Mit „Match & Merge“ zur GND? Erprobung von Anreicherungsverfahren im Archivportal-D-Projekt (*Veröffentlicht in: Archivar 71 (2017), H. 2*)

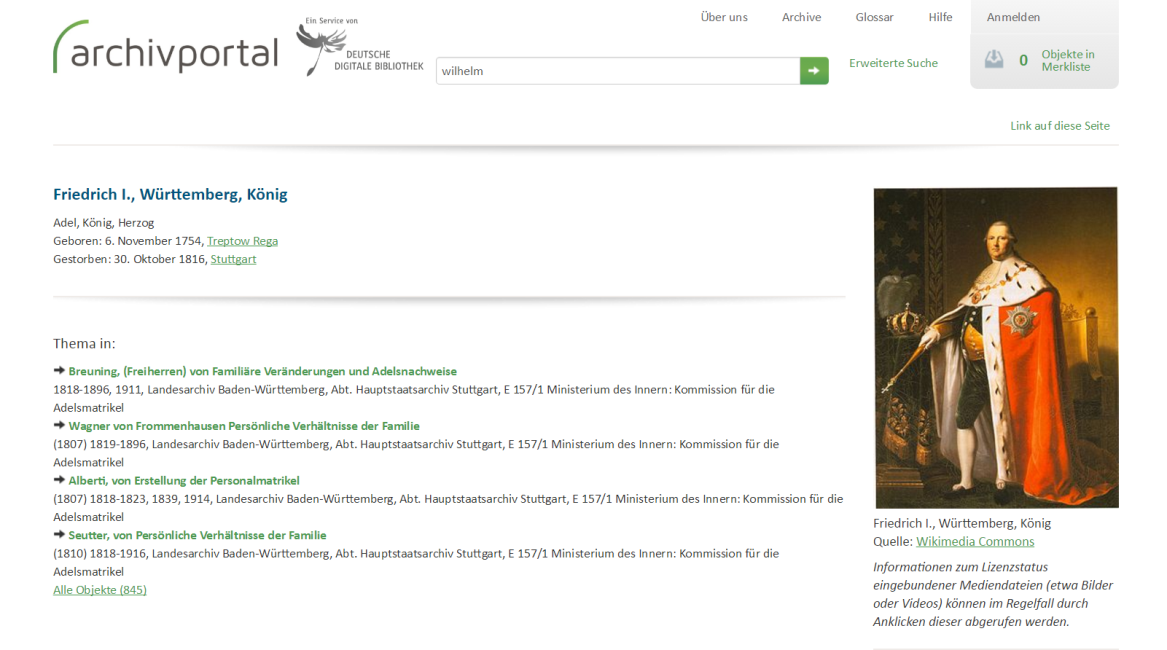
*Von Daniel Fähle und Nadine Seidu*

Es ist ein verlockender Gedanke: Statt bei der Verzeichnung oft mühsam den zutreffenden Personeneintrag in der Gemeinsamen Normdatei (GND) recherchieren zu müssen, um dann die Normdatenverknüpfung in die eigene Datenhaltung zu übernehmen, könnten bereits vorliegende Findbücher automatisiert mit der Gemeinsamen Normdatei abgeglichen („Matching“) und die Ergebnisse ins Erschließungssystem rückgespielt werden („Merging“). Viele Hunderte, ja Tausende Normdatenreferenzen ließen sich so gegebenenfalls in kurzer Zeit generieren. Bislang ist die Datenbasis hierzu im Archivbereich überschaubar. Mit nachgelagerten, automatisierten Anreicherungsverfahren – so die Hoffnung – könnten Findmittel und Archivgut einrichtungsübergreifend in großem Maßstab maschinell miteinander verknüpft und damit inhaltliche Verbindungen zwischen Archivalien unterschiedlicher Provenienz aufgezeigt werden. Zudem könnten die neu eingerichteten Personenseiten des Archivportals-D (siehe Abbildung),[[1]](#footnote-1) die auf der Basis von Normdaten qualifizierte Einstiegspunkte zur Recherche von Personenbezügen in Archivgut bieten, rasch mit Inhalten gefüllt werden.

**Das Projekt**

Doch ist eine automatisierte Anreicherung von Archivdaten überhaupt möglich? Und wenn ja, unter welchen Voraussetzungen? Diese Fragen waren Gegenstand eines Schwerpunkt-Arbeitspakets in der zweiten Förderphase des Archivportal-D-Projekts.[[2]](#footnote-2) Im Mittelpunkt stand die Erprobung von Text-Mining-Prozessen bzw. eine Adaption des bei der Deutschen Nationalbibliothek (DNB) für den Metadatenabgleich eingesetzten Standardverfahrens „Match & Merge“ für Archivdaten. In einem ersten vorbereitenden Schritt wurde ein Set an Beständen und Ausgangsformaten definiert, das grundsätzlich geeignet für eine Anreicherung erschien. Die Auswahl umfasste unterschiedliche Datenstrukturen und Erschließungstiefen, um die Varianz möglicher archivischer Ausgangsdaten nachzubilden. Formal sollte ein möglichst breites Spektrum an Personenbezügen berücksichtigt werden, also neben dezidiert strukturierten Personenakten auch Personen-Indexbegriffe bzw. -Deskriptoren, aber auch strukturierte Nennungen in Titel- und Enthält-Feldern sowie sonstigen Fließtexten in Erschließungsinformationen.

Außer dem Landesarchiv Baden-Württemberg stellte das Hessische Landesarchiv[[3]](#footnote-3) sowie das Archiv der Max-Planck-Gesellschaft Daten für die Evaluierung zur Verfügung. Insgesamt wurden 13 Bestände und zwei Deskriptorenlisten bzw. Personen-Indices analysiert. Es handelte sich bei den Beständen u.a. um Nachlässe, Justizakten, Personalakten der Finanzverwaltung, Patienten-/Personalakten eines Hospitals sowie Bilder- und Zeitungsauschnittsammlungen.



[[*Abbildung: Screenshot der Personenseite zu König Friedrich I. von Württemberg im Archivportal-D*]]

**Das Verfahren**

An dieser Stelle kann das Prozedere nicht in allen technischen Details geschildert werden.[[4]](#footnote-4) Zusammenfassend lässt es sich folgendermaßen beschreiben: Bei dem „Match & Merge-Verfahren“ der DNB werden gelieferte Datensätze gegen die vorhandenen GND-Einträge abgeglichen.[[5]](#footnote-5) Die Daten müssen dazu im bibliothekarischen PICA-Format vorliegen. Notwendig wurde also eine Konvertierung der proprietären archivischen Lieferformate oder auch der Daten im EAD(DDB)-Profilformat. Hierzu mussten zunächst die relevanten personenbezogenen Informationen in den Beständen identifiziert und anschließend extrahiert werden. Die Herausforderung bestand darin, jeweils die Struktur bzw. das Muster der enthaltenen Informationen wie Vor- und Nachname, Geburts- und Sterbedatum, Geburts- und Sterbeorte sowie Berufe oder Funktionen zu erkennen, um eine Regel („Mapping“) für die Transformation der Daten definieren zu können. Eine derartige Regel kann aber nur erstellt werden, wenn die Erschließungsdaten überhaupt strukturiert erfasst wurden und diese Strukturierung pro Datenlieferung einheitlich ist.

Bei dem eigentlichen Metadatenabgleich werden die extrahierten personenbezogenen Informationen dann mit denen vorhandener GND-Einträge abgeglichen. Dabei wird ein sog. Match- oder Konfidenzwert ermittelt, der die Wahrscheinlichkeit einer Übereinstimmung mit einer GND-Entität wiedergibt. Hohe Matchwerte[[6]](#footnote-6) ergeben sich, wenn über Namensübereinstimmungen hinaus zudem das Geburts- oder Sterbejahr identisch ist sowie zusätzlich der Beruf oder der Geburts-/Sterbeort. Mindestvoraussetzung für ein Match ist, dass sowohl im gelieferten Datensatz als auch im GND-Datensatz zumindest ein Lebensdatum vorhanden sein und übereinstimmen muss.

Als Resultat des Abgleichs ergibt sich demnach eine Liste 1) mit den gelieferten Personen, 2) diesen im Falle eines Matches zugeordnet eine GND-Personenentität sowie 3) ein Konfidenzwert zur Bewertung der Übereinstimmung. Aufgrund der Erfahrungswerte der DNB-Kollegen, denen im Übrigen ein Großteil der Aufgaben des Projekts zukam und denen mithin die meisten Ergebnisse des Projekts zu verdanken sind, kann bei hohen Konfidenzwerten mit großer Wahrscheinlichkeit von verlässlichen Zuordnungen ausgegangen werden. Letztgültige Sicherheit bringen aber naturgemäß nur intellektuelle Prüfung und Bestätigung der Zuordnungsempfehlungen. Stichproben und Plausibilitätskontrollen sind daher in jedem Fall unerlässlich.

**Die Ergebnisse**

Von den insgesamt 13 betrachteten Beständen erwiesen sich im Projektverlauf letztlich drei als ungeeignet für eine automatisierte Anreicherung, da die Extraktion der relevanten Personeninformationen ein sehr aufwändiges Text-Mining unter Einbeziehung externer Datenquellen zur Kontextualisierung erfordert hätte. Die übrigen 10 Bestände sowie die zwei Personen-Indices wurden aufgrund der einheitlichen Strukturierung und Reichhaltigkeit an individualisierenden Personenmerkmalen (Lebensdaten, Berufe) als „gut“ bzw. „sehr gut“ geeignet klassifiziert. Aus diesen gelieferten Daten konnten insgesamt 205.466 Personen(-bezüge) extrahiert werden, zu denen via „Match & Merge“ 9.204 GND-Einträge mit hohem Konfidenzwert zugeordnet werden konnten. Die Zuordnungsquote betrug damit aufs Ganze gesehen 4,48 Prozent. Jedoch variieren die Quoten bei den einzelnen Beständen stark, von unter 0,1 Prozent (Patienten-/Personalakten eines Krankenhauses) bis zu 23 Prozent (Zeitungsauschnittsammlung der Personengeschichte). Bei den meisten Beständen ist die geringe Quote jedoch nicht auf Datendefizite wie unzureichende Strukturierung oder fehlende individualisierende Personenmerkmale zurückzuführen, sondern schlicht auf den erwartbaren Umstand, dass in der GND die entsprechenden Personenpendants nicht vorhanden sind. Denn verlässlich enthält die Normdatei (bislang) nur Prominente bzw. Personen hoher Bekanntheit, bei anderen Personengruppen variiert die GND-Abdeckung stark.

Die aus dem Projekt gewonnen Erkenntnisse lassen sich demnach wie folgt zusammenfassen:

1. Das evaluierte Anreicherungsverfahren der DNB konnte für den Archivbereich adaptiert werden. Die Extraktion der Personendaten aus den archivischen Erschließungsinformation war jedoch recht aufwändig, da aufgrund fehlender übergreifender Standards in diesem Bereich für jeden Bestand ein unterschiedliches Regelwerk erarbeitet werden musste.
2. Archivdaten eignen sich grundsätzlich für eine nachgelagerte Anreicherung, wenn die folgenden Kriterien erfüllt sind:
   1. Es ist eine ausreichende Datentiefe vorhanden, d.h. Personen sind mit vollständigem Namen erfasst und durch weitere Merkmale wie Geburts-/Sterbedatum, Geburts-/Sterbeort sowie Berufe/Funktionen individualisierbar.
   2. Die personenbezogenen Daten werden gleichförmig strukturiert erschlossen (z.B. über entsprechende Formulare) und die einzelnen Informationen werden in der Archivsoftware in separaten Feldern vorgehalten. Damit können diese Daten für Abgleiche normiert ausgespielt werden.
3. Für die Anreicherung eignen sich angesichts des aktuellen GND-Personenbestands insbesondere solche Bestände, die prominente Persönlichkeiten enthalten.
4. Eine vollständige Automatisierung der Anreicherung ist unrealistisch, allenfalls kann von semi-automatischen Verfahren gesprochen werden. Datenanalyse und -extraktion, die Definition von Abgleichsalgorithmen und die individuelle Bewertung von Konfidenzwerten sind intellektuelle Prozessschritte.
5. Auch die Rückführung der mit GND-Verknüpfungen angereicherten Daten in das Ursprungssystem, z.B. in Form einer Konkordanzliste, muss von Beginn an mit bedacht werden.

**Der Ausblick**

Aus den Projektergebnissen heraus ergeben sich Handlungsbedarfe und auch Anknüpfungspunkte für konkrete weitere Schritte und Planungen. Wie dargelegt basiert der Erfolg von Anreicherungsverfahren ganz allgemein gesprochen auf einer entsprechend guten Erschließung. Einheitlichkeit fördernden Richtlinien, Strukturvorgaben und der Verwendung von Standards kommt in diesem Bereich größte Bedeutung zu.[[7]](#footnote-7) Sie bilden die Grundlage erfolgversprechender maschineller Verarbeitung. Es wird auch zu überlegen sein, wie z.B. beim bestehenden archivischen Standard-XML-Profil EAD(DDB) einheitliche Maßgaben zur Binnenstrukturierung von Personeninformationen (und auch mit Blick auf weitere Entitäten) verankert werden können. Über syntaktische Vorgaben könnte EAD(DDB) im Sinne eines generischen Ansatzes für einen neu zu konzipierenden Anreicherungsdienst „fit gemacht“ werden.

Die Etablierung eines derartigen Dienstes wird bereits im spartenübergreifenden Kontext diskutiert: Auch vor dem Hintergrund der Erfahrungen im Archivportal-D-Projekt entwickelte sich seitens der DNB die Idee, die angestrebte Öffnung der GND für nicht-bibliothekarische Kultursparten in einem neuen Projekt anzugehen. Die Konzeption und Realisierung eines für Archive, Museen, Mediatheken etc. gleichermaßen praxistauglichen Self-Service-Dienstes für Datenanreicherungen ist dabei nur ein Schwerpunkt. Denn ein weiteres Hindernis bei der GND-Referenzierung soll angegangen werden: das Fehlen benötigter GND-Einträge. Es wird dazu nötig werden, dass auch Archive aktiv an der GND als zukünftig spartenübergreifendem Kooperationsprojekt mitarbeiten. Archivische Einrichtungen werden sich in diesem Zusammenhang allerdings überlegen müssen, für welche Personen sie Normdaten benötigen und anlegen lassen wollen, denn nicht jede identifizierbare Person wird einen eigenen Normdateneintrag rechtfertigen.[[8]](#footnote-8)

Abschließend lässt sich somit festhalten, dass das Projekt zur Erprobung von GND-Anreicherungsverfahren neben den tatsächlich erzielten Normdatenverknüpfungen zu wichtigen Erkenntnissen führte, die in die Zukunft weisen. Viele anfänglich bestehende Vermutungen und Annahmen wurden bestätigt, jedoch lag der Gewinn insbesondere in der Konkretisierung der relevanten Parameter. Damit wurde eine solide empirische Ausgangsbasis für anstehende Lösungsansätze im Bereich der nachgelagerten GND-Anreicherung von Archivdaten geschaffen.

1. Im Zuge des Teilprojekts wurden auch die Entitäten- bzw. Personenseiten innerhalb des Archivportals-D konzipiert und umgesetzt. Diese Seiten weisen sämtliche in den Erschließungsinformationen enthaltenen GND-Referenzen auf einzelne Personen aus. [↑](#footnote-ref-1)
2. Zum Projekt „Aufbau eines Archivportals-D“ siehe zuletzt u. a.: Daniel Fähle u. a.: Archivportal-D. Funktionalität, Entwicklungsperspektiven und Beteiligungsmöglichkeiten. In: Archivar 68 (2015), H. 1, S. 10-19 sowie Gerald Maier u. Christina Wolf: Das Archivportal-D. Neue Zugangswege zu Archivgut innerhalb der Deutschen Digitalen Bibliothek. In: Handbuch Kulturportale. Online-Angebote aus Kultur und Wissenschaft. Hg. von Ellen Euler u. a. Berlin/Boston 2015, S. 180-190. [↑](#footnote-ref-2)
3. Vgl. Peter Sandner: Hessisches Landesarchiv. Pilotprojekt Normdaten. In: Newsletter HessenArchiv aktuell 08/2016, S. 2. [↑](#footnote-ref-3)
4. Die vollständige Dokumentation des Teilprojekts wird auf den Seiten des Archivportals-D zum Projektende veröffentlicht. [↑](#footnote-ref-4)
5. Das Verfahren findet im OCLC/PICA-CBS statt, dem Katalogisierungssystem der GND. [↑](#footnote-ref-5)
6. Der Maximalwert ist 1.00 (entspricht einer vollständigen Übereinstimmung aller Vergleichselemente). Der Schwellwert für Matches im Projekt lag bei 0.80. Die DNB geht ab einem Wert von 0.70 von sehr wahrscheinlichen Zuordnungen aus. [↑](#footnote-ref-6)
7. Im Landesarchiv Baden-Württemberg gibt es z.B. seit 2011 Richtlinien für die Verwendung von Orts- und Personendeskriptoren. Diese enthalten ein Regelwerk zur einheitlichen, maschinenlesbaren Ansetzung von Deskriptoren. [↑](#footnote-ref-7)
8. Es wird also eine Relevanzdiskussion geführt werden müssen. Erste Überlegungen gehen in die Richtung, nur für Personen, zu denen mehrere eindeutig identifizierbare Einträge in einem Archiv vorhanden sind, auch Normdaten anzulegen. [↑](#footnote-ref-8)