

## ABI Technik Frage

Fabian Steeg

# Wie kann ich Daten aus einem Digitalisierungsprojekt mit Normdaten anreichern?



<https://doi.org/10.1515/abitech-2022-0012>

**Fabian Steeg ist Softwareentwickler im Bereich Datenverarbeitung und Webentwicklung in der Gruppe Offene Infrastruktur am Hochschulbibliothekszentrum des Landes Nordrhein-Westfalen (hbz). Er ist Open-Source-Contributor im Umfeld von Metafacture, JSON-LD und OpenRefine sowie Co-Chair der W3C Entity Reconciliation Community Group. Er beantwortet die aktuelle ABI-Technik-Frage: Wie kann ich Daten aus einem Digitalisierungsprojekt mit Normdaten anreichern?**

Normdaten spielen speziell im Hinblick auf die Qualität der Inhaltserschließung bibliografischer und archivarischer Ressourcen eine wichtige Rolle. Sie ermöglichen Einheitlichkeit und Eindeutigkeit und so eine bessere Auffindbarkeit der Ressourcen. Daher sollten Daten aus Digitalisierungsprojekten mit Normdaten angereichert werden.

Eine Möglichkeit stellt die Verwendung des Werkzeugs OpenRefine dar. OpenRefine ist – auch im Bibliotheksbereich – eines der meistbenutzten und -bewährten Werkzeuge für Datenbereinigung und -abgleich und wird als *Essential Open Source Software for Science* gefördert.

OpenRefine ist ein generisches tabellenbasiertes Werkzeug, vergleichbar mit klassischen Tabellenkalkulationen wie Excel. Es bietet zahlreiche Funktionalitäten zur Bereinigung und Transformation von Daten, sowie zum Abgleich (Reconciliation) und zur Anreicherung mit externen Datenquellen.

Neben dem in OpenRefine integrierten Wikidata-Service gibt es eine größere Anzahl von Datenquellen, für die ein Reconciliation-Dienst angeboten wird,<sup>1</sup> darunter auch die Gemeinsame Normdatei (GND). Das Protokoll zur Kommunikation mit diesen Diensten kann auch von anderen Clients als OpenRefine verwendet werden (z. B. direkt aus dem Bibliotheksmanagementsystem Alma über *Alma Refine*).

Was bedeutet das nun für die Arbeit in Digitalisierungsprojekten konkret? Der erste Schritt ist hier die manuelle oder automatische Eigennamenerkennung (Named Entity Recognition) in den Digitalisaten, z. B. von Personen oder Orten. Als Ergebnis haben wir Listen von Namen. Diese können in OpenRefine eingelesen werden und mittels

eines Reconciliation-Dienstes mit Normdaten abgeglichen werden. Besonders gute Treffer werden automatisch abgeglichen, andere müssen manuell überprüft werden. Dazu bietet OpenRefine in Abhängigkeit der konkreten Service-Implementierung Unterstützung in der Oberfläche an, für die GND bei Personen etwa in Form von Lebensdaten, Berufen und Abbildungen. Zur Verbesserung des automatischen Abgleichs können zusätzlich zu den Namen disambiguierende Merkmale mitgeschickt werden.

Nach erfolgreichem Abgleich verfügen wir nicht nur über die Identifikatoren aus der jeweiligen Normdatei, sondern können weitere Felder der abgeglichenen Normdatensätze in unseren lokalen Daten anreichern. Schließlich können wir die abgeglichenen, angereicherten Daten in verschiedene Formate exportieren und erhalten so für unsere Digitalisate Metadaten in hoher Qualität.

Um das alles einmal auszuprobieren empfiehlt sich ein Blick in unsere Tutorials zur Reconciliation mit lobid-gnd<sup>2</sup> sowie in die Anleitung von OpenRefine.<sup>3</sup>

## Autoreninformationen

### Fabian Steeg

Hochschulbibliothekszentrum NRW (hbz)  
Offene Infrastruktur  
Jülicher Straße 6  
50674 Köln  
[steeg@hbz-nrw.de](mailto:steeg@hbz-nrw.de)  
[orcid.org/0000-0001-8829-1989](https://orcid.org/0000-0001-8829-1989)



<sup>1</sup> <https://reconciliation-api.github.io/testbench/> (07.12.2021).

<sup>2</sup> <https://lobid.org/gnd/reconcile> (07.12.2021).

<sup>3</sup> <https://docs.openrefine.org/> (07.12.2021).