Alexander Haffner

Institutionenübergreifende Integration von Normdaten (IN2N)

Gestiegene Anforderungen an Bibliotheken durch das Web

Mit der Bereitstellung von Medien und vor allem deren Erschließungsangaben im Web haben sich die Anforderungen an Bibliotheken über die Zeit geändert. Eindeutige Identifizierbarkeit ist eine der immer lauter werdenden Forderungen – die Angabe des Autorennamens oder auch des Verlags sind nicht mehr ausreichend. Dem Gedanken semantischer Vernetzung folgend werden für Mitwirkende am Medienwerk, wie auch für das Medienwerk selbst, eindeutige Beschreibungen angelegt, die im Web referenzierbar und untereinander verlinkt sind. Verlässliche Nachweise zu Publikationslisten von Verlagen, Reputationen von Wissenschaftlern oder auch Diskografien von Interpreten sind das Ergebnis, welches durch technische Neuerungen nun nicht mehr ausschließlich dem Benutzer einer Bibliothek, sondern allen Menschen im Web sowie Maschinen zur freien Nachnutzung zur Verfügung stehen.

Eindrucksvolles Beispiel GND

Mit der Gemeinsamen Normdatei (GND) hat das deutschsprachige Bibliothekswesen eindrucksvoll gezeigt, wie Beschreibungen für Personen, Körperschaften, Schlagwörter etc. in gigantischem Umfang – derzeit über 10 Mio. Eintragungen – innerhalb eines semantischen Netzes bereitgestellt werden können. An dieser Stelle ist der damit verbundene Arbeitsaufwand der Kolleginnen und Kollegen mit dem ihm zustehendem Respekt hervorzuheben. Allerdings ist auch die Frage zu stellen, wo bei der aktuellen Praxis und dem hohen Qualitätsanspruch Grenzen der bibliothekarischen Leistungsfähigkeit erreicht werden und ob es spartenfremde Akteure gibt, die ähnlich agieren und ihre Bemühungen mit denen der GND-Kooperationspartner bündeln können.

Das Projekt IN2N

IN2N ist ein von der Deutschen Forschungsgemeinschaft (DFG) gefördertes Kooperationsprojekt zwischen der Deutschen Nationalbibliothek (DNB) und dem Deutschen Filminstitut (DIF). Das Projekt startete im Dezember 2012 und hat eine Laufzeit von zwei Jahren.

Projektbeschreibung und Projektziele

Ziel von IN2N ist es, ein Kooperationsmodell für eine domänenübergreifende Normdatenpflege zu entwickeln und zu erproben. Zur Zielgruppe gehören nicht bibliothekarische Einrichtungen, die sich in ihrer täglichen Arbeit ebenfalls der Erschließung mithilfe von Normdaten widmen. Diese Einrichtungen sollen von der bereits in der GND getätigten Arbeit profitieren sowie durch die angestrebte Kooperation zum Ausbau und der Qualitätserhöhung der GND beitragen.

Oberste Prämisse des neu zu etablierenden Kooperationsmodells ist die Vereinfachung der derzeitigen Prozesse zur Normdatenpflege. Aktuelle Datenstrukturen, Schnittstellen und Redaktionskonzepte sind bisher vollkommen auf bibliothekarische Systeme und Nutzer zugeschnitten. Die Ergebnisse aus IN2N sollen dazu beitragen, eine Alternative für domänenfremde Akteure zu schaffen, die die bislang notwendige Spiegelung der GND-Daten und die damit verbundene Übereinstimmung des Datenmodells und des Datenformates durch innovative und zeitgemäße Lösungen ersetzt.

Neu zu etablierendes Kooperationsmodell

Exemplarisch sollen hierzu die Personendaten des Deutschen Filminstituts (DIF), die im Internetportal zum deutschen Film[1)] zugänglich sind, mit den entsprechenden Personendaten der GND zusammengeführt und über die Projektlaufzeit hinaus kooperativ gepflegt werden.

Verallgemeinert betrachtet wird das Kooperationsmodell zwei Phasen für neue Kooperationspartner bereithalten:

- Initialer Datenabgleich zwischen dem Datenbestand des Partners und der GND sowie anschließender beiderseitiger Import von Informationen, die durch die Gegenseite erwünscht sind, aber bislang nicht lokal existieren.
- Ein redaktioneller Routinebetrieb über das Web, wobei durch den Partner in Echtzeit in der GND

gesucht wird sowie Änderungen im Bestand des Partners über Differenzmeldungen in die GND übermittelt werden.

Die entwickelten Verfahren, Werkzeuge und Dienste werden auf andere Szenarien und weitere Partner übertragbar sein und damit eine allgemeine Grundlage für den domänenübergreifenden Einsatz der GND-Normdaten bilden.

Initialer Datenabgleich und initiale Datenübernahme

Die Realisierung einer kooperativen Redaktion verlangt zunächst, die bestehenden Datensets der Kooperationspartner initial auf Übereinstimmungen innerhalb der Entitätenbeschreibungen zu prüfen. Im Fall von IN2N bedeutet das, zu den rund 180.000 Personen aus filmportal.de ein Äquivalent in den knapp 2,9 Mio. Personendatensätzen der GND zu identifizieren.

Kooperative Redaktion ist auf übereinstimmende Entitätenbeschreibungen angewiesen

Um den initialen Abgleich performant zu halten, wurde ein Kernset bestehend aus Elementen für Namensformen, Zeit- und Ortsangaben zu Geburt und Ableben, Berufen und Geschlecht gebildet, welches als Grundlage für den angewandten Matching-Algorithmus dient.

Die angestellten Berechnungen kategorisieren Personen aus filmportal.de in drei Klassen:

- exakt ein Äquivalent in der GND identifiziert,
- ein oder mehrere potenzielle Äquivalente in der GND identifiziert,
- kein Äquivalent in der GND identifiziert.

Im Rahmen der initialen Dateneinspielung werden alle Entitätenbeschreibungen des Partners in die GND übernommen, sofern sie exakt einer oder keiner existierenden GND-Entität zugewiesen sind. Des Weiteren müssen die Daten dem GND-Minimalset für individualisierte Personen genügen.

Übernahme der Entitätenbeschreibungen in die GND

Folglich gilt es, Personen der Kategorie 2 in die Gruppen 1 und 3 aufzulösen. Hierfür ist eine intellektuelle Mitwirkung unausweichlich. Eine effiziente Durchführung der intellektuellen Zuweisung wird mittels eines webbasierten Werkzeuges erreicht. Die Applikation bietet dem Redakteur die wichtigsten Informationen zur zuzuweisenden Person aus filmportal.de sowie zu allen potenziellen GND-Äquivalenten auf einen Blick sowie eine direkte Verlinkung zu den Ursprungsportalen. Darüber hinaus sieht der Redakteur die berechnete Match-Score, worauf aufbauend die Entscheidung gefällt werden soll, welche Treffer zuerst einer Prüfung unterzogen werden.

Ein Nachteil des Abgleichverfahrens ist, dass keine Titeldaten einbezogen werden können, da filmografische Werke nur begrenzt im Bestand der GND und DNB nachgewiesen sind. Um die Matchergebnisse zu verfeinern, wurde ein zusätzlicher Abgleich der Daten aus filmportal.de mit den Personenartikeln aus Wikipedia, welche umfangreiche Filmografien auflisten, unternommen. Unter den bislang 230.000 mit der GND verlinkten Personen der Wikipedia konnten auf diesem Wege bereits über 11.000 eindeutig zu filmportal.de als äquivalent identifiziert werden.

Titeldaten werden nicht in den Abgleich einbezogen

Bei der initialen Dateneinspielung in die GND werden neben dem Kernset auch weitere Informationen wie biografische und historische Angaben, Affiliationen etc. importiert. Bei Vorhandensein eines äquivalenten GND-Datensatzes werden die Charakteristika, sofern noch nicht Teil des GND-Satzes, ergänzt. Falls keine passende GND-Person existiert, wird ein neuer Datensatz für die Person aus filmportal.de angelegt.

Filmportal.de hingegen übernimmt aus der GND nur Ergänzungen zu den bereits in ihrem System vorhandenen Daten.

Mit Abschluss der initialen Einspielung gehen die Partner in den redaktionellen Routinebetrieb über. Zu diesem Zeitpunkt stehen allen bisherigen GND-Kooperationspartnern die Personendaten aus filmportal.de auf ihren gewohnten Datenbezugswegen zur Verfügung.

Redaktioneller Routinebetrieb

Der dem Routinebetrieb zugrundeliegende Anwendungsfall nimmt als Basis einen Redakteur einer nicht bibliothekarischen Einrichtung an, der in seinem lokalen Redaktionssystem seiner Arbeit nachgeht. Falls dieser Änderungen an Normdaten vornimmt, die ein Pendant in der GND besitzen, sollen diese ohne weiteres Zutun des Redakteurs in die GND übernommen werden. Falls keine passende Person im lokalen System existiert, sucht der

Aufgaben des Redakteurs

Redakteur auf Basis des Namens und ggf. der Lebensdaten eine entsprechende Beschreibung in der GND. Die Eingabe der Suchcharakteristika, wie auch die Präsentation der Ergebnismenge, findet im lokalen Redaktionssystem statt. Durch bequeme Auswahl kann der Redakteur die Informationen einer GND-Person nachnutzen und bei Bedarf ergänzen.

Was verbirgt sich technisch hinter diesem Anwendungsfall? Nach dem initialen Datenabgleich besitzt jede Person aus filmportal.de eine Referenz auf die zugehörige GND-Person. Aktualisierungen der verknüpften GND-Entitäten werden einerseits regelmäßig und zusätzlich vor jeder Bearbeitung dieser in filmportal.de übernommen. Dies garantiert die notwendige Synchronität.

Inhaltsbasierte Suche in der GND

Die Suche aus dem bzw. der Datenbezug durch das Redaktionssystem wird über SRU (Search / Retrieve via URL) realisiert. SRU ist ein standardisiertes Webservice-Protokoll, um bibliothekarische Datenbanken im Internet abzufragen. Über die Schnittstelle stehen Datenformate wie MARC 21, aber auch GND/RDF bereit. SRU bietet einen einfachen Mechanismus für die inhaltsbasierte Suche in der GND, auch wenn an dieser Stelle mit dem Vorhaben, bibliothekarische Schnittstellen und Formate für domänenfremde Akteure zu vermeiden, gebrochen wird.

Hervorzuheben ist, dass filmportal.de nur ein Subset der GND-Beschreibung übernimmt. Beispielsweise werden keine GND-Ländercodes nachgenutzt, da für die GND andere Vergaberichtlinien existieren. Folglich gilt es, für schreibende Aktionen Alternativen zum derzeitig praktizierten Datensatzbasierten Ansatz zu finden.

REST-Schnittstelle für schreibenden Zugriff

Für den schreibenden Zugriff soll eine neue REST-Schnittstelle für inkrementelle Updates etabliert werden. Ressourcen werden mittels einer HTTP-basierten Anfrage angesprochen und durch eine PUT-Operation geändert bzw. falls noch nicht existent, neu angelegt. Auf Property-Ebene (entspricht im bibliothekarischen Format einem Feld mit Unterfeldern) wird die Schnittstelle drei Operationen für die Datenmanipulation anbieten: hinzufügen (add), ändern (change) und löschen (delete) von Objektcharakteristika. Die Änderungsoperationen für einen bestimmten Datensatz werden in einem JSON-Request eingebettet und mit der HTTP-Anfrage mitgesandt. Für Personen steht bislang ein Set von ungefähr 25 Datenelementen zur Verfügung, die zur Datensatzanpassung einsetzbar sind.

Die Innovation der neu zu gestaltenden Schnittstelle findet sich in dem Ansatz, keine kompletten Datensätze zu harvesten, nachfolgend zu manipulieren und final in die GND zurückzuschreiben, sondern lediglich Differenzen zum aktuellen Datensatz zu übermitteln.

Neue Anwendungsszenarien: Normdatenmitgestaltung ohne Kenntnis des Datensatzes

Dadurch bestünde sogar die Möglichkeit für Akteure, die keine GND-Daten beziehen, bibliothekarische Normdaten mitzugestalten. Vorstellbar wäre die Übernahme von Informationen aus Online-Plattformen, insofern die Angaben in der GND nicht vorliegen. Falls beispielsweise ein Wikipedia-Artikel mit einer GND-Person verknüpft ist und ein Sterbedatum eingetragen wird, könnte dies einfach und bequem in die GND eingepflegt werden. Ähnliches gilt für soziale Netzwerke von Wissenschaftlern, in denen Wissenschaftler ihre personenbezogenen Daten sowie ihre Publikationstätigkeit selbst verwalten. Der Transfer von Informationen kann ohne Kenntnis des eigentlichen GND-Datensatzes, ausschließlich unter Angabe des Uniform Resource Identifier (URI) der GND-Ressource geschehen.

Die neue Schnittstelle verlangt wie die bisherige GND-Kooperation eine Registrierung der Partner. Änderungs- und Löschoperationen sind bei diesem Ansatz mit besonderer Vorsicht zu genießen und sollten nur durch explizite Rechtezuweisung gestattet sein.

Herausforderungen innerhalb des Projektes

Insbesondere der Abgleich mit Normdaten anderer Domänen ohne Einbeziehung von Titeldaten stellt eine große Herausforderung dar. Entsprechend können für den Abgleich ausschließlich Charakteristika der Personenbeschreibungen miteinander verglichen werden.

Herausforderung: Abgleich mit Normdaten anderer Domänen

Der Übereinstimmung von Namensformen kommt dadurch eine sehr große Bedeutung zu. Da filmportal.de und die GND unterschiedliche Namensbestandteile in ihren Datenmodellen vorsehen und ihnen unterschiedliche Regeln zugrunde liegen, wurde ein mehrstufiges Abgleichverfahren entwick-

elt, welches von vollständiger Einbeziehung aller Namensbestandteile in einen nachgestellten Abgleich mit ausgewählten Namensbestandteilen übergeht. Durch geschicktes Kombinieren der Namensbestandteile konnten vergleichbare Strings geschaffen werden.

Bei den weiteren Match-Charakteristika mussten lediglich syntaktische Anpassungen vorgenommen bzw. kontrollierte Vokabulare für den Abgleich zur Verfügung gestellt werden. Zu den literalen Angaben von Geburts- und Sterbeorten aus filmportal.de konnten über die bereits bekannten 11.000 Äquivalenzpaare bereits mehr als 1.000 geografische Entitäten der GND zugeordnet werden. Hierfür wurde ein semi-automatisches Verfahren eingesetzt.

Eine weitere noch offene Frage stellt sich für das initiale Einspielen der Daten in die GND. Aktuell werden verschiedene Ansätze evaluiert, wobei seitens der DNB die Anforderung besteht, möglichst die durch OCLC zur Verfügung gestellten Module nachzunutzen. Durch filmportal.de ist angedacht, die Mechanismen des Routinebetriebs auch für die initiale Dateneinspielung einzusetzen.

Module von OCLC sollen für die Einspielung in die GND nachgenutzt werden

Neue Anwender bringen neue Workflows. Die Öffnung der GND ist unumstritten. Welche Nutzergruppen Berücksichtigung finden müssen und welche Auswirkungen die neuen Anwendungsfälle auf das bestehende Redaktionskonzept haben, wird gegenwärtig geprüft. Eine Abstimmung mit Expertengruppen und Gremien für die Normdatenerschließung im deutschsprachigen Raum wird einer der essenziellen Schritte sein, um domänenfremde Akteure in der kooperativen Normdatenpflege zukünftig begrüßen zu dürfen.

IN2N-Zeitplan

Bis Ende 2013 ist die Fertigstellung des initialen Matches sowie der Module für die initiale Dateneinspielung vorgesehen. Anfang 2014 wird die intellektuelle Zuweisung für die eindeutige Zuordnung von Personen aus filmportal.de und der GND vollendet und die Migration zum neuen Redaktionssystem mit GND-Anbindung bei filmportal.de realisiert. Für das zweite und dritte Quartal 2014 sind eine Evaluation der Ergebnisse und etwaige Verbesserungen an den Komponenten angedacht. Zeitgleich werden die Projektergebnisse verstärkt kommuniziert und es wird in eine Akquise- und Beratungsphase zur Gewinnung neuer Partner übergegangen. Es gilt, Aktivitäten zur Einbindung neuer Kooperationspartner innerhalb der DNB zu verstetigen.

Fazit

Das IN2N-Projekt ist bestrebt, die technischen und organisatorischen Voraussetzungen für eine domänenübergreifende Normdatenkooperation zu schaffen. Nicht bibliothekarischen Einrichtungen wird die Möglichkeit eröffnet, ohne Kenntnis von hochkomplexen bibliotheksspezifischen Formaten wie auch ohne tiefgründige Regelwerkskenntnis am Erschließungsprozess teilzuhaben.

Zukünftige Ausdehnung der kooperativen Normdatenpflege auf weitere Institutionen

Die Kooperation mit dem DIF wird exemplarisch die Leistungsfähigkeit des verfolgten Ansatzes und des darauf aufbauenden Kooperationsmodells zeigen. Die Projektergebnisse sollen weiteren Einrichtungen aus Wissenschaft, Verlagswesen, Social Web, Kultur, aber auch aus dem Bibliothekswesen aufzeigen, wie eine kooperative Normdatenpflege aus dezentral organisierten Datenbeständen über das Web in der Praxis angewandt werden kann.

Die exemplarische Umsetzung für Personendaten ist ein erster Schritt. Das Konzept ist auf weitere Entitätentypen wie auf weitere Properties für die Manipulation einzelner Entitäten erweiterbar.

Die Einführung und Verbreitung der IN2N-Ergebnisse ist neben dem eigenen Projekterfolg stark davon abhängig, inwieweit das Bibliothekswesen für eine nicht bibliothekarische Öffnung bestimmter Teile der GND bereit ist.

Anmerkungen

1 <http://www. filmportal.de>