Datenintegration mit D:SWARM

Werkstattbericht zum Management bibliothekarischer Daten an der SLUB Dresden

von **FELIX LOHMEIER**

Europa fördert Sachsen.
EFRE
Europäischer Fonds für regionale Entwicklung
Europäische Union

Nach desillusionierenden Erfahrungen mit den auf dem Markt erhältlichen Discovery-Systemen und den darin enthaltenen Normalisierungsfunktionen haben wir im Juni 2013 begonnen, eine technisch weit in die Zukunft reichende Vision zu realisieren: Kulturerbe-Institutionen sollen Daten aus unterschiedlichsten Quellsystemen mit einem integrierten Werkzeug auf einfache, intuitive Weise miteinander verknüpfen und anreichern können. Die Datenqualität soll signifikant verbessert werden. Daten werden in einem Graphenformat verarbeitet, der entstehende spezifische Wissensgraph soll als zentrale Datenhaltung für vorhandene Kataloge und neue Präsentationssysteme dienen und gleichzeitig die bibliothekarischen Daten als Linked Open Data für andere Einrichtungen zur Nachnutzung bereitstellen. Alles intuitiv und einfach nutzbar, interoperabel und auf Basis von Open Source-Technologien. Dank der Finanzierung aus Mitteln der Europäischen Union und des Freistaates Sachsen (EFRE) konnten wir an dieser Vision mit zusätzlichen Software-Entwicklern und gemeinsam mit einer auf Big Data spezialisierten Dresdner Firma, Avantgarde Labs GmbH, arbeiten. Dieser kurze Artikel soll über den bis Mai 2015 erreichten Entwicklungsstand und die geplanten nächsten Meilensteine informieren.

Yet another ...?

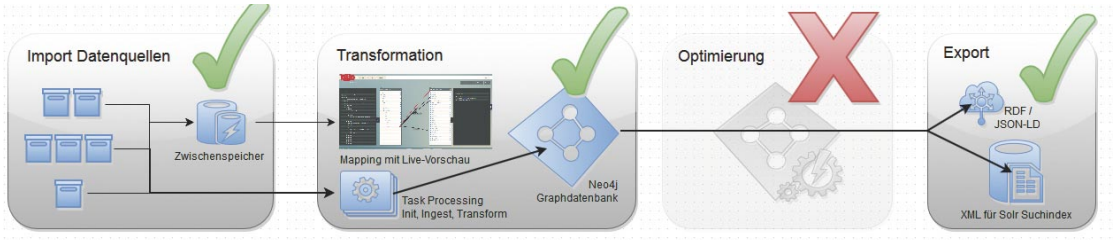
Open Source-Werkzeuge für Datenmanagement gibt es reichlich, auch speziell für bibliothekarische Daten. In Sachsen wird beispielsweise die Software VuFind mit zahlreichen Erweiterungen der UB Leipzig erfolgreich in der finc Nutzergemeinschaft eingesetzt, um die vielfältigen Daten wissenschaftlicher Hochschulbibliotheken in Sachsen zu integrieren (vgl. www.finc.info). Der technische Ansatz von d:swarm unterscheidet sich von den gängigen Lösungen in zwei Punkten:

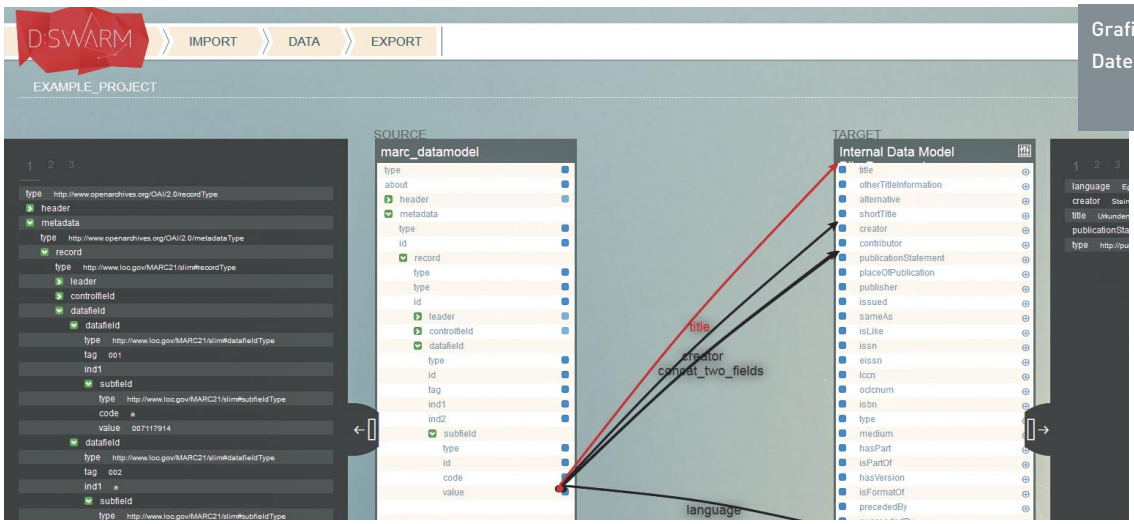
1. Die Integration von Datenquellen kann über eine grafische Oberfläche erfolgen. Beim Mapping vom Quell- auf das Zielschema können intuitiv Pfeile gezogen werden und es gibt eine Live-Vorschau auf die Ergebnisse der Transformationen. Damit kann dieser Arbeitsschritt perspektivisch aus den Händen von Informatikern in diejenigen von Bibliothekaren gegeben werden.
2. Alle Datenquellen werden inklusive ihrer Provenienz in einen zentralen Property-Graph importiert. Deduplizierung, FRBRisierung und Anreicherungslogiken können über den gesamten Datenbestand erfolgen. Damit sind intelligentere Logiken auf Basis von Datenanalysen möglich als bei der üblichen skriptbasierten Einzelbetrachtung jeder Datenquelle.

Wie aus der Grafik auf Seite 89 ersichtlich, ist bislang nur das erste Alleinstellungsmerkmal realisiert. Die intelligenten Logiken zur Optimierung der Datenqualität im Wissensgraphen sind noch nicht implementiert.

Entwicklungsstand im Mai 2015

Innerhalb der Projektlaufzeit (bis September 2014) konnte die gesamte Infrastruktur prototypisch für kleine Datensätze implementiert werden. Zentrales Element für die Anwender ist eine grafische Oberfläche zum Import und zur Modellierung der Datentransformationen (siehe Screenshot S. 89). Dabei stehen insbesondere die im Projekt Culturegraph





Grafische Oberfläche zum Datenimport

(http://culturegraph.github.io/) entwickelten Funktionalitäten zur Verfügung. Die Speicherung der Daten erfolgt in einer Graphdatenbank. Abschließend können die Daten in verschiedenen Formaten (XML, RDF, JSON-LD, ...) exportiert und beispielsweise in einen Suchindex für einen Bibliothekskatalog geladen werden.

Anlässlich der Konferenz Semantic Web in Libraries haben wir im Dezember 2014 eine Beta-Version von d:swarm veröffentlicht. Der gesamte Quellcode und die Dokumentation liegt bei Github (https://github.com/dswarm). Die grafische Oberfläche für die Modellierung der Transformationen kann auf einer Demo-Installation ausprobiert werden (http://demo.dswarm.org). Seitdem entwickeln wir das System weiter, um es zunächst an der SLUB Dresden in den Produktivbetrieb zu bringen (Version 0.9, Juni 2015) und anschließend abgerundet anderen Einrichtungen zur Verfügung stellen zu können (Version 1.0, ~2016). Detaillierte Arbeitspläne sind im Ticketsystem öffentlich einsehbar (https://jira.slub-dresden.de/browse/DD/).

Verbliebene Herausforderungen

Für die Verarbeitung von rund 1,5 Millionen Datensätzen (Deutsche Fotothek, E-book Library, E-Book-Pakete von Wiley, deGruyter usw.) verwenden wir eine einfache Task Processing Unit, die initial als erste Erweiterung für d:swarm von Hans-Georg Becker (Universitätsbibliothek der TU Dortmund) entwickelt wurde. Damit wird, wie bei anderen Systemen auch, jede Datenquelle isoliert prozessiert und die Graphdatenbank nur als Zwischenspeicher genutzt. Um der Vision eines zentralen Wissensgraphen näher zu kommen, müssen wir zunächst aufgetretene Skalierungsprobleme lösen. Derzeit erfordert die vollständige Verarbeitung der genannten Datensätze noch rund zehn Stunden Laufzeit und einen Speicherbedarf von zehn Gigabyte, was für unsere Zielstellung (~ 100 Millionen Datensätze vollständig im Graph und tägliche Updateroutinen) noch nicht hinreichend skaliert. Wir planen eine Zusammenarbeit mit dem Big Data Kompetenzzentrum Dresden/Leipzig (ScaDS), um die Infrastruktur zu optimieren.

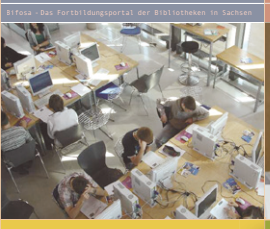
Weiterhin müssen die (vorhandenen) Konzepte für Deduplizierung, FRBRisierung und Anreicherungslogiken noch aufwändig implementiert werden, damit sich der technische Ansatz auszahlen kann. Gemeinsam mit der UB Leipzig evaluieren wir derzeit, wie sich die beiden Systeme finc und d:swarm elegant verbinden lassen, um Ressourcen zu bündeln und die jeweiligen Vorteile beider Systeme nutzen zu können. Dazu mehr im folgenden BIS-Heft 3/2015.

www.dswarm.org

http://demo.dswarm.org

Die Datenmanagement-Plattform d:swarm bildet das Herzstück der zukünftigen Datenverarbeitungsinfrastruktur der SLUB Dresden und wird konsequent weiterentwickelt. Im Sinne von Openness und der damit einhergehenden Flexibilität und Unabhängigkeit von kommerzieller Software in einem innovationsfernen Marktsegment verfolgt die SLUB Dresden weiterhin verstärkt Open-Source-Ansätze.

FELIX LOHMEIER



BIS

Das Magazin der Bibliotheken in Sachsen

Dieser Text (nicht die Bilder) steht unter der

Creative Commons Namensnennung - Weitergabe unter gleichen Bedingungen 4.0 International Lizenz