Die Normdatei als ein Mittel in der Erschließung von Archivbeständen[1]

Von Gerhard Müller

Abstract

Die Normdatei ist heute ein unverzichtbares Instrument für die Katalogisierung in Bibliotheken. Aufgrund ihrer Charakteristik wurde die Normdatei zudem zu einem entscheidenden Element in der Entwicklung des Web of Data, indem ihre Nutzung zur Strukturierung des online verfügbaren Datenbestandes beiträgt. Die Charakteristika, die dies unterstützten, sind: gemeinsame Regeln der Datenerfassung, Standardformate und -schnittstellen, die Bereitstellung eines eindeutigen, von einer einzelnen Anwendung und Nutzung unabhängigen Identifier je Entität, das ist etwa eine Person, eine Körperschaft, ein Geographikum, ein Ereignis oder eine Sache. Durch die Normdatei gelingt es, diese Entitäten sowohl für Menschen mittels identifizierender Merkmale als auch für Maschinen mittels eines eindeutigen Identifier zu disambiguieren. Daraus resultiert das Potenzial, die Suche nach einer bestimmten Entität in umfangreichen Datenbeständen weiterzuentwickeln. Die Normdatei dient somit nicht mehr nur der Nachnutzung (Reuse) bei der wiederholten Katalogisierung eines Werkes in Bibliotheken, sondern auch der Vernetzung von Daten verschiedener Provenienz (Linked Data). Mit dem neuen Katalog des Kalliope-Verbundes gelang es überdies, mithilfe von Normdaten das Potenzial von Metadaten von Findbüchern und Katalogen für quantitative Methoden zur Erforschung von historischen Zusammenhängen zu demonstrieren (Data Analysis). Der Ursprung der Normdatei im Bibliotheksumfeld ist für die Entscheidung über ihre Nutzung oder Nichtnutzung genauso nachrangig wie der Verweis auf eingeübte Prinzipien, die ihre Wurzeln im 18. und 19. Jahrhundert haben. Vielmehr erzwingt die Digitalisierung der Arbeit auch ein höheres Standardisierungsniveau bei der Erschießung. Dabei liegt der Fokus der Erschließung weiterhin auf dem Bestand, aber durch die konsequente Nutzung von Normdaten tritt die bestandsübergreifende Perspektive hinzu. Bestehende Hürden des Zugangs zur Normdatei sind durch geeignete Modelle und Redaktionsverfahren abzubauen.

Einleitung

Sie suchen nach Rudolf Schmidt? Welchen suchen Sie? Den Ingenieur? Den Künstler? Den General? Den Unternehmer? Den Journalisten? Eine einfache Suche im Archivportal-D zeigt 467[2], das Europäische Archivportal 918[3] und die Europäische Digitale Bibliothek, Europeana, gar 3.685 Treffer[4]. In den Volltexten des europäischen Zeitungsportals: 1.023 Treffer[5]. Doch welcher Rudolf Schmidt? – Die Konversion der Findmittel und zunehmend auch der Quellen in maschinenlesbare Formate reduziert unzweifelhaft den Aufwand für die Recherche. Bequem lassen sich die Datenbanken und Findbücher online mit Stichwörtern von jedem beliebigen Standort mit Zugang zum Internet aus durchsuchen. Ob Metadaten der Kataloge, Repertorien und Inventare oder gar Volltexte von „Born Digital"-Quellen, durch Optical Character Recognition (OCR) oder aufwendig erarbeitete Transkriptionen etwa im Rahmen von Editionsprojekten – die online verfügbare Datenmenge nimmt stetig zu. Zeitgleich änderten die Nutzerinnen und Nutzer mit der Etablierung heute allgemein bekannter Suchmaschinen ihr Verhalten bzw. emanzipierten sich; denn niemand will mehr für die Recherche die Funktionsweise eines Kataloges oder die grundlegenden Prinzipien der Ordnung, Gliederung und Erschließung der Bestände studieren. Bibliotheken suchen eine Antwort auf diese Herausforderung etwa mit Discovery Systemen (vgl. Arndt, 2013). Ob Gefecht im Rückzug oder valide Unternehmung, das Alte mit dem Neuen zu verbinden, wenn etwa für den Erhalt klassischer Sucheinstiege argumentiert wird (vgl. Spinnler-Dürr 2013, 58 ff.), muss sich zeigen. Tatsache ist, dass das, was online nicht zu finden oder zu identifizieren ist, nicht existiert; im Zweifelsfall geht es im Meer der Daten unter. Suchende sind bis heute allein mit der Entscheidung gelassen, ob – um am Beispiel zu bleiben – der Rudolf Schmidt in den Ergebnislisten der Gesuchte ist und ob in den nicht berücksichtigten Daten doch etwas Wesentliches übersehen worden sein kann.

Bereits in der Mitte der 1990er Jahre entwickelte Tim Berners-Lee, Erfinder des World Wide Web, die Idee des Semantic Web. In diesem sollen Informationen eine eindeutige Bedeutung für die Vereinfachung der Interaktion zwischen Mensch und Maschine erhalten. Hieraus ging das Konzept eines Web of Data hervor. Es beruht auf referenzierten Daten (Linked Data) und den dafür entwickelten Standards (vgl. Gradmann et al. 2012, 18 f.). Eine notwendige Bedingung für das Web of Data oder, praxisnäher, für die eindeutige Identifikation von Rudolf Schmidt für Mensch und Maschine sind Merkmale, die

1 *Vortrag auf dem 17. Brandenburgischen Archivtag am 8. Mai 2014. Der Autor ist seit 2012 Leiter des Kalliope-Verbunds, Abteilung Überregionale Bibliographische Dienste, in der Staatsbibliothek zu Berlin – PK. Er studierte Informationswissenschaften in den Fächern Archiv und Dokumentation an der Fachhochschule Potsdam 1998-2002 sowie am Otto-Suhr-Institut Politikwissenschaft, Freie Universität Berlin, 2003-2009. In 2011 war er Referent für die EU-Projektberatung im Kompetenznetzwerk für Bibliotheken (KNB).*

2 *https://www.archivportal-d.de/ (31.03.2015).*

3 *https://www.archivesportaleurope.net/ (31.03.2015).*

4 *http://www.europeana.eu/ (31.03.2015).*

5 *http://www.theeuropeanlibrary.org/tel4/newspapers/ (31.03.2015).*

uns diese Identifikation ermöglichen. Dabei sind die Bedürfnisse von Mensch und Maschine – wenig überraschend – verschieden: Als Menschen ist uns Rudolf Schmidt mit seinen Lebensdaten 1875-1943 und ergänzenden Hinweisen zum Beruf (Journalist in Eberswalde) ausreichend identifiziert, und wir können mit diesen Angaben ohne größere Anstrengungen Artikel und Werke dieser Person, aber auch etwa seinen Nachlass im Kreisarchiv Barnim mithilfe geeigneter Verzeichnisse ermitteln. Für eine Maschine müssten allein hierfür komplexe Algorithmen entwickelt werden, um die vorhandenen Daten zu und von dieser Person zu identifizieren. Eine Alternative ist ein Identifier, hier: 117514608, der Gemeinsamen Normdatei (GND, http://d-nb.info/gnd/117514608). Wird die Entität, das ist eine Person, eine Körperschaft, ein Geographikum, ein Ereignis oder eine Sache, etwa in Lexika wie der Wikipedia[6] oder der Deutschen Biographie[7], in Katalogen, wie dem des Kalliope-Verbunds[8] usw. mit diesem Identifier ausgezeichnet (mark-up), können die vielen unterschiedlichsten, unabhängigen Informations- und Datenangebote etwa mittels BEACON-Dateien[9] vernetzt werden. Dieser Identifier unterstützt somit als Bestandteil einer überregionalen Infrastruktur die Strukturierung einer stark wachsenden Datenmenge. Zugleich flankiert die Auszeichnung von Entitäten die Weiterentwicklung von Algorithmen für die Durchsuchung und Analyse der Daten, etwa mit der Abfragesprache SPARQL[10]. Es ist die Verbindung von Strukturierung mittels Markup und Abfrage mittels Algorithmen, die zu einer systematischeren Aufbereitung der verfügbaren Datenangebote führt, wohingegen die Fokussierung auf Suchmaschinen für schwach- und nicht strukturierte Daten (Volltext-/Stichwortrecherche) zur Bevorzugung einzelner Angebote führen kann (vgl. Introna/Nisssenbaum 2000, 171 f.); denn einem Algorithmus liegen Annahmen über die Beschaffenheit von Datenmengen zugrunde, die auf die eine Teilmenge anwendbar ist, sich aber für eine zweite Teilmenge als unzureichend erweist. Mit dem Formulieren einer Aussage darüber, dass eine Entität in verschiedenen Teilmengen mit demselben Identifier immer auch dieselbe Entität ist, gelingen neue, nutzerorientierte transparente Informationsangebote.

Durch die Nachnutzung von Datensätzen einer Normdatei für ein und dieselbe Entität wird somit Eindeutigkeit in der wachsenden Datenmenge erzielt. Durch diesen Fakt ist ein weiterer Aspekt von hohem Interesse: die Quantifizierbarkeit. Ein nur scheinbar starkes Argument für die Digitalisierung der Bestände von Bibliotheken, Archiven und Museen ist die Möglichkeit, diese online lesen respektive betrachten zu können. Dadurch, dass diese Quellen online sind, reduzierten sich zwar die Kosten etwa für An- und Abreise zum Studium der Quellen vor Ort; dennoch ist dieses Argument bei genauer Betrachtung ein schwaches: Mit Ausnahme der monetären Kosten, die auf den Anbieter der digitalisierten Quellen verlagert werden, wird die eingesparte Zeit kaum sinnvoll für das Studium einiger weniger weiterer digitalisierter Seiten genutzt werden können. Die menschliche Aufnahme- und Verarbeitungsfähigkeit ist begrenzt. Dennoch soll dies kein Argument gegen die weitere Digitalisierung sein. Das Gegenteil ist der Fall: Nicht nur sind die analogen Quellen in Bilddaten zu konvertieren, sondern es sind auch die Verfahren der weiteren Konversion in maschinenlesbare Formate voranzutreiben. Durch die Verfügbarkeit der Metadaten[11] und Texte in Formaten, die auch von Maschinen verarbeitet werden können, entsteht ein Datenschatz für eine Vielzahl von Forschungsdisziplinen. Der Datenschatz wird umso wertvoller, je strukturierter er ist, das heißt die in ihm vorkommenden Entitäten disambiguiert, also sowohl für Mensch als auch Maschine eindeutig identifiziert sind. Diese Daten lassen sich grundsätzlich mit statistischen Verfahren analysieren und erweitern so die Fähigkeit, auch größere Datenmengen systematisch zu verarbeiten. Eindeutig identifizierte Entitäten eröffnen etwa in Verbindung mit Raum- und Zeitangaben erstmals die Möglichkeit, in signifikantem Umfang quantitative und qualitative Methoden für die Analyse historischer Ereignisse fruchtbar anzuwenden. Nicht nur lassen sich Beobachtungen in statistischen Werten ausdrücken, sondern Informationen in Form von Graphen visualisieren. Dem Rückgriff auf Normdaten bei der Erschließung von Archivbeständen kommt in diesem Kontext eine besondere Bedeutung und, wie zu zeigen sein wird, auch Verantwortung zu; das Erschließen hat einen dualen Nutzen – für die Recherche und das Identifizieren relevanter Quellen im Kontext ihrer Entstehung (primärer Nutzen) sowie die Datenanalyse mit statistischen Verfahren für historische Forschungen (sekundärer Nutzen).

Es geht also bei der Frage, ob Normdaten genutzt werden sollen oder nicht, zunächst weniger um die Diskussion grundlegender, etablierter Methoden der Erschließung als vielmehr um eine differenzierte Perspektive auf die Art und Weise der Datenerfassung. Der Zugang zur Normdatei selbst ist in mehrerer Hinsicht voraussetzungsvoll und beginnt etwa bei den eingesetzten Systemen zur Datenerfassung oder aber dem Verständnis darüber, für welche Entitäten, etwa Personen oder Körperschaften, Datensät-

6 *http://de.wikipedia.org/wiki/Rudolf_Schmidt_%28Journalist%29 (31.03.2015).*

7 *http://www.deutsche-biographie.de/pnd117514608.html?anchor=index (31.03.2015).*

8 *http://kalliope-verbund.info/de/eac?eac.id=117514608 (31.03.2015).*

9 *https://de.wikipedia.org/wiki/Wikipedia:BEACON (31.03.2015).*

10 *http://www.w3.org/TR/sparql11-query/ (31.03.2015).*

11 *Wenn nicht explizit erwähnt, wird unter Metadaten auch das Findbuch subsummiert. Es ist letztlich eine Metainformation über einen Bestand und seine Teile.*

ze in der Normdatei angelegt und gepflegt werden. Es ist die weitverbreitete Annahme, dass dies nur Autoren publizierter Werke vorbehalten ist – eine Annahme, die nicht aufrechterhalten werden kann.

Die Gemeinsame Normdatei

Als Gemeinsame Normdatei (GND) wird ein Dienst der Deutschen Nationalbibliothek (DNB) bezeichnet. Dieser Dienst umfasst eine bei der DNB gehostete Datenbank (Technikebene), ein kooperativ überregional geführtes Redaktionswesen (Qualitätssicherung) sowie eine GND-Arbeitsstelle in der DNB (Koordinierung und Betreuung). Die GND vereinigt in sich die seit den 1970er Jahren entstandenen Dateien für Körperschaften (Gemeinsame Körperschaftsdatei, GKD), Personen (Personennormdatei, PND) und Schlagworte (Schlagwortnormdatei, SWD). Diese werden heute in einer Datei und auf Grundlage eines gemeinsamen internationalen Regelwerks, den Resource Description and Access (RDA), weitergeführt (vgl. Behrens-Neumann 2012, 96). Mit der Ablösung der Regeln zur Alphabetischen Katalogisierung (RAK) durch das neue Regelwerk RDA erfolgte nach langer Vorbereitung 2012 die Zusammenführung der drei Dateien in der GND mit sieben Teildatenbeständen:

Teildatenbestand	Typ	Quantität, Stand 4.03.2015
Personennamen, nicht individualisiert	Tn	4.856.857
Personen, individualisiert	Tp	3.696.006
Körperschaften	Tk	1.274.361
Kongresse und Veranstaltungen	Tv	615.101
Geographika	Tg	312.581
Sachbegriffe	Ts	205.242
Werke	Tw	216.829

Tabelle 1: Teildatenbestände der GND.

Die Intention für den Aufbau der Normdateien war die Unterstützung der Katalogisierung durch regelkonforme Ansetzungen einzelner Entitäten wie Personen und Körperschaften. Im Fall der PND, deren Entwicklung mit Mitteln der Deutschen Forschungsgemeinschaft (DFG) erfolgte, war die „vorrangige Zweckbestimmung, die Ansetzung von Personennamen aus der Zeit vor 1850“ um somit die „Altbestandskatalogisierung zu beschleunigen und zu harmonisieren“ (Rinn 1995, 617). Die Autographen- und Nachlasserschließung ist explizit Bestandteil der Zielsetzung (vgl. ebd. 618). Trotz dokumentierter Vorbehalte gegenüber der PND selbst innerhalb des deutschen Bibliothekswesens, konnte diese sich schließlich erfolgreich etablieren (vgl. Fabian 1995, 6).

Über lange Zeit waren die Normdateien der DNB Hilfsmittel und keine Selbstverständlichkeit für die Katalogisierung. Alle drei genannten Dateien sind historisch „zu unterschiedlichen Zeiten als Hilfsdateien der Formal- und Inhaltserschließung“ gewachsen (Hengel-Dittrich 2010, 35). Hinzu kam, dass häufig eine normierte, regelkonforme Ansetzung und ggf. Individualisierung etwa von Personen als hinreichend für die Bezeichnung als Normdatensatz galten, ohne aber aus einer Normdatei übernommen zu sein. Einzelne Einrichtungen gingen soweit, sogenannte lokale Normdateien aufzubauen. Entscheidend für einen Normdatensatz ist jedoch neben einem gemeinsamen Regelwerk für die Datenerfassung der Fakt, dass die Aufnahme der Entität in einer überregional geführten Normdatei – im deutschsprachigen Raum die GND – erfolgt.

Die Notwendigkeit für das Vorliegen beider Bedingungen bzw. die Definition beider Bedingung als Voraussetzung für einen Normdatensatz – Regelwerk und überregionale Datei – liegt darin begründet, dass nur die Datei einen eindeutigen Identifier für einen Normdatensatz unabhängig von einem bestimmten Anwendungsfall und einer speziellen Anwendung bietet. Dieser Identifier kann überregionale Gültigkeit durch Verbindlichkeit und Persistenz beanspruchen, d. h., selbst wenn ein Datensatz für eine Entität gelöscht oder durch Dublettenbereinigung umgelenkt wird, bleibt der Identifier erhalten und weist bei der Dublettenbereinigung auf den neuen Datensatz. Die überregionale Gültigkeit wiederum erst ermöglicht die Vernetzung von Daten unabhängig von Provenienz und Kontext, indem Entitäten mit diesem Identifier ausgezeichnet bzw. relational verknüpft sind. Nur durch sie sind Entitäten für Maschinen eindeutig zu identifizieren und Informationsangebote für Menschen zu gestalten. Die Identifier der GND finden wir in Katalogen und Verzeichnissen, Lexika oder Editionen wie z. B. Berliner Intellektuelle um 1800-1830[12].

Im Ergebnis dieser Darstellung und auch aus ökonomischen Gründen wäre nur eine einzige Normdatei vorstellbar. Da die Normdatenstrukturen zumeist in nationalen Zusammenhängen mit unterschiedlichen Regelwerken und Formaten gewachsen sind, kennen wir jedoch eine Vielzahl von Normdateien. Sie werden über einen Dienst des Online Computer Library Center (OCLC), in der Virtual Integrated Authority File (VIAF) aggregiert und abgeglichen[13]. Ausschlaggebend für die Qualität und Quantität, das heißt für die Pflege und Neuaufnahme des Normdatenbestandes, sind die Normdateien einschließlich ihrer Redaktionsstrukturen, die wie bei der DNB oder der Library of Congress (LoC)[14] kooperativ dezentral or-

12 http://tei.ibi.hu-berlin.de/berliner-intellektuelle/ (31.03.2015).

13 http://viaf.org/ (31.03.2015).

14 http://www.loc.gov/aba/pcc/ (31.03.2015).

ganisiert sind. Durch einen lesenden Zugriff ist die Nachnutzung bereits erfasster Entitäten möglich; durch einen schreibenden Zugriff wird die Datenbasis erweitert – auch etwa um Personen, die selten und nur lokal vorkommen. Neuere Entwicklungen bei der GND zielen darauf ab, die Möglichkeiten zu erweitern, online über Maschinenschnittstellen lesend und schreibend auf die Normdatei zuzugreifen (vgl. etwa das Projekt SCUSI[15]), oder über ein Online-Formular einen schreibenden Zugriff auf die GND einzurichten. Die Arbeiten an online verfügbaren Schnittstellen konzentrieren sich insbesondere auf den Teildatenbestand für die individualisierten Personen. Sukzessive werden diese Dienste auch auf die weiteren Teildatenbestände ausgedehnt.

Mit dem Zusammenlegen der drei Normdateien PND, GKD und SWD in der GND wurde zudem ein wegweisendes selbstreferenzielles System aufgebaut. Selbstreferenziell heißt, dass zur Erfassung von Merkmalen einer Entität, wiederum Entitäten der GND selbst Verwendung finden. Am Beispiel von Personendaten kann dies einfach verdeutlicht werden:

- Geburts-, Wirkungs- und Sterbeorte werden mit Entitäten des Teildatenbestands Geographika verknüpft
- Berufs- und Funktionsangaben werden mit Entitäten des Teildatenbestands Sachbegriffe verknüpft
- familiäre Beziehungen werden mit Entitäten des Teildatenbestands individualisierte Personen verknüpft
- Affiliationen, das sind Beziehungen zu einer Körperschaft, werden mit Entitäten des Teildatenbestands Körperschaften verknüpft

Durch diese Selbstreferenz finden für die gleichen Ausprägungen von Merkmalen dieselben Entitäten Verwendung, was wiederum die Möglichkeit etwa für künftige mehrsprachige Angebote der Normdatei bietet.

Eindeutigkeit von Entitäten und Datenqualität

Die Bezeichnung Datenqualität bzw. die Aussage darüber, ob Daten eine geringere oder höhere Qualität haben, ist zunächst sehr frei und liegt im Auge des Betrachters; denn sie ist abhängig vom Anwendungsfall, in dem die Nutzung von Daten vorgestellt wird. Den Anwendungsfall als Maßstab für Datenqualität zu nehmen, hat zwei charmante Seiten: eine empirische und eine ökonomische. Empirisch deswegen, weil uns die Kenntnis über den Anwendungsfall, für den Daten erstellt wurden, dabei hilft, zu erklären, warum Daten über eine bestimmte Qualität verfügen. Ökonomisch, weil der Anwendungsfall als Maßstab dafür gelten kann, welche Daten in welchem Umfang über eine Einheit erfasst werden müssen, um dem Anwendungsfall zu genügen. Die Erfassung von Daten kann in Zeit und somit monetär ausgedrückt werden. Je umfassender die Erfassung von Daten, desto mehr Zeit wird je Einheit benötigt.

Diese recht simple Aussage kann an einem Beispiel verdeutlicht werden: So kann noch immer die Feststellung, dass der Datenbestand der Zentralkartei der Autographen, ZKA, über eine hohe Qualität verfügte, Bestand behalten: Ihr Zweck war es, Forschung mit einem zentralen Instrument zu helfen, den Standort von Autographen von Personen in den vielen Bibliotheken, Archiven und Museen zu ermitteln und somit eine Übersicht über die Überlieferungssituation zu gewinnen. Für dieses Ziel erfolgte von 1966 an mit Unterstützung der DFG der Aufbau der ZKA, ein alphabetischer Katalog mit über 1,2 Millionen Karteikarten zum Ende der 1990er Jahre. Die Karteikarten, die von Bibliotheken, Archiven und Museen in Abständen an die Arbeitsstelle in der Staatsbibliothek zu Berlin (SBB) geschickt oder durch Informationen aus Zeitungsartikeln und sonstigen Quellen erstellt wurden, enthielten meist sehr rudimentäre Angaben:

- Autor (ggf. mit Lebensdaten) (Merkmal für die alphabetische Einordnung in die Kartei)
- Signatur (Merkmal für die Bestellung vor Ort)
- Materialart, meist Brief oder Manuskript (Merkmal für die Relevanzbewertung)
- Titel, wenn es sich um Manuskripte handelte (Merkmal für die Relevanzbewertung)
- Entstehungsdatum und -ort (Merkmal für die Relevanzbewertung)

Mit Eingang der Karten bei der Arbeitsstelle der ZKA wurden diese mit einem Sigel gestempelt. Im Fall von Briefen erhielten Adressaten Verweisungskarten, die an der jeweiligen Stelle im Katalog alphabetisch eingeordnet wurden.

Ein einziges Merkmal war für die Einordnung der Karten in die Kartei im Unterschied zur lokalen Kartei in der bereitstellenden Bibliothek, dem Archiv oder dem Museum zusätzlich erforderlich: das Anbringen des Sigels. Nur mithilfe des Sigels ließ sich eindeutig und unabhängig vom Spezialwissen über die Signaturenschemata der Vielzahl von Gedächtnisinstitutionen allein in der Bundesrepublik Deutschland gewährleisten, dass die Aufbewahrungsorte schnell identifiziert werden konnten. Mit der Kartei gelang es somit, effizient die an sie gestellte Aufgabe zu erfüllen. Mit der Konversion der ZKA und dem Aufbau des Kalliope-Verbundes zeigten sich jedoch besondere Herausforderungen. Die Daten auf den Karteikarten waren für die Datenerfassung nicht immer ohne Autopsie der – nicht vorliegenden – Quellen eindeutig zuzuordnen. Zudem war für die 180.000 identifizierten Personen auf den 1,2 Millionen Karteikarten eine nach-

15 http://in2n.de (31.03.2015).

haltige Datenerfassung zu finden, die letztlich in der PND erfolgte. Ein weiteres Problem bei der Weiterentwicklung der Verbunddatenbank Kalliope konnte weder durch Regelwerke noch Datenformate gelöst werden: Dubletten im Datenbestand für Personen und Körperschaften; denn selbst dann, wenn Daten für den Import in die Datenbank regelwerkskonform erfasst und in dem vereinbarten Datenformat bereitgestellt wurden, gelang ein Import nur, wenn auch eine Normdatensatznummer mitgeliefert wurde. Anderenfalls würden wiederholt neue Datensätze für ein und dieselbe Entität mit jedem Datensatz, jeder Datenlieferung und jedem Datenlieferanten erstellt werden – eine Herausforderung, die in der analogen Welt der Zettelkataloge und Findbücher nicht bestand; die Ein- und Zuordnung von Entitäten erfolgte durch Menschen.

Selbiges gilt schließlich für alle Findmittel: Durch den Verlust des unmittelbaren räumlichen Bezugs von Findmitteln steigen die Anforderungen an die Datenqualität. Personen, die im Findbuch vor Ort im Archiv eindeutig durch den räumlichen Kontext identifiziert werden können, sind in überregionalen Katalogen und Portalen nicht ohne weitere Hilfsmittel eindeutig. Schon aus ökonomischen Gründen ist es daher kaum förderlich, als Qualität von Daten ein besonders elaboriertes Datenmodell mit einer Vielzahl von nur noch schwer zu differenzierenden Datenfeldern anzusehen. Dies gilt vor allem auch, weil der primäre Anwendungsfall für die Erschließung noch immer das Suchen und Identifizieren von Quellen für ein bestimmtes Informationsbedürfnisses ist und sicherlich auch bleibt. An diesem Anwendungsfall müssen sich Art und Umfang der Beschreibung rechtfertigen lassen. Der Maßstab für die Datenqualität ist die Disambiguierung der in der Verzeichnungseinheit genannten Entitäten sowohl für Mensch als auch Maschine; denn werden Daten überregional, das heißt außerhalb des direkten räumlichen Bezugs zu den beschriebenen Quellen und gemeinsam mit den Daten aus weiteren Institutionen erfasst oder aggregiert wie etwa beim Kalliope-Verbund oder dem Europäischen Archivportal, ist es erforderlich, dass zwischen Rudolf Schmidt, dem 1875 in Dillingen geborenen Journalisten in Eberswalde, und etwa Rudolf Schmidt, dem 1886 in Berlin geborenen General, eindeutig unterschieden werden kann.

Erschließung von Archivbeständen mithilfe der Normdatei

Um dieses Ziel zu erreichen, ist eine wichtige Anforderung, dass die traditionelle Erstellung von Indizes für Findbücher fortgeführt wird. Die Volltextsuche, so verführerisch einfach und praktisch sie in der eigenen Datenbank vor Ort erscheint, stößt, wie bereits argumentiert, in größeren Kontexten schnell an ihre Grenze (s. a. Krauth 2015, 8). Die wenige eingesparte Zeit bei der Datenerfassung erhöht die Kosten für die Recherche exponentiell mit der anwachsenden Datenmenge. Doch Indizes wie in den analogen bzw. gedruckten Findbüchern sind ebenfalls noch nicht hinreichend. Vielmehr ist jedem Indexbegriff mindestens ein eindeutiger Identifier für die Entität mitzugeben. Dieser Identifier muss auch außerhalb der lokalen Datenbank verstanden werden, das heißt, dass der Identifier gegen eine allgemein anerkannte Normdatei referenziert und somit die Entität eindeutig identifiziert werden kann. Diese Datei ist im deutschsprachigen Raum, das heißt auch für Österreich und die Schweiz, die GND.

Wie schon gezeigt, erlaubt die Normdatei, Entitäten mit zusätzlichen Merkmalen auszustatten, die ihre eindeutige Identifizierung ermöglichen. Im Fall von individualisierten Personen, Tp-Sätze in der GND, sind dies etwa:

- Geschlechtsangabe (nicht jeder Name lässt einen Rückschluss auf das Geschlecht zu)
- Verweisungsformen, das sind etwa die in Quellen gefundenen abweichenden Namensformen zur Ansetzungsform
- Pseudonyme
- Geburts- und Sterbedaten
- Geburts- und Sterbeorte
- Nennung von Berufen und Funktionen
- Nennung von familiären Beziehungen
- Nennung von Beziehungen mit Körperschaften
- ergänzende biographische Hinweise
- Nennung der Quellen für die Angaben

Diese Angaben sollen, mit Ausnahme der Lebensdaten, nicht in der Verzeichnungseinheit und schon gar nicht als Fließtext, sondern in der GND für die kontextunabhängige Nachnutzung erfasst werden. Für die Verzeichnungseinheit selbst sind wenige Angaben hinreichend, das sind am Beispiel einer EAD-kodierten Verzeichnungseinheit: normierte Ansetzungsform des Namens (@normal), die Normdatensatznummer (@authfilenumber), Quelle der Normdatensatznummer (@source) sowie nach Möglichkeit die Funktion bzw. die Rolle, in der die Person im Kontext der Quelle genannt ist (@role). Dabei sind alle Attributwerte aus den normierten Dateien oder kontrollierten Vokabularien zu übernehmen (z. B. die Relator Codes[16] für Funktionen).

In dem in Abbildung 1 gezeigten Beispiel sind die folgenden Entitäten mittels normierter Referenzsysteme disambiguiert:

- die **Verzeichnungseinheit (./c)** selbst mit einem eindeutigen Identifier des Datensatzes, zusammengesetzt aus einem Identifier für die Herkunftsdatenbank

16 *http://www.loc.gov/marc/relators/relaterm.html (31.03.2015).*

und dem Identifier des Datensatzes in der Datenbank (z.B. hier die ISIL des Kalliope-Verbundes, ein Namensraum, ein interner Identifier: DE-611-HS-20156)

- die **Bestandshaltende Institution (./repository)** mit dem International Identifier for Libraries (ISIL), der auch für Archive und Museen Verwendung findet und bei der nationalen ISIL-Agentur bei der Staatsbibliothek zu Berlin beantragt werden kann[17]
- die **Personen (./controlaccess/persname)**, die im Index des Findbuchs aufgenommen sind
- die **Orte (./controlaccess/geogname)**, die im Index des Findbuchs aufgenommen sind

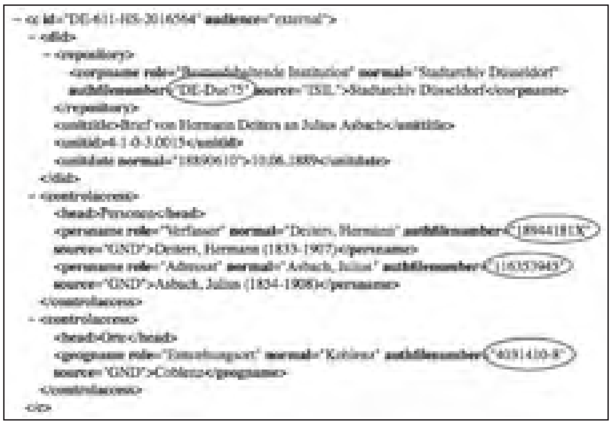

Abbildung 1: Verzeichnungseinheit in EAD.

In jedem Fall soll die Quelle (@source) des Referenzsystems etwa ISIL oder GND genannt werden, um die Identifier (@authfilenumber) auch maschinell eindeutig zuordnen zu können. Für Personen werden in diesem Beispiel die Lebensdaten mit in die Verzeichnungseinheit übernommen, sodass sie unmittelbar im Findbuch angezeigt werden können und für Menschen die Entität schnell und einfach ohne weitere Klicks zu identifizieren sind.

Weitere Entitäten, für die eindeutige Referenzsysteme, mindestens aber kontrollierte Vokabularien existieren, sind:

- Sachschlagworte (GND)
- Gattungen und Materialarten (GND)
- Sprachen (ISO 639-2)
- Länder (ISO 3166-1)

Die kommende Version des internationalen Formatstandards Encoded Archival Description (EAD) sieht vor, dass auch Umfangsangaben normiert, mindestens aber normalisiert mit definierten Maßeinheiten erfasst werden können.

17 http://sigel.staatsbibliothek-berlin.de (31.03.2015).

Zugang zur Normdatei

Die Normdatei wird unter der CC0-Lizenz[18] kostenfrei von der Deutschen Nationalbibliothek (DNB) zur Verfügung gestellt. Auf den aktuellen Datenbestand kann über das Portal der DNB oder aber über Maschinenschnittstellen (z. B. Z39.50, SRU, OAI) zugegriffen werden. Jedoch ist es nicht nur das Ziel, Daten aus der Normdatei zu übernehmen, sondern auch zur Normdatei beizutragen, indem fehlende Daten ergänzt, falsche Angaben korrigiert oder neue Datensätze für fehlende Entitäten angelegt werden. Die Normdatei ist dementsprechend eine Kooperative mit Redaktionsstellen in den teilnehmenden Bibliotheken und Bibliotheksverbünden. Der Kalliope-Verbund nahm von Beginn an redaktionell an der GND teil. Es ist jedoch ein bekanntes Problem, dass Institutionen, die nicht über einen Bibliotheksverbund direkt an die Normdatei angebunden sind, bisher zwar über einen lesenden – mindestens über das Portal der DNB –, aber über keinen direkt schreibenden Zugriff auf die GND verfügen. Im Rahmen des von der DFG geförderten Projektes IN2N – Institutionenübergreifende Integration von Normdaten – wird an Schnittstellen und Verfahren zur kooperativen Normdatennutzung und -pflege für Archive und Museen gearbeitet[19]. So gelang es im Rahmen dieses Projektes, die Regisseure, Darsteller, Drehbuchautoren, Produzenten etc. des Filmportals[20] des Deutschen Filminstituts (DIF) in die GND aufzunehmen. Über die Schnittstelle SCUSI verfügt das Filminstitut nunmehr über eine direkte Maschinenschnittstelle zur GND. Ebenfalls arbeitet die DNB an der Entwicklung eines Online-Formulars, mit dessen Hilfe zunächst Personendaten neu in die GND aufgenommen werden können – ohne intensive Einführungen in Regelwerke und Formate. Funktionen für die redaktionelle Bearbeitung bestehender Personennormdatensätze oder die Erfassung von Daten in weiteren Normdatensegmenten wie Körperschaften folgen zu einem späteren Zeitpunkt.

Mehrwert der Normdatei

Es klang bereits an, dass ein wesentliches Ziel, das mit der Normdatei verfolgt wird, die Disambiguierung von Entitäten und natürlich die Vermeidung von doppelter Erfassungsarbeit durch Nachnutzung derselben in unterschiedlichen Kontexten ist. Das Bedürfnis, Entitäten eindeutig zu identifizieren, ist mit der Entwicklung des Internets, speziell der Idee des Semantic Web verbunden, die wiederum das Bedürfnis nach Struktur durch Systematisierung im sonst recht unübersichtlichen World Wide Web zum Ausdruck bringt. Für Maschinen wird mithilfe der Normdatei nicht nur deutlich, dass ein Textstring eine

18 http://www.dnb.de/DE/Standardisierung/GND/gnd_node.html (31.03.2015).
19 http://in2n.de (31.03.2015).
20 http://www.filmportal.de (31.03.2015).

Entität ist, sondern diese Entität eine Person und diese Person eine ganz bestimmte Person ist, mit der Daten über Bilder, Briefe, Nachlässe, Manuskripte verknüpft sind – ob als Fotograf, Adressat, Bestandsbildner oder Autor. Die Kenntnis über die Entität ermöglicht es, die Daten unabhängig vom Kontext zu vernetzen und auf dieser Grundlage neue Dienste für Forschende zu entwickeln. Eine denkbar einfache Variante ist die Verlinkung von Online-Ressourcen etwa mithilfe von so genannten BEACON-Dateien. Hierbei handelt es sich um schlichte Textdateien, die eine Liste der Identifier der GND enthalten, die von einem Online-Dienst wie Wikipedia, Bibliothekskataloge, Zentrale Datenbank Nachlässe, Kalliope-Verbund etc. verwendet werden. Über wenige Metadaten im Kopf der Datei werden Kurzangaben zum Dienst und zur URL mitgegeben, über welche die Identifier aufgelöst werden können. Auf dieser Grundlage ist es etwa der Deutschen Biographie möglich, Lexikoneinträge mit einer Vielzahl an unterschiedlichen und voneinander unabhängigen Informations- und Datendiensten zu verbinden[21].

Anspruchsvollere Ansätze sind die Extraktion von Informationen über historische soziale Netzwerke. So initiierte das Institute for the Advanced Technologies in the Humanities 2010 das Projekt Social Networks and Archival Context (SNAC), aus dem ein Prototyp für ein Archivportal hervorgegangen ist[22]. In diesem Projekt wurde eine beachtliche Anzahl von Findbüchern, die in maschinenlesbaren Formaten vorliegen, ausgewertet. Entitäten (Personen und Körperschaften) wurden extrahiert, in XML auf Basis des internationalen Formatstandards Encoded Archvial Context – Corporate Bodies, Persons, Families (EAC-CPF)[23] je Entität kodiert und diese schließlich mithilfe von Algorithmen unter Einbeziehung von Normdateien wie die der Library of Congress und der Virtual International Authority File (VIAF) versucht, zu disambiguieren. In diesem sogenannten „Match and Merge"-Verfahren werden Merkmale über eine Entität strukturiert in den EAC-CPF kodierten Dateien gespeichert. Dies sind neben den klassischen identifizierenden Merkmalen wie Lebensdaten und Berufsangaben auch Beziehungen zu Archivbeständen, Sekundärliteratur, aber auch zu weiteren Entitäten, z. B. Korrespondenzpartner oder allgemein assoziierte Personen und Organisationen. Der entwickelte Algorithmus arbeitet u.a. mit der Annahme, dass die unter einem Bestandsbildner, das heißt in einem Findbuch aufgeführten Personen und Organisationen mit der Person des Bestandsbildners in einer Beziehung stehen. Diese Beziehungen lassen sich als Graphen wiederum visualisieren.

Einen vergleichbaren Ansatz verfolgt der Kalliope-Verbund mit dem Online-Katalog: Zunächst werden die in EAD-kodierten Findbücher und Autographenkataloge indiziert. Die beschriebenen Korrespondenzen werden ausgewertet und die daraus resultierenden Beziehungsinformationen in einem Graph visualisiert. Der Graph eines Korrespondenznetzwerks in Abbildung 2 beruht auf Indexeinträgen mit GND-Identifier (Personen und Organisationen) für Verzeichnungseinheiten in Findbüchern für Nachlässe, Verlagsarchive oder Autographenkatalogen. Die Kodierung in EAD entspricht der Abbildung 1. Im Kontext des Katalogs dient die Visualisierung Forschenden, Querbeziehungen zwischen den Beständen einfacher als bisher erkennen zu können.

Erschließungsdaten für die historische Forschung

Das Potenzial der Erschließungsdaten reicht jedoch über diesen Ansatz der Visualisierung als flankierendes Rechercheinstrument hinaus: So können mit Standarddaten für Briefe zeitliche (Entstehungsdatum) und räumliche (Entstehungsort) Verteilungen in die Visualisierung einbezogen werden. Werden die Normdaten einbezogen, lassen sich die Graphen zudem statistisch beschreiben und Varianzen über Zeit werden sichtbar. Für die statistische Beschreibung steht z. B. der Modus für die Geschlechtsverteilung (wie viele Männer, wie viele Frauen gehören einem Netzwerk an), für die Altersverteilung oder auch für die Verteilung von Berufen zur Verfügung[24]. Die Erschließung wandelt sich vor diesem Hintergrund von der reinen Datenerfassung zu einer Art Datenerhebung. Für Forschende standen diese Daten und somit die Methodik der sozialen Netzwerkanalyse bisher nicht bzw. nur sehr eingeschränkt zur Verfügung; die Erhebung zumindest von umfangreichen Daten im Rahmen eines Forschungsprojekts ist nahezu prohibitiv teuer. So müssten die relevanten Bestände in den Bibliotheken, Archiven und Museen identifiziert und die Daten mühsam erfasst werden. Dagegen ist es möglich, ohne Eingriff in die Methodik der Erschließung und ohne die Einführung neuer Datenelemente, sondern einzig und allein auf der Grundlage normierter und disambiguierter Entitäten relevante Daten für Forschungsprojekte zu erarbeiten. Es muss an dieser Stelle betont werden, dass die Netzwerkanalyse sicherlich nicht das einzige Anwendungsszenario für statistische Verfahren ist; vielmehr ist davon auszugehen, dass bei genauer Untersuchung eine Vielzahl von Möglichkeiten für quantifizierte Analysen mithilfe von Erschließungsdaten denkbar und möglich sind – auch für institutionelle Überlieferungen, sodass neue Perspek-

21 http://www.deutsche-biographie.de (31.03.2015).

22 http://socialarchive.iath.virginia.edu (31.03.2015).

23 http://eac.staatsbibliothek-berlin.de, s. auch zu EAC-CPF: Sonderausgabe des Journal of Archival Organization: Identity Matters. Describing and Interconnecting with EAC-CPF, 2014.

24 Zur Netzwerkanalyse s. etwa Jansen, Dorothea (2003): Einführung in die Netzwerkanalyse. Grundlagen, Methoden, Forschungsbeispiele. Opladen; historische Forschung, bspw. Medizingeschichte: Fangerau, Heiner (2010): Spinning the scientific web. Jacques Loeb (1859-1924) und sein Programm einer internationalen biomedizinischen Grundlagenforschung. Berlin.

tiven auf historische Ereignisse und Abläufe gewonnen werden können. Dass die Analyse historischer sozialer Netzwerke eine bedeutende Methodik für die historische Forschung ist, zeigt eine Reihe von jüngeren Projekten:

- Mapping the Republic of Letters (Stanford University)[25]
- Networking the Republic of Letters (University of Oxford)[26]
- Vernetzte Korrespondenzen (Institut für Informatik der Martin-Luther-Universität, Halle)[27]
- Visualisierung von Beziehungen in der Deutschen Biographie[28]

Um die für statistische Verfahren erforderliche Qualität der Daten zu erzeugen, sind jedoch moderne Instrumente für die Erfassung von Daten erforderlich, das heißt ein direkter Zugang zur Gemeinsamen Normdatei einschließlich einer schreibenden, redaktionell betreuten Schnittstelle sowie die Nutzung kontrollierter Vokabularien. Das Potenzial für Archive im digitalen, vernetzten Zeitalter ist sehr weitreichend. Mithilfe neuer Werkzeuge für die Erschließung stehen zugleich Optionen für die Weiterentwicklung von Abläufen und Dienstleistungen etwa in der Bildungs- und Forschungszusammenarbeit offen. Ausgangspunkt für diese Dienste sind die etablierten Prinzipien der Überlieferungsbildung, Ordnung und Erschließung der Quellen (s. Brenneke 1953, 25 ff.). Hinzu tritt jedoch ein Wechsel der Perspektive auf die Nutzungsmöglichkeit von Daten resppektive die Entwicklung von Werkzeugen zur Erhöhung der Datenqualität vor dem Hintergrund neuer Anwendungsfälle.

Letztlich ist der Wunsch zu einer stärkeren Strukturierung insbesondere der kontextsensitiven Daten zu Bestandsbildnern durch archivische Normdaten nicht neu und kann zunächst helfen, verstreute Bestände einer Provenienzstelle zusammenzuführen (vgl. Brübach 2008, 8). Ob aber der Aufbau einer Normdatei für Archive erfor-

25 *http://republicofletters.stanford.edu (31.03.2015).*
26 *http://www.culturesofknowledge.org (31.03.2015).*
27 *http://www.informatik.uni-halle.de/ti/forschung/ehumanities/neli/ (31.03.2015).*
28 *http://www.deutsche-biographie.de/ueber (31.03.2015).*

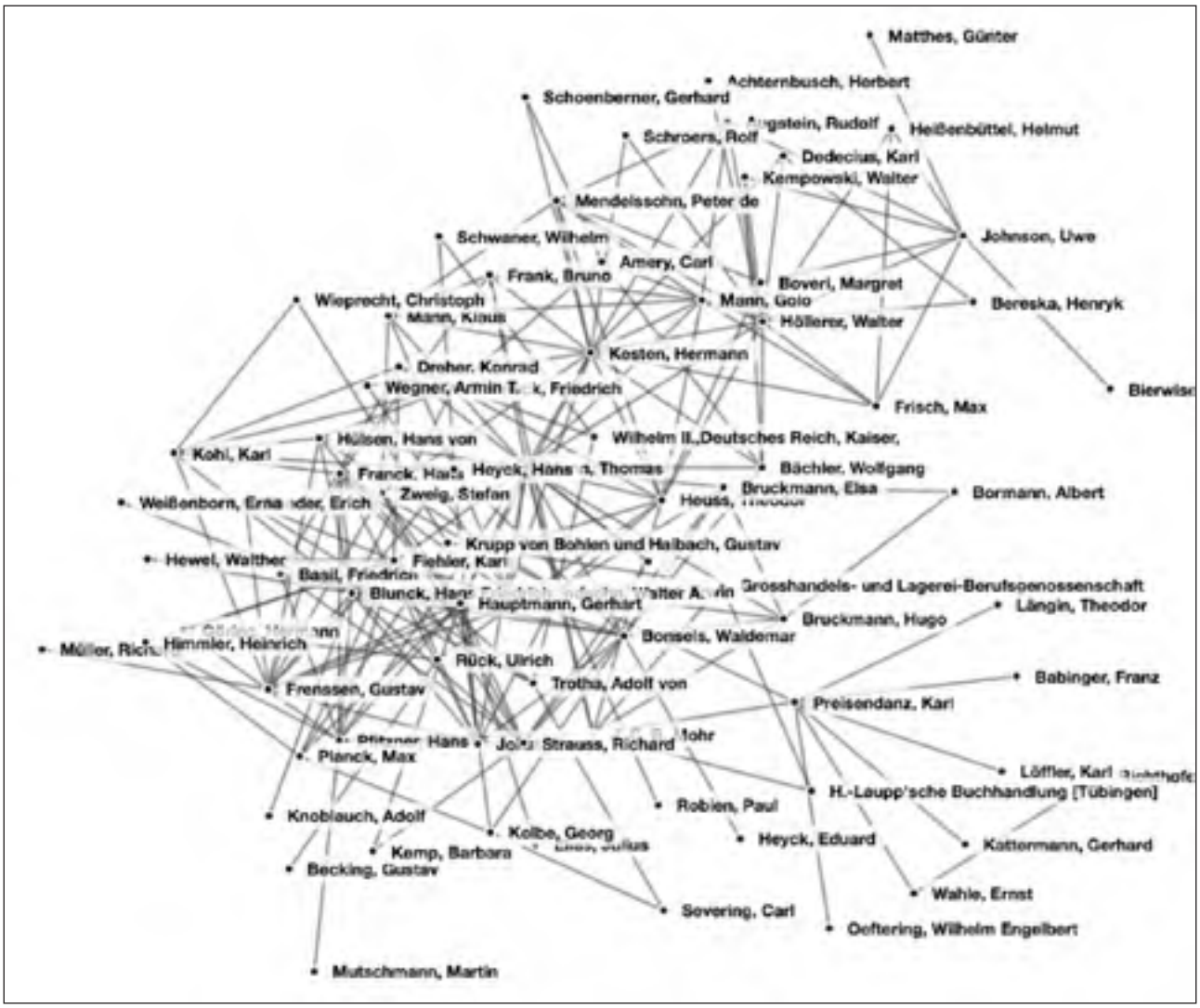

Abbildung 2: Beispiel eines Korrespondenznetzwerks.

derlich ist, wie dies etwa internationale Regelwerke wie die International Standard Archival Authority Record for Corporate Bodies, Persons, and Families (ISAAR-CPF)[29] nahelegen, ist streng zu prüfen. Vielmehr sollte die fachliche Nachfrage nach überregionalen Diensten, die aus einem hier skizzierten modifizierten Anforderungsprofil resultieren, hinsichtlich der Option zur Mitnutzung existierender Instrumente geprüft werden. Das Beispiel der ZKA respektive des Aufbaus des Kalliope-Verbundes zum überregionalen Verbund und zum nationalen Nachweisinstrument für Nachlässe und Autographensammlungen sollte vor allem verdeutlichen, dass Regelkonformität und Datenformate noch keine hinreichenden Bedingungen für die Interoperabilität von Daten sind. Dies erfordert einen Identifier, der unabhängig von einer Datenbankanwendung Gültigkeit beansprucht bzw. aufgrund gemeinsamer sozialer Übereinkunft Gültigkeit beanspruchen kann. Die gilt heute für die GND, die auch außerhalb der Bibliothekswelt breite Nutzung findet – Wikipedia, ADB/NDB, ZDN etc. – und dies resultiert aus ihrer Verbindlichkeit und Persistenz.

Fazit

Die Bedeutung der Normdatei liegt in ihrer Entwicklung zu einem Instrument für das Web of Data. Die Normdatei ist entscheidend für die überregionale eindeutige Identifikation von Informations- und Datenangeboten zu einer Entität. Vorrangig im Bereich der Erschließung von Nachlässen konnte sich die Normdatennutzung bereits als fruchtbar zeigen und ist nicht mehr wegzudenken. Dabei ist es zweitrangig, ob Personen von lokaler, regionaler oder überregionaler Bedeutung sind. Die Chancen der Disambiguierung von Entitäten für die Strukturierung der Datenmengen und für die damit verbundenen Chancen zur Vernetzung der Daten verschiedener Angebote, der Visualisierung oder gar der Nutzung der Daten für historische Forschungsmethoden, die bisher nicht zur Verfügung standen, liegen in modernen und attraktiven Diensten und Dienstleistungen für verschiedene Nutzergruppen. Diese Dienste können dazu beitragen, dass der Zugang zur Geschichte über die sozialen Beziehungen neue Perspektiven auf die Quellen und damit die historischen Ereignisse hervorbringt. Durch neuere Möglichkeiten bzw. Methoden der Präsentation von Informations- und Datenangeboten kann es ebenfalls gelingen, an Attraktivität und Aufmerksamkeit für unsere Geschichte zu gewinnen. In jedem Fall kann mithilfe von Normdaten sichergestellt werden, den gesuchten Rudolf Schmidt eindeutig zu identifizieren und den Aufwand für die Suche zu reduzieren.

29 *http://www.ica.org/10203/standards/isaar-cpf-international-standard-archival-authority-record-for-corporate-bodies-persons-and-families-2nd-edition.html (31.03.2015).*

Literaturverzeichnis

Arndt, Irina: Der Weg zum Wissen. Einführung eines Discovery Systems in fünf Max-Planck-Bibliotheken. In: Forschungsbericht 2013. Max Planck Digital Library. http://www.mpg.de/6708968/JB_2013.

Behrens-Neumann, Renate (2012): Das Projekt Gemeinsame Normdatei (GND). In: Zeitschrift für Bibliothekswesen und Bibliographie. 59, 2. 96–99.

Brenneke, Adolf (1953): Archivkunde. Ein Beitrag zur Theorie und Geschichte des europäischen Archivwesens. Leipzig.

Brübach, Nils (2008): Entwicklung von internationalen Erschließungsstandards. Bilanz und Perspektiven. In: Der Archivar. 61, 1. 6–13.

Fabian, Claudia (1995): Entwicklung und Aufbau der Personennamendatei in Deutschland. Bericht über Konzeption und Realisierung seit 1989. In: Zeitschrift für Bibliothekswesen und Bibliographie. 42, 6. 605– 615.

Gradmann, Stefan/Hennicke, Steffen/Olensky, Marlies (2012): Linked Data. In: cms-journal. 35. urn:nbn:de:kobv:11-100200851. 18–22.

Introna, Lucas D./Nissenbaum, Helen: Shaping the Web. Why the Politics of Search Engines Matters. In: The Information Society. 16, 3. 169–185.

Hengel-Dittrich, Christel (2010): Das Projekt Gemeinsame Normdatei. GND. In: Dialog mit Bibliotheken. 22, 1. 35–38.

Krauth, Wolfgang (2015): Archive und Online-Portale. Thesen für den weiteren Erfolg. In: Der Archivar. 68, 1. 6–9.

Rinn, Reinhard: Die überregionale Normdatei für Personennamen. In: Zeitschrift für Bibliothekswesen und Bibliographie. 42, 6. 617–637.

Spinnler-Dürr, Alice (2013): Die Diktatur der Suchmaschinen. In: 027.7. Zeitschrift für Bibliothekskultur. 2, 1. 58–66.

Kontakt

Staatsbibliothek zu Berlin – Preußischer Kulturbesitz
Abteilung Überregionale Bibliographische Dienste
Gerhard Müller
Potsdamer Straße 33, 10785 Berlin
Tel.: 030 266-435119

gerhard.mueller@sbb.spk-berlin.de
http://kalliope.staatsbibliothek-berlin.de/
http://www.staatsbibliothek-berlin.de/