semantischen Netz dar, an den andere Communities, wie z. B. Archive und Museen, anknüpfen können. Wenn man davon ausgeht, dass ein Großteil der Werksätze in Zukunft nicht intellektuell, sondern maschinell entstehen wird, würde ein Verzicht auf die Hinterlegung der maschinell generierten Werksätze in der GND bedeuten, dass ein großer Anteil für das semantische Netz nicht verfügbar sein wird. Allerdings müssen Qualitätssicherungsmechanismen geschaffen werden, um einer Überschwemmung der GND mit redundanten oder falschen Werknormdatensätzen vorzubeugen. Als kooperativ geführte Normdatei bietet die GND etablierte Verfahren für die Erstellung und Pflege von Normdaten, zum Datentausch und zur Nachnutzung in unterschiedlichen Systemen, die allerdings für ein automatisiertes Verfahren zu erweitern sind. Denkbar wäre zum Beispiel ein zentrales Clusterverfahren, in das die Bestände aller Bibliotheksverbünde des deutschen Sprachraums einbezogen werden.

Wenn man bei der Entscheidung die übergeordneten Ziele für die Erstellung von Metadaten berücksichtigt, nämlich das Auffinden von relevanten Informationen und ein sicheres Navigieren durch das „unendliche“ Netz der Informationen zu ermöglichen, spricht alles dafür, die GND als Speicherort auch für maschinelle erstellte Werknormdatensätze zu verwenden.

2.5. Die nächsten Schritte

Im RDA-Projekt soll das Thema Werkclustering 2017 wieder aufgegriffen werden. Dazu sollen in einem ersten Schritt ein Konzept erstellt sowie grundsätzliche Fragen geklärt und abgestimmt werden. Zentrale Fragestellungen betreffen dabei die Bedingungen zur Verwendung der GND, wie z. B. Qualitätssicherungsmechanismen, oder die Frage, wie die Teilnehmenden an der GND möglichst weitgehend Clusterergebnisse nachnutzen können, sowie die Konsequenzen für den Datentausch. Damit eng verbunden ist die Frage, ob es sinnvoll ist, einen nationalen Clusterdienst unter Einbeziehung der Daten der Verbünde aufzubauen.

Ferner ist näher festzulegen, auf welche Daten das Clusterverfahren angewendet werden soll. Dabei ist z. B. zu klären, ob nur RDA-Daten oder auch die nach RAK erschlossenen Altdaten mit einbezogen und ob bestimmtes Material, wie z. B. Zeitschriften, ausgeschlossen werden sollen. Auch die Einbeziehung internationaler Werkdateien, etwa OCLC-Works[6], ist zu untersuchen.

Ein weiteres Arbeitspaket ist die Überarbeitung des schon in der ersten Phase des Projekts erarbeiteten Algorithmus für RAK-Daten hinsichtlich RDA. Für das Werkclustering im DNB-Bestand sind die Festlegungen für die Behandlung der Werkangaben im Katalogisat Grundlage. Ziel ist es, alle Katalogisate in das Clusterverfahren einzubeziehen, unabhängig davon, in welcher Form die Werkangaben bei der intellektuellen Erschließung erfasst wurden (Verknüpfung zum GND-Werknormsatz, textuelle Erfassung oder implizites Vorliegen in den Angaben der Manifestation).

3. Zukunftsszenario Sucheinstieg über die Werkebene in nationalen und internationalen Katalogen

Zukünftig sollen Suchumgebungen die Werkebene als zentralen Einstiegspunkt anbieten. In nationalen und internationalen Katalogen ist die FRBR-gemäße Darstellung von bibliografischen Daten ein Mehrwert, der den Nutzerinnen und Nutzern den Einstieg über das der Ressource zu Grunde liegende Werk anbietet und sie zu anderen Expressionen und Manifestationen führt. Auch für die Linked-Data-Cloud stellen Werke einen wichtigen Knotenpunkt dar: Sie spannen einen Informationsraum aus den an ihnen hängenden Expressionen und Manifestationen auf, der mit anderen Entitäten verbunden werden kann. Die Beziehungen zwischen Werken sowie zu Entitäten der FRBR-Gruppe 2 (Personen, Familien und Körperschaften) und Themen können zu anderen interessanten Inhalten führen.

6 Die von OCLC entwickelten Algorithmen wurden zur Veröffentlichung der „Worldcat Works“ als Linked Data verwendet.

Dabei ist die Diskussion um die Regeln zur Bestimmung der Werkebene, die Mitarbeit an der Weiterentwicklung des Regelwerks und die Kenntnis über Suchstrategien und Nutzerbedürfnisse von hoher Wichtigkeit.

Mit der Einführung von RDA, der damit einhergehenden Harmonisierung der Regelwerke für Formal- und Sacherschließung im Bereich der Werke, der Einführung der obligatorischen Angabe des Werks in Katalogdaten und dem zukünftigen Einsatz maschineller Verfahren können die Recherche für die Nutzerinnen und Nutzer in Suchumgebungen und die Vernetzung in der Linked-Data-Cloud verbessert werden. Aber auch die Erschließung wird vom Vorliegen größerer Mengen von RDA-Katalogdaten und maschinell erzeugten Werknormsätzen bzw. nachnutzbaren maschinellen Clusterergebnissen profitieren.

Literaturverzeichnis

- Gatenby, Janifer, Richard O. Green, W. Michael Oskins und Gail Thornburg. „GLIMIR: Manifestation and Content Clustering within WorldCat." *Code4Lib journal* 17 (2012). Zuletzt geprüft am 26.07.2016. http://journal.code4lib.org/articles/6812

- Hickey, Thomas B. und Jenny Toves. *FRBR Work-Set Algorithm*. Version 2.0. Dublin, Ohio: OCLC Online Computer Library Center, Inc. (Research disivion), 2009. Zuletzt geprüft am 26.07.2016. http://www.oclc.org/research/activities/past/orprojects/frbralgorithm/2009-08.pdf

- Pfeffer, Magnus. „Using Clustering Across Union Catalogues to Enrich Entries with Indexing Information." In *Data Analysis, Machine Learning and Knowledge Discovery*, herausgegeben von Myra Spiliopoulou, Lars Schmidt-Thieme und Ruth Janning, 437–445. Cham: Springer International Publishing, 2014. https://dx.doi.org/10.1007/978-3-319-01595-8

- Wiesenmüller, Heidrun und Magnus Pfeffer. „Abgleichen, anreichern, verknüpfen: das Clustering-Verfahren – eine neue Möglichkeit für die Analyse und Verbesserung von Katalogdaten." *BuB: Forum Bibliothek und Information*, Nr. 9 (2013): 625–629.