Automat. Normdaten Anreicherung: Erfahrungen mit Archivdaten

# **Ortsnormdaten**

In vielen Fällen enthalten Erschließungsinformationen von Archiven bereits Indexbegriffe zu Orten. Diese sind jedoch nicht normiert erfasst und können daher nicht eindeutig zugeordnet werden. Handelt es sich nun bei Kirchheim um jenes unter Teck oder um das am Neckar? Nur über den Namen lässt sich also keine gute Zuordnung erreichen.

Die Wahrscheinlichkeit einer korrekten Identifizierung lässt sich allerdings signifikant erhöhen, indem die Suche auf einen bestimmten räumlichen Bereich eingegrenzt wird. Gibt es innerhalb eines Gebiets doch meist nur eine begrenzte Anzahl an identen Ortsnamen, weswegen man bei Treffern davon ausgehen kann, dass der richtige Ort identifiziert wurde. Gerade bei kommunalen Archiven ist dies gut möglich, ist doch der Zuständigkeitsbereich primär über die räumliche Zuordnung definiert, die sich mehr oder minder genau abgrenzen lässt.

Dass dabei natürlich nie eine 100% vollständige und korrekte Identifizierung erreicht werden kann, ist klar. Können sich doch etwa die räumlichen Zuständigkeiten eines Archivs im Laufe der Zeit verändert haben, weswegen nicht so einfach ist, geographische administrative Einheiten (Kreis, Stadt, …) als Sucheinstieg zu nutzen. Außerdem kommt es natürlich vor, dass auch Orte außerhalb des Archivsprengels in kommunalen Erschließungsinformationen erwähnt werden, was im schlimmsten Fall nicht nur zu fehlender Identifizierung, sondern sogar zu falschen Zuordnungen führen kann.

So gibt es etwa einen Vorort von Karlsruhe mit dem Namen Ulm. Bei dem in den Erschließungsinformationen des Stadtarchivs Karlsruhe erwähnten Ulm handelt es sich allerdings um die Universitätsstadt an der Donau. Da durch die Eingrenzung des Suchraums auf den Raum Karlsruhe allerdings nur ein Ulm gefunden wurde, würde bei einer automatischen Anreicherung, davon ausgegangen werden, dass es sich um den Vorort von Karlsruhe handeln würde. Das kann vermieden werden, indem die Anreicherungslisten vor der Anreicherung noch einmal durchgesehen werden bzw. automatisiert bedeutende Stadtnamen „black-gelistet“ werden und für diese keine Anreicherung stattfindet.

Zu klären könnte daher in einem Folgeprojekt anhand einiger kommunaler Archive, die mit Ortsindexbegriffen verzeichnen, sein, wie hoch der Prozentsatz an angereicherten Indexbegriffen ist, wobei auch hier nur selten Werte über 20% zu erwarten sind (da etwa auch viele allgemeine Begriffe wie „Stadtgarten“ oder Straßennamen hier angegeben sind) und wie viele davon „false positives“, wie das zuvor erwähnte Ulm, wären. Davon ausgehend wäre zu evaluieren, ob kommunale Archive, trotz eines gewissen Fehler-Quotients, mit einer automatischen Anreicherung ihrer Daten für Portale einverstanden sind, bzw. ob diese Daten auch ins eigene System, idealerweise nach einer manuellen Kontrolle, zurückgespielt werden sollen.

# **Personennormdaten**

Meist noch zentraler als die Identifikation des Orts, welche idealerweise (zusätzlich) auch über Koordinaten erfolgen sollte, ist die eindeutige Identifizierung und Zuordnung von Personen. Um diese eindeutig bestimmen und zuordnen zu können, sind in der Praxis bei den Archivbeständen gerade zwei Aspekte von Bedeutung. Einerseits ein exaktes Geburts- oder Sterbedatum, andererseits Vor- und Familiennamen, unter Umständen noch die Profession.

Gerade in größeren, insbesondere staatlichen Archiven oder Spezialarchiven werden diese Informationen getrennt verzeichnet und sind dadurch leicht strukturiert auszulesen. Diese Bestände eignen sich daher sehr gut für Abgleiche und eine automatische Generierung von Normdaten, wie es auch im Rahmen des Archivportal-D Projekts gezeigt wurde. Idealerweise sollten diese Daten zum Abgleich normiert ausgespielt werden, zusätzlich sollte aber eine Syntax aufgestellt werden, nach der diese über einem standardisierten Abgleich mit der GND angereichert werden können.

Wenn man von Normdaten spricht, ist aber natürlich die automatische Anreicherung nur der erste Schritt, mit der bereits geleistete Erschließungsarbeit Web 3.0 fähig gemacht werden kann. Ein vieler wichtigerer Schritt ist natürlich die Erschließung mit Normdaten auch in der Praxis zu verankern. Hier kommt neben den begrenzten Ressourcen im Archivbereich natürlich die Frage auf, wie auch aktiv an der GND mitgearbeitet werden kann, da eine Vielzahl an für den Archivbereich relevanten Personen, eben nicht Teil der GND ist.

Archive müssen sich in diesem Zusammenhang allerdings ebenfalls überlegen, für welche Personen sie Normdaten anlegen wollen. Nur weil etwa in einer Personalakte eine Person eindeutig identifiziert werden kann, muss es noch keinen Sinn machen, Normdaten für diese Person in die GND einzuspielen. Ist es doch unwahrscheinlich, dass Informationen zu dieser Person auch in anderen Archiven vorhanden sind und über Normdaten verknüpft werden können.

Vielmehr könnte etwa zuvor ein Clustering erfolgen, und festgelegt werden, dass nur für Personen zu denen mehrere eindeutig identifizierbare Einträge in einem Archiv vorhanden sind, Normdaten angelegt werden. Diese Entscheidungsprozesse müssten evaluiert und definiert werden, was im Rahmen eines Folgeprojekts geschehen könnte.

In den meisten kleineren Archiven liegen diese Daten allerdings nicht so strukturiert vor. Vielmehr sind die Namen und Geburts- bzw. Sterbedaten im Titel oder den Enthältvermerken enthalten und müssen aus diesen herausgeparst werden. Da diese in den meisten Fällen durch Marker oder -zeichen wie die folgenden bereits in reguläre Ausdrücke übersetzten Indikatoren

* qr/(\\*)|((geb\.)|(Geburtsdatum)|(geboren))[( am)|(:)]?|(Registraturbildner:)/i;
* qr/(✝)|(\+)|((gefallen)|(gef\.)|(gest\.)|(Todesdatum)|(Sterbedatum)|(gestorben)|(verst\.)|(verstorbenen)|(verstorben))[( am)|(:)]?|(Registraturbildner:[\s\d\.]\*-)/i;

und eine mehr oder minder definierte Zahlenfolge für Datumsangaben vorgegeben sind, können diese in vielen Fällen verhältnismäßig gut extrahiert und auch erweitert werden.

Über die so extrahierten Geburtsdaten kann bereits eine SPARQL-Abfrage an Wikidata/Dbpedia/GND gesendet werden und der umgebende Text nach den erhaltenen Nach- und Vornamen durchsucht werden und so relativ sicher, Personen eindeutig identifiziert werden.

Zentral ist aber natürlich, dass die genauen Geburts- und/oder Sterbetage vorhanden sind. Die ersten Abgleiche haben hier allerdings gezeigt, dass nur eine begrenzte Anzahl an automatisieren Anreicherungen zu erreichen ist. Sind doch nur in den wenigsten Fällen exakte Geburts- oder Sterbedaten vorhanden (selbst Angaben nur über Jahreszahlen sind verhältnismäßig selten). Vielmehr kann dieses Anreicherungsverfahren nur als erster Schritt gesehen werden, um auch kleineren Archiven anhand einzelner Beispiele Normdaten und die damit verbundenen Möglichkeiten näher zu bringen bzw. auch um Kennzeichen für jene Bestände zu extrahieren und definieren, in denen viele Lebensdaten enthalten sind und die sich gut für diese automatischen Anreicherungsprozesse eignen würden.