Informationspraxis Bd. 3, Nr. 2 (2017)

DOI: https://doi.org/10.11588/ip.2017.2.43235

Regina RETTER, Christian WILKE

Zitationsdaten extrahieren: halbautomatisch, offen, vernetzt. Ein Workshopbericht

Zusammenfassung

Bericht zum Linked Open Citation Database Workshop an der Universität Mannheim am 07. November 2017.

Schlüsselwörter

Linked Open Data; Zitation; Semantic Web; Workshop

Linked Open Citation Data and its Semiautomatic Extraction

Abstract

Report on the Linked Open Citation Database Workshop at the University of Mannheim, November 7th 2017..

Keywords

linked open data; citation; semantic web; workshop

Inhaltsverzeichnis

1 Einleitung

Die Erschließung von Zitationen stand im Mittelpunkt des Linked Open Citation Database Workshops am 07.11.2017 an der Universität Mannheim. Anlässlich der Halbzeit ihres DFG-Projekts Linked Open Citation Database (LOC-DB)[1] gaben die Universitätsbibliothek Mannheim, die Hochschule der Medien Stuttgart, die Deutsche Zentralbibliothek für Wirtschaftswissenschaften – Leibniz-Informationszentrum Wirtschaft Kiel und das Deutsche Forschungszentrum für Künstliche Intelligenz Kaiserslautern einen Einblick in den aktuellen Entwicklungsstand und suchten Austausch mit ähnlichen Initiativen. Im LOC-DB-Projekt werden Arbeitsabläufe und Werkzeuge entwickelt, die es Bibliotheken ermöglichen, Zitationsbeziehungen zwischen einzelnen Werken halbautomatisiert zu erfassen und nachnutzbar zu machen. Hierzu werden Literaturangaben aus gedruckten und elektronischen Medien extrahiert und die Zitationen in einer Datenbank und als Linked Open Data zur Verfügung gestellt. Damit werden Zitationen Teil des Semantic Web.

2 Gemeinsame Datenproduktion – Zitationsdaten auf dem Weg in das Semantic Web

In ihrer Einführung erläuterte Annette Klein von der Universitätsbibliothek Mannheim die Grundidee des Projekts LOC-DB und stellte das Programm des Workshops vor. Die Materialien einer Bibliothek vollständig mit allen ihren Beziehungen zueinander zu erschließen und mit der Wissensbasis des Semantic Web zu verbinden sei die Zukunftsvision der Linked-Open-Data-Technologie in Bibliotheken. Das LOC-DB-Projekt ziele darauf ab zu evaluieren, was konkret an Personal und Infrastruktur notwendig sei, um diese Vision zu realisieren. Zentral sei hierfür einerseits die Nachnutzung vorhandener Metadaten, andererseits eine kooperative Erschließung, die weitgehend durch Automatisierung unterstützt werde. Nachnutzbar sind bereits die Titeldaten der Verbundkataloge und der Zeitschriftendatenbank (ZDB) zu Monographien und Zeitschriften. Metadaten zu einzelnen Artikeln lassen sich u.a. über die DOI-Vergabe-Organisation Crossref beziehen; dank der *Initiative for Open Citations* I4OC (https://i4oc.org/) verfügen diese Datensätze in steigendem Umfang (derzeit knapp zur Hälfte) auch über offene Zitationsdaten. Eine weitere Metadatenquelle für offene verlinkte Zitationsdaten stellt das Repositorium *OpenCitations* (http://opencitations.net/) dar, das derzeit ca. 11 Millionen Zitationen vor allem aus den Lebenswissenschaften umfasst und sich langfristig zu einem Hub für offene Zitationsdaten entwickeln will.

Was jetzt noch fehlt, sind nach Klein Zitationsbeziehungen zwischen gedruckten Publikationen, eine einheitliche Strukturierung und Qualität der vorhandenen Daten und die Verknüpfung von bibliothekarischen Nachweissystemen und anderen Quellen. Wie LOC-DB mittels automatisierter Prozesse diese Lücken füllen kann, etwa durch ein Redaktionssystem mit eingebauten Schnittstellen zur Datenübernahme von externen Anbietern und durch eine Texterkennungs- und Segmentierungskomponente zur Erschließung gedruckter Ressourcen, sei ebenso Gegenstand der folgenden Vorträge wie die Erfahrungen bei Standardisierung und

1 https://locdb.bib.uni-mannheim.de/blog/de/, sämtliche Präsentationen des Workshops sind online abrufbar unter: https://locdb.bib.uni-mannheim.de/blog/de/workshop/

Verbesserung der Datenqualität aus anderen Projekten. Kleins Einführung endete mit der Vision einer kooperativen Datenproduktion durch ein Netz verteilter LOC-DB-Instanzen und der Frage nach den Akteuren, die hierfür zuständig sind.

3 Wieviel Mehraufwand entsteht für Bibliotheken? - Eine Ressourcenabschätzung

Philipp Zumstein und Laura Erhard, ebenfalls UB Mannheim, schilderten in ihrem Beitrag zu LOC-DB-Daten und -Workflows das Projekt aus bibliothekarischer Perspektive. Die Rolle der UB bestehe neben der allgemeinen Projektplanung vor allem in der Entwicklung von Workflows für die Digitalisierung der Literaturverzeichnisse und die Bearbeitung der Metadaten im Redaktionssystem, aber auch in der Erstellung einer Kosten-Nutzen-Analyse für das Gesamtprojekt.

Zunächst habe die Bibliothek als Datengrundlage einen Ausschnitt des Fachbestands Soziologie ausgewählt. In einer ersten Phase seien sämtliche Literaturverzeichnisse der Neuerwerbungen des Jahres 2011 – Monographien und Zeitschriften ebenso wie Sammelwerke – erfasst worden, in der zweiten Phase seien die Workflows im laufenden Betrieb an den Neuerwerbungen aus 2017 getestet worden. Erhard erläuterte exemplarisch den Workflow für Sammelwerke im Detail. Bei einem Werk, das bei Bestellung für das LOC-DB-Projekt vorgemerkt wurde, wird nach Eingang und Katalogisierung zunächst das Literaturverzeichnis gescannt und anschließend das Buch für die Nutzer eingestellt. Der mit Pica-Produktionsnummer (PPN) versehene Scan wird ins Redaktionssystem hochgeladen, wo die zugehörigen Metadaten für das Gesamtwerk aus dem Katalog des Südwestdeutschen Bibliotheksverbunds (SWB) und für die einzelnen Kapitel falls vorhanden aus Crossref eingespielt und zusammengeführt werden. Dann verknüpft ein Mitarbeiter die aus den Scans extrahierten Zitationen, wobei er auf automatisierte Vorschläge aus Fremddaten des SWB und von Crossref zurückgreifen kann. Sofern keine Fremddaten verfügbar sind, wird manuell nachgearbeitet.

Zumstein präsentierte die Ergebnisse einer ersten Ressourcenabschätzung für den Jahrgang 2011: weniger als 135,5 Hiwi-Stunden wurden schätzungsweise für das Scannen von 13.550 Seiten Literaturverzeichnissen aufgewandt, 26,6 Stunden betrug der Mehraufwand im Geschäftsgang (3 Minuten pro Buch). Der Aufwand für die Arbeit im Redaktionssystem ist derzeit noch nicht bezifferbar. Es ist gelungen, einen semi-automatischen Workflow für alle Mediendaten zu schaffen, der die Nachnutzung vorhandener Daten ermöglicht und Verknüpfungen durch Identifikatoren gewährleistet. Als Desiderate verbleiben weiterhin die Verknüpfung mit Normdatensätzen und die Nutzung weiterer Datenquellen.

4 Die Datenbank und ihre Benutzeroberfläche – zur Funktionsweise von LOC-DB

Die Architektur von Backend und Frontend der Datenbank waren Thema eines weiteren Vortrags. Anne Lauscher von der Universität Mannheim, die betreut von Kai Eckert (Hochschule der Medien) das Backend entwickelt hat, erläuterte die LOC-DB Pipeline als Abfolge von Auffinden der Zitationen im gescannten Bild, Texterkennung, Zitationsextraktion, Informationsgewinnung aus internen Datenbanken und externen Ressourcen, Verlinkung der

Ressourcen mit menschlicher Prüfung und schließlich der dezentralen Speicherung nach einem Datenmodell von OpenCitations. Das Backend der Datenbank bestehe aus den Komponenten Swagger/Express, Node.js, Elastic Search und MongoDB. Wenn im Backend der Scan einer gedruckten Monographie mit der PPN des Verbunds eingeht, werden zunächst die Metadaten aus dem Verbund abgefragt und gesichert, dann die OCR-Verarbeitung angestoßen und Referenzen extrahiert und gespeichert. Anschließend werden externe und interne Vorschläge für die Referenzen generiert, und letztlich die Referenzen verlinkt und die Verlinkung persistiert. Bei elektronischen Ressourcen wird die DOI zur Identifikation verwendet. Die Titeldaten sowie gegebenenfalls schon vorhandene Zitationsdaten werden aus Crossref übernommen. Die Funktionsweise des Frontends, des sogenannten Redaktionssystems, wurde von Lukas Galke von der ZBW Kiel präsentiert. Die drei Kernfunktionalitäten des Systems sind die Eingabe der Scans, die Auswahl von Ressourcen und Referenzen aus dem Literaturverzeichnis und schließlich die Auswahl des korrekten Ziels der Zitation aus wenigen Vorschlägen. Es muss ein Datenmanagement-System mit zahlreichen Schnittstellen zur Verfügung stellen, das von Mitarbeitern mit getrennten Zuständigkeiten bedient werden kann und die Verknüpfung von zwei bibliografischen Ressourcen ermöglicht.

Wie das Redaktionssystem diesen Anforderungen gerecht wird, veranschaulichte Galke durch Snapshots aus dem Redaktionssystem. Während die funktionalen Anforderungen im Grundsatz bereits abgebildet sind, müssen Effizienz, Effektivität und Nutzerfreundlichkeit noch evaluiert werden. In Praxissessions hatten die Teilnehmer des Workshops später die Gelegenheit, die Workflows für elektronische Medien und Printwerke selbst zu testen.

5 The magic – visuelle Zitationserkennung mit Hilfe künstlicher neuronaler Netze

Im Zentrum des Vortrags von Sheraz Ahmed vom Deutschen Forschungszentrum für Künstliche Intelligenz Kaiserlautern (DFKI) stand das, was für den Verlauf des Workshops unter dem geflügelten Wort „the magic“ firmierte: die OCR-Komponente von LOC-DB, mittels derer die Zitationen extrahiert werden. Ahmed erläuterte zunächst den Ablauf der Zitationsextration für Scans aus gedruckten Büchern, digitalen PDF-Dateien und strukturierten XML-Formaten mit Hilfe des Programms ParsCit (https://github.com/knmnyn/ParsCit). Bei gescannten Buchseiten ist das Verfahren besonders komplex. Zunächst werden die Bilder binarisiert, also in Schwarz-Weiß-Werte transformiert, dann als ein- oder zweispaltige Dokumente klassifiziert und schließlich mit OCR in Text umgewandelt, bevor die Segmentierung einzelner Zeichenfolgen als Zitationen mit ParsCit erfolgt.

Als Alternative zu diesem textbasierten Vorgehen mit ParsCit präsentierte Ahmed schließlich eine eigene Entwicklung, die Zitationen in Scans und PDFs noch präziser voneinander abgrenzen kann: das Programm DeepBibX (vgl. BHARDWAJ et al 2017), das auf künstlichen neuronalen Netzen basiert und die Methode des Deep Learning anwendet. Statt zuerst den Text auszuwerten, imitiert DeepBibX menschliches Sehen und trennt die Zitationen voneinander, indem es Muster in den Bildern erkennt. So erfolgt im ersten Schritt die Segmentierung der einzelnen Zitationen durch ein visuelles Verfahren, erst anschließend wird eine Texterkennungskomponente eingesetzt. DeepBibX ist nicht nur sprachunabhängig anwendbar,

sondern auch deutlich treffsicherer als ParsCit. In einem Test an 286 Bilddateien mit insgesamt 5090 Zitationen hat DeepBibX 84,9% der Zitationen korrekt extrahiert, ParsCit hingegen nur 71,7%. In einem Expertengespräch am Nachmittag stand Ahmed für Rückfragen zur Verfügung und zeigte Beispiele für Probleme bei der Segmentierung.

6 EXCITE – eine neue Lösung zur Texterkennung von Literaturverzeichnissen?

Behnam Ghavimi (GESIS – Leibniz-Institut für Sozialwissenschaften) berichtete von dem DFG-Projekt *EXCITE*, das gemeinsam von der GESIS und dem *Institute for Web Science and Technologies* (WeST) der Universität Koblenz-Landau durchgeführt wird. Darin wird versucht, dem Mangel an Zitationsdaten Abhilfe zu schaffen, der vor allem in den deutschen Sozialwissenschaften anders als in weiten Teilen der Naturwissenschaften herrsche. Die Projektgruppe entwickelt daher seit September 2016 eine Software, die solche Daten aus PDF-Dokumenten deutschsprachiger Fachpublikationen auslesen soll. Sie verfolgt dabei einen eigenen Ansatz: Bestehende Softwarelösungen – wie ParsCit, Cermine (https://github.com/CeON/CERMINE) oder GROBID (https://github.com/kermitt2/grobid) – versuchen in einem ersten Schritt, das Literaturverzeichnis zu identifizieren, um dann in einem zweiten Schritt das gegebenenfalls fehlerhaft selektierte Verzeichnis in einzelne Einträge (*reference strings*) zu zerlegen. Nach dem EXCITE-Ansatz werden die bestehenden Programme modifiziert und trainiert, um zunächst linguistische und typografische Informationen des vorliegenden Textes auszulesen, zum Beispiel Großschreibung am Zeilenbeginn, Einrückungen und Zeilenlängen. Auf dieser Basis entscheidet dann das neu entwickelte Programm *RefExt* (https://github.com/exciteproject/refext), ob eine Zeile zu einer Literaturangabe (*reference string*) gehört oder nicht.

7 Automatisierte Erschließung von Zeitschriftenaufsätzen – der FID Theologie

Timotheus Chang Whae Kim, Fachreferent für Theologie und Koreanistik an der Universitätsbibliothek Tübingen, stellte Verfahren zur halbautomatischen Katalogisierung unselbständiger Werke innerhalb des Fachinformationsdienstes (FID) Theologie vor. Die Hälfte jener 1349 Zeitschriften, deren Inhaltsverzeichnisse bibliografische Daten zu Aufsätzen liefern sollen, liegt nur gedruckt vor. Derzeit werde diese enorme Aufgabe zum Teil automatisch bewältigt, und zwar mithilfe des kommerziellen Softwarepakets C-3 der Firma ImageWare Components GmbH (vgl. FASSNACHT 2017). Sie leiste nicht nur eine OCR-Erkennung der gescannten Zeitschriften-Inhaltsverzeichnisse, sondern übertrage diese auch mithilfe angepasster Templates einzelner Zeitschriften in strukturierte Daten (Autorenname, Titel, Seitenzahl). Diese Aufsatzdaten könnten manuell ergänzt (Erscheinungsjahr, Heftnummer u.ä.), im XML-Format gespeichert und schließlich direkt in das im Südwestdeutschen Bibliotheksverbund (SWB) genutzte Pica-Format exportiert werden.

Im Fall von Online-Zeitschriften komme, so Kim, eine Erweiterung des Literaturverwaltungsprogramms Zotero (https://www.zotero.org/) zum Einsatz. Zotero dient dabei nicht selbst als Katalogisierungsclient, sondern als Instrument, das bibliografische Metadaten zum Beispiel aus Verlagswebseiten oder Bibliothekskatalogen sammelt (Zotero

Picker). Um diese in das Pica3-Format der SWB-Datenbank zu exportieren und mit den verbundinternen Normdaten anzureichern, wurde in Zusammenarbeit mit der UB Mannheim die Zotero-Programmerweiterung zotkat (https://github.com/UB-Mannheim/zotkat) entwickelt, die als Vorstufe zum SWB-Katalogisierungsclient WinIBW geschaltet wird (vgl. KIM 2016).

8 Zukunft der Zitationsdaten – eine Einschätzung der ZBW

Tamara Pianos, Leiterin der Abteilung Informationsvermittlung des ZBW – Leibniz-Informationszentrum Wirtschaft, sieht große Chancen, aber auch einige Hindernisse für den künftigen Einsatz von Zitationsdaten im ZBW-Fachportal *EconBiz* (https://www.econbiz.de/). Einerseits können Angaben zu den aktiven und passiven Zitationen eines Werks (Werk zitiert A, B, C und wurde zitiert von D, E, F) im Katalog angezeigt werden und so die Recherche für Nutzer effizienter machen. Solche bibliometrischen Informationen können ferner für die Visualisierung und die Analyse von Zitationsnetzwerken und nicht zuletzt zur Anreicherung von wissenschaftlichen Autorenprofilen genutzt werden, wie man sie auf ResearchGate (https://www.researchgate.net/) oder Google Scholar (https://scholar.google.de/) finden kann. Andererseits gab Pianos auch mit Blick auf diese Portale einiges zu bedenken. Zum einen sind die Quellen zur Erstellung von Zitationsdaten keineswegs vollständig; entweder weil sie nicht erfasst wurden oder weil sie (noch) nicht frei zugänglich gemacht werden dürfen. Zum anderen entstehen derzeit auch auf renommierten Zitationsdatenbanken wie dem Social Science Citation Index viele Fehler bei der Disambiguierung vorhandener Daten, also etwa der eindeutigen Zuordnung von Autorennamen zu einzelnen Personen, und damit auch bei der Zählung von Zitationen (vgl. TÜÜR-FRÖHLICH 2016). Um das große Potential zur Nutzung von Zitationsdaten besser auszuschöpfen als bisher, komme alles auf die richtige Verknüpfung von Linked Data mit Personennormdaten an. Diese könne jedoch nicht ohne den Einbezug menschlicher Intelligenz bewerkstelligt werden: sei es durch die Forschenden selbst, die ihre Autoren- und Zitationsprofile mitpflegen sollten, sei es durch die seit Langem bewährte Metadatenkompetenz der Bibliothekare.

9 OpenCitations – eine Konkurrenz zum Web Of Science?

Per Web-Konferenz wurde schließlich David Shotton (Universität Oxford) mit einem Vortrag zu *OpenCitations* zugeschaltet. Shotton ist zusammen mit Silvio Peroni (Universität Bologna) Leiter *von OpenCitations*, einer Infrastrukturorganisation, die schon 2011 damit begann, ein frei zugängliches Repositorium für Zitationsdaten aufzubauen, das sogenannte *OpenCitations Corpus* (OCC). Derzeit liegen zwar weit weniger Zitationen vor als die 1,25 Milliarden des kommerziellen Marktführers Web of Science, nämlich nur 11 Millionen Zitationen aus ca. 270.000 Open-Access-Artikeln von PubMed Central, einer Open-Access-Publikations-Datenbank der Medizin und Biologie. *OpenCitations* hat aber erst einen Bruchteil (ca. 16 %) der dort frei verfügbaren Artikel ‚abgegrast' und ist damit schon heute die größte frei zugängliche Zitationsdatenbank der Welt.

Das OCC lasse sich laut Shotton mit zusätzlicher Hardware, also zusätzlichem Geld, leicht um ein Vielfaches vergrößern. Das zentrale Hindernis sei jedoch, dass ein Großteil der

Zitationsdaten aus Zeitschriftenaufsätzen gegenwärtig weder frei zugänglich sei noch frei nachgenutzt werden dürfe. Künftig sollten Zitationsdaten in einen Teil der *commons*, also der frei zugänglichen Werke, verwandelt werden. In dieser Absicht war *OpenCitations* (neben *Wikimedia Foundation*, PLOS, *eLife* und *DataCite*) auch Mitbegründer der im April 2017 gestarteten *Initiative for Open Citations* (I4OC), einer Vereinigung für die freie Zugänglichkeit und innovative Aufbereitung von Zitationsdaten. Ihr sind mittlerweile auch namhafte Wissenschaftsverlage wie *Wiley*, *Springer Nature*, *De Gruyter* und *Cambridge University Press* beigetreten, so dass der Anteil der frei zugänglichen Zitationsdaten aller Publikationen, die diese Verlage an CrossRef übermitteln, sprunghaft von 1% auf über 45 % (Stand Juni 2017) angestiegen ist.

10 Abschließende Podiumsdiskussion

Der Workshop schloss mit einer Podiumsdiskussion, die von Annette Klein moderiert wurde und an der Kai Eckert, Tamara Pianos, Timotheus Kim sowie Jakob Voß (Verbundzentrale des GBV, VZG) teilnahmen. Insgesamt herrschte Konsens darüber, dass Zitationsdaten ein neues Aufgabenfeld für Bibliotheken darstellen. Auch in der Frage, wer für die Speicherung und die Verteilung dieser Daten zuständig sei, bestand Einigkeit: Die Bibliotheksverbünde sollten hier Verantwortung übernehmen. Das Desiderat sei jetzt, dass die Verbünde eine Möglichkeit schaffen, Zitationen als einzelne Daten im jeweiligen Datenformat der Verbünde (MARC21, Pica) abzuspeichern. Schließlich wurde die zentrale Bedeutung von offenen Zitationsdaten angesichts der Fehleranfälligkeit und der Intransparenz von kommerziellen Zitationsdatenbanken hervorgehoben. Nur offene Zitationsdaten könnten Forschungsrichtungen wie die Bibliometrie auf eine solide Grundlage stellen und darüber hinaus die wissenschaftliche Recherche verbessern.

Der Workshop bot den Teilnehmern Gelegenheit, einen Blick auf eine bibliothekarische Zukunftstechnologie zu werfen, die sich noch mitten im Entstehungsprozess befindet. An der offenen Gesprächsatmosphäre merkte man deutlich, dass es den Organisatoren nicht nur um die Vernetzung von Daten ging, sondern auch darum, den Austausch mit Interessierten zu suchen, die Kooperation unter den Experten zu vertiefen und zukünftige Anwendern frühzeitig in die Entwicklung miteinzubeziehen.

Quellen

Bhardwaj, Akansha, u.a. 2017. DeepBIBX: Deep Learning for Image Based Bibliographic Data Extraction, in Liu, Derong, u.a. (Hg.): Neural Information Processing. Cham: Springer International Publishing. (10635). DOI: 10.1007/978-3-319-70096-0_30.

Faßnacht, Martin & Gebhard, Winfried 2016. Index Theologicus – neue Produktionsverfahren bei der Bibliographieerstellung. *B.I.T. Online* 19(6), 511–514. URL: http://www.b-i-t-online.de/heft/2016-06-nachrichtenbeitrag-fassnacht.pdf.

Kim, Timotheus C.-w. & Zumstein, Philipp 2016. Semiautomatische Katalogisierung und Normdatenverknüpfung mit Zotero im Index Theologicus. Libreas.Library Ideas (29). URL: http://libreas.eu/ausgabe29/05kim/.

Tüür-Fröhlich, Terje 2016. The non-trivial effects of trivial errors in scientific communication and evaluation. Dissertation. Als Manuskript gedruckt. (Schriften zur Informationswissenschaft, Bd. 69).

AutorInnen

Regina RETTER
Universitätsbibliothek Mannheim
Schloss Schneckenhof West
68131 Mannheim

https://www.bib.uni-mannheim.de/

regina.retter@bib.uni-mannheim.de

Christian WILKE
Universitätsbibliothek Mannheim
Schloss Schneckenhof West
68131 Mannheim

https://www.bib.uni-mannheim.de/

christian.wilke@bib.uni-mannheim.de