Themenkreis 5: Fokus Lehre & Forschung

Von der Schneeflocke zur Lawine: Möglichkeiten der Nutzung freier Zitationsdaten in Bibliotheken

Annette Klein, Universitätsbibliothek Mannheim

Zusammenfassung:
Zitationen spielen eine wichtige Rolle im wissenschaftlichen Diskurs, in der Recherchepraxis sowie im Bereich der Bibliometrie. In jüngster Zeit gibt es zunehmend Initiativen, die Zitationen als Open Data zur freien Nachnutzung verfügbar machen. Der Beitrag beschreibt den Stand der Entwicklung dieser Initiativen und zeigt, dass in nächster Zeit eine kritische Masse von Daten entstehen könnte, aus denen sich gerade für Bibliotheken neue Perspektiven ergeben. Als konkrete Möglichkeit zur Partizipation für Bibliotheken wird das DFG-Projekt Linked Open Citation Database (LOC-DB) vorgestellt.

Summary:
Citations play an important role in scientific discourse, in the practice of information retrieval, and in bibliometrics. Recently, there have been a growing number of initiatives which make citations freely available as open data. The article describes the current status of these initiatives and shows that a critical mass of data could be made available in the near future. New opportunities could arise from that, especially for libraries. The DFG funded project Linked Open Citation Database (LOC-DB) is presented as a practical way for libraries to participate.

Zitierfähiger Link (DOI): https://doi.org/10.5282/o-bib/2017H4S127-136
Autorenidentifikation: Klein, Annette: GND 128819146
ORCID: http://orcid.org/0000-0001-8825-6446
Schlagwörter: Zitat, Open Data, Bibliothek

1. Einleitung

Zitationen bilden ein wesentliches Element des wissenschaftlichen Diskurses. Die Auflistung zitierter Literatur in einem Literaturverzeichnis ist eine Anforderung der „guten wissenschaftlichen Praxis", weil sie nachvollziehbar macht, welche fremden Inhalte von der Autorin bzw. dem Autor des zitierenden Werkes rezipiert worden sind. Auf diese Weise werden inhaltliche Beziehungen zwischen wissenschaftlichen Publikationen transparent gemacht. „Citations are the links that knit together our scientific and cultural knowledge", stellt die Initiative for Open Citations (I4OC) zusammenfassend fest.[1]

Zitationen sind aber auch wichtig für Bibliotheken. In vielen Einführungen in die Literaturrecherche empfehlen Bibliothekarinnen und Bibliothekare Studienanfängerinnen und Studienanfängern das „Schneeballsystem", das darauf beruht, dass ausgehend von nur einer passenden Quelle zu einem

1 „About," I4OC, zuletzt geprüft am 28.11.2017, https://i40c.org/#about.

bestimmten Thema durch die Auswertung der zitierten Literatur rasch eine große Anzahl weiterer relevanter Quellen gefunden werden kann. Jeder, der dieses Verfahren einmal selbst mit gedruckter Literatur praktiziert hat, weiß allerdings auch, wie mühsam es in der Praxis sein kann. Es ist daher naheliegend, die Funktionalität des „Verfolgens" von Zitationsbeziehungen direkt in Recherchesysteme wie z.B. die Online-Kataloge wissenschaftlicher Bibliotheken einzubauen. Sind Zitationen umfassend in ein Recherchesystem eingebunden, kann im Gegensatz zum analogen Verfahren nicht nur die zitierte (und damit der Publikation zeitlich vorangehende) Literatur verknüpft werden, sondern auch die späteren Publikationen, die die Quelle ihrerseits zitieren. Solche anklickbaren Zitationsnetzwerke bieten den Nutzerinnen und Nutzern bei der Recherche einen echten Mehrwert.

Darüber hinaus sind Zitationsdaten grundlegend für die meisten quantitativen Publikationsanalysen im Bereich der Bibliometrie. Seit einiger Zeit beschäftigen sich zunehmend auch Bibliotheken mit solchen bibliometrischen Analysen und bieten Services zur Forschungsevaluierung an.[2] In der Regel werden als Datengrundlage hochpreisige kommerzielle Zitationsdatenbanken wie Scopus und Web of Science genutzt. Dabei wird immer wieder kritisiert, dass diese Analysen aufgrund der mangelnden Vollständigkeit und Qualität der Daten nur eingeschränkt zuverlässig sind,[3] ihnen gleichwohl jedoch in einigen Disziplinen eine existentielle Bedeutung für die Karriere von Nachwuchswissenschaftlerinnen und Nachwuchswissenschaftlern zukommt.[4] Um die bekannten Mängel der kommerziellen Zitationsdatenbanken auszugleichen, werden an vielen Einrichtungen zusätzlich zu den Lizenzkosten noch einmal erhebliche Ressourcen in die Aufbereitung und Verbesserung der Daten investiert.[5]

Bei all dem ist nicht zu vergessen, dass es sich bei Zitationen letztlich nicht um Forschungsinhalte, sondern ‚nur' um Metadaten bzw. Beziehungen zwischen Metadaten handelt. Dass diese kostenfrei und ohne Hindernisse zugänglich sein sollten, darf eigentlich als breiter Konsens zwischen vielen Akteuren des Publikationsmarktes gelten. Im Folgenden soll dargestellt werden, wie sich ausgehend von einigen Initiativen in jüngster Zeit geradezu ein Trend zur massenhaften Produktion und Freigabe von Zitationsdaten entwickelt hat, und welche Perspektiven sich hieraus für Bibliotheken ergeben könnten.

2 So z.B. die TU München (vgl. „Bibliometrie," TUM, zuletzt geprüft am 28.11.2017, https://www.ub.tum.de/bibliometrie) und die UB Wien (vgl. „Bibliometrie an der Universität Wien," Universität Wien, zuletzt geprüft am 30.08.2017, https://bibliothek.univie.ac.at/bibliometrie/).

3 Vgl. Benjamin Walker et al., „Inter-rater Reliability of H-Index Scores Calculated by Web of Science and Scopus for Clinical Epidemiology Scientists," *Health Information and Libraries Journal* 33, Nr. 2 (2016): 140–149, http://doi.org/10.1111/hir.12140.

4 Vgl. Maximilian Fochler, Ulrike Felt und Ruth Müller, „Unsustainable Growth, Hyper-Competition, and Worth in Life Science Research: Narrowing Evaluative Repertoires in Doctoral and Postdoctoral Scientists' Work and Lives," *Minerva* 54, Nr. 2 (2016): 175–200, http://doi.org/10.1007/s11024-016-9292-y.

5 Die Universität Wien, die in diesem Bereich einen sehr professionellen Service liefert, hat beispielsweise eine „Abteilung Bibliometrie und Publikationsstrategien" mit sechs Mitarbeiter/inne/n.

2. Die Schneeflocke: CiteSeer/CiteSeerX

Bereits seit 1998 existiert für wissenschaftliche Artikel in der Informatik der Dienst CiteSeer, der aktuell unter dem Namen CiteSeerX[6] vom College of Information Sciences and Technology der Pennsylvania State University betrieben wird. Die Datenbank enthält über 6 Millionen Dokumente und 120 Millionen Zitationen, die unter der Lizenz CC-BY-NC-SA angeboten werden.

Die Daten werden mit Hilfe eines Webcrawlers automatisch aus online verfügbaren pdf-Dateien extrahiert. Die Software ist als Open Source verfügbar[7] und wird u.a. von RePEc (Research Papers in Economics) nachgenutzt. Durch die frühe Entwicklung eines massentauglichen, automatisierten Verfahrens zur Erschließung von Publikationen und Zitationen hat CiteSeer einen wichtigen Anstoß für die Erfassung freier Zitationsdaten gegeben – es war sozusagen die erste Schneeflocke, die im Verbund mit vielen anderen das Potential hat, eine Lawine auszulösen.

Das vollautomatische Verfahren, das auf der Grundlage unstrukturierter Daten – dem Text aus den pdf-Dateien unterschiedlichster Provenienz – arbeitet, hat allerdings seine Grenzen. Schon die grundlegenden Metadaten (Autor und Titel) der ausgewerteten Artikel werden nicht immer erkannt (vgl. Abb. 1). Komplexe Beziehungen zwischen verschiedenen hierarchischen Ebenen einer Publikation und den damit verknüpften Zitationen sind mit der eingesetzten Technologie wohl grundsätzlich kaum zu erschließen. Abbildung 2 illustriert einen solchen Fall: Die automatisch extrahierten Daten suggerieren, die Autorin Wendy Hall habe einen Aufsatz mit dem Titel „Linked Open Data“ veröffentlicht, und dieser Artikel enthalte keine Zitationen. Tatsächlich ist „Linked Open Data“ aber der Titel eines Sonderhefts der Zeitschrift *Ercim News*, in dem die Autorin einen einseitigen Artikel mit dem Titel „Linked Data – the Quiet Revolution“[8] verfasst hat, der immerhin eine Literaturliste mit sechs Einträgen enthält. Weder diese Zitationen noch der eigentliche Aufsatz sind in CiteSeerX zu finden. Alle genannten Angaben waren auf dem Titelblatt des Sonderhefts aufgeführt – sie sind nur falsch zusammengesetzt bzw. nicht mit den erschließungsrelevanten Informationen im Inneren des Zeitschriftenhefts in Beziehung gesetzt worden.

Probleme dieser Art können nur mit einem neuen methodischen Ansatz gelöst werden. So kann man beispielsweise auf bereits vorhandenen strukturierten Metadaten zur Beschreibung der zitierenden Werke aufsetzen und damit zumindest Fehler bei deren Identifikation ausschließen. Dies ist möglich, wenn man die Erschließung auf bestimmte Datenquellen beschränkt, in denen elektronische Volltexte einschließlich beschreibender Metadaten vorhanden und mit Lizenzen versehen sind, die eine automatisierte Weiterverarbeitung erlauben.

6 *CiteSeerX*, http://citeseerx.ist.psu.edu/index.
7 *CiteSeerX*, https://github.com/SeerLabs.
8 Wendy Hall, „Linked Data: The Quiet Revolution,“ *ERCIM News* 96 (2014): 4, zuletzt geprüft am 17.07.2017, https://ercim-news.ercim.eu/en96/keynote.

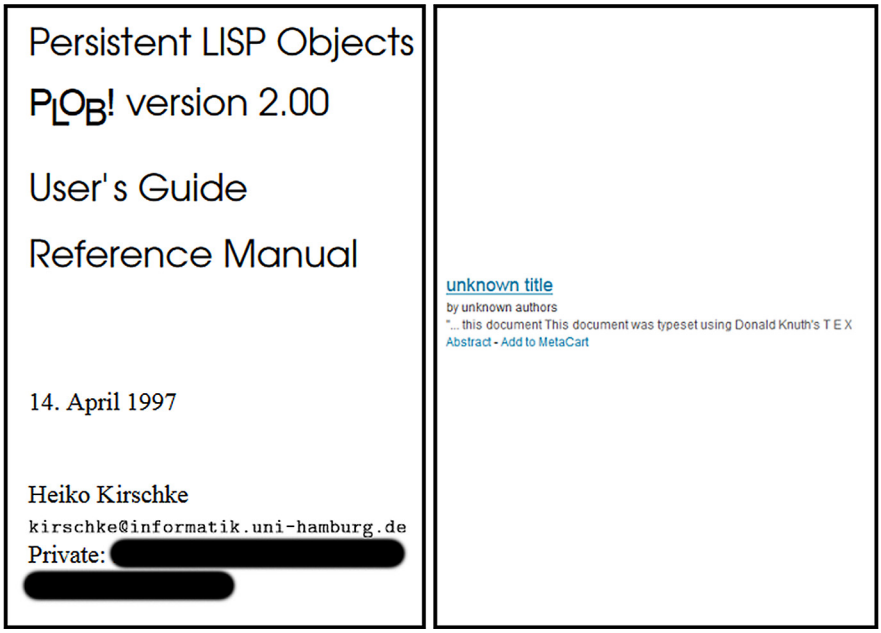

Abb. 1: Links: Titelseite eines ausgewerteten Papers; rechts: Metadaten dieses Papers in CiteSeerX

Abb. 2: Vermischung unterschiedlicher hierarchischer Ebenen in erkannten Metadaten von CiteSeerX

3. Der Schneeball: OpenCitations

Dieser Ansatz ist von OpenCitations[9] gewählt worden. Seit 2010 treibt der Biologe David Shotton (Oxford) die Erstellung eines Repositoriums offener Zitationsdaten mit Fokus auf den Lebenswissenschaften voran; seit Oktober 2015 zusammen mit Silvio Peroni (Bologna). Betreiber ist die britische gemeinnützige Gesellschaft Infrastructure Services for Open Access[10], unter deren Dach auch das Directory of Open Access Journals (DOAJ) angesiedelt ist.

9 *OpenCitations*, http://opencitations.net/.
10 *Infrastructure Services for Open Access*, https://is4oa.org/.

OpenCitations erschließt aktuell die Open-Access-Artikel in der Datenbank PubMed Central mit vollautomatischen Verfahren. Die Ergebnisse werden als Linked Open Data in einem Triple Store mit SPARQL Endpoint unter der Lizenz CC0 (public domain) angeboten. Derzeit handelt es sich um ca. 8,4 Mio. Zitationen aus 198.000 Artikeln (Stand: 18.07.2017). Für die Zukunft ist eine deutliche Ausweitung von Inhalt und Funktionalität angestrebt: "our ultimate objective [...] is to provide an open alternative to Web of Knowledge and Scopus, covering all the disciplines."[11] Zu diesem Zweck sollen neue Datenquellen wie ArXiv und Crossref in die Erschließung einbezogen und die Verarbeitungsgeschwindigkeit deutlich erhöht werden. Zudem sind neue Tools zur Visualisierung, für die Suche und das Browsing geplant.

Schon jetzt lässt sich sagen, dass mit OpenCitations eine neue Qualität bei der Erschließung von freien Zitationsdaten erreicht worden ist: Durch die Nutzung strukturierter Daten aus definierten Datenquellen ist die Datenqualität besser als beispielsweise bei CiteSeerX. Durch den Einsatz von Linked-Data-Technologie und CC0-Lizenz sind außerdem die Voraussetzungen für die Nachnutzung und weitere Verlinkung und Vernetzung der Daten sehr gut. Zur Abdeckung weiterer Fachgebiete und zur Entwicklung neuer Funktionalitäten sind Kooperationen mit anderen Akteuren naheliegend. Die Schneeflocken der früheren Initiativen verdichten sich sozusagen mit dem neuen technologischen und methodischen Ansatz von OpenCitations zu einem soliden Schneeball, mit dem ein gezielter Wurf auf das Ziel einer umfassenden freien Zitationsdatenbank nicht mehr unmöglich erscheint.

Tatsächlich sind im vergangenen Jahr mehrere Projekte gestartet, die sich unter verschiedenen Fragestellungen ebenfalls mit Zitationen befassen. Mit dem Projekt Linked Open Citation Database (LOC-DB) werden wir uns später noch ausführlich befassen, ergänzend dazu seien hier noch das DFG-Projekt EXCITE[12] und die Initiative WikiCite[13] der Wikimedia Foundation genannt.

4. Der Beginn einer Lawine? Die „Initiative for Open Citations"

Eine ganz neue Dynamik entwickelte sich im Frühjahr 2017 durch die Initiative for Open Citations (I4OC).[14] Verschiedene Akteure (darunter die Wikimedia Foundation, PLOS, DataCite und OpenCitations) haben sich zusammengeschlossen, um wissenschaftliche Verlage dazu zu bringen, die Zitationen in den von ihnen veröffentlichten Publikationen über die Plattform Crossref[15] frei verfügbar zu machen. Mitte Juli 2017 beteiligen sich bereits 46 Verlage, darunter Cambridge University Press, De Gruyter,

11 David Shotton, private E-Mail an die Verfasserin vom 15.05.2017.

12 Ziel des Projekts EXCITE ist die Entwicklung verbesserter Softwarekomponenten zur automatischen Extraktion von Zitationen aus Texten in existierenden Fachdatenbanken, vgl. „DFG-Project: EXCITE - Extraction of Citations from PDF Documents," Universität Koblenz-Landau, Institute for Web Science and Technologies, zuletzt geprüft am 28.11.2017, http://west.uni-koblenz.de/en/research/excite.

13 Vgl. *WikiCite*, https://meta.wikimedia.org/wiki/WikiCite. Die Initiative WikiCite befasst sich zunächst grundsätzlich mit bibliographischen Metadaten in den verschiedenen Projekten der Wikimedia-Foundation. Zitationen sind hier jedoch von besonderer Bedeutung, da sie ein Indiz für die Zuverlässigkeit von Inhalten liefern können (vgl. Dario Taraborelli, „WikiCite: The Journey and the Road Ahead," zuletzt geändert am 23.05.2017, http://doi.org/10.6084/m9.figshare.5032235.v1).

14 *I4OC*, https://i4oc.org/.

15 Crossref (https://www.crossref.org/) wird von einer gemeinnützigen Organisation betrieben und vergibt DOIs für elektronische Publikationen. Die Crossref-Plattform und -Schnittstellen dienen dazu, die DOIs mit den zugehörigen Metadaten recherchierbar und nutzbar zu machen, z.B. für Link Resolver.

Sage, Springer, Taylor & Francis und Wiley. Damit sind bereits 45% der ca. 35 Mio. Publikationen in Crossref inklusive Referenzen offen und über die Crossref REST-API abrufbar. Über OpenCitations sollen diese Daten in Zukunft auch im RDF-Format als Linked Open Data verfügbar werden.

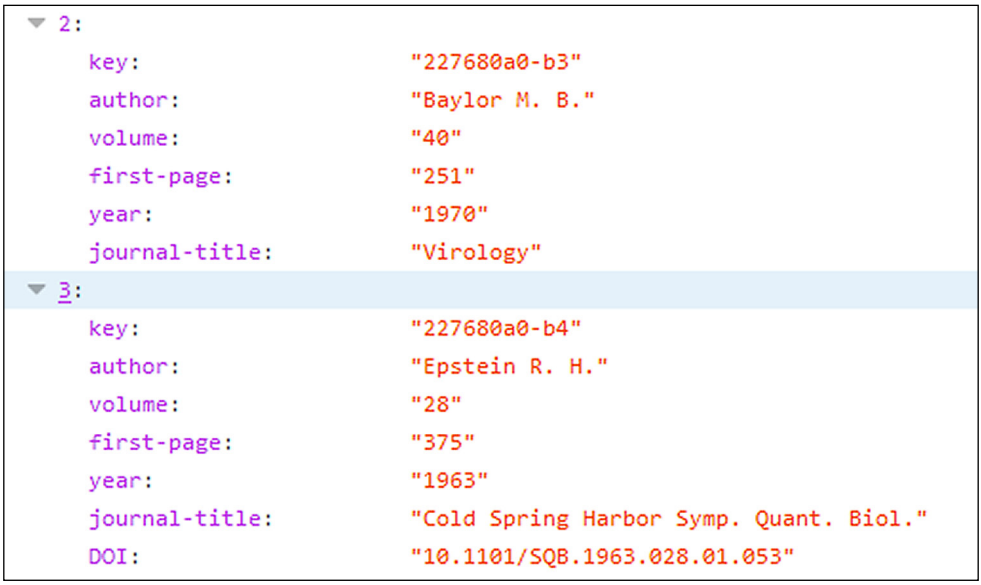

Abb. 3: Zitationsdaten bei Crossref[16]

Ist damit die Lawine bereits ausgelöst, und die Tage der kommerziellen Zitationsdatenbanken sind gezählt? Die schiere Masse der offengelegten Daten eröffnet zweifellos ganz neue Möglichkeiten. Dennoch ist das Ziel noch nicht erreicht: Zum einen ist Crossref auf elektronische Publikationen fokussiert, da die primäre Funktion ja in der Vergabe von DOIs besteht. Zum anderen ist die Datenqualität der Zitationsdaten sehr unterschiedlich. Im Beispiel in Abbildung 3 fehlt bei beiden aufgeführten Referenzen der Titel des Aufsatzes. Nur im zweiten Fall ist eine DOI verknüpft, so dass die Publikation eindeutig identifiziert, mit zusätzlichen Metadaten angereichert und verlinkt werden kann. Die DOI kann entweder bereits in den vom Verlag gelieferten Metadaten enthalten gewesen sein, oder sie wurde nachträglich durch einen von Crossref angebotenen Verlinkungsdienst ergänzt.[17] Dies funktioniert jedoch nur, wenn die Qualität der Ausgangsdaten ausreichend und die Zielpublikation tatsächlich bei Crossref registriert ist – was beim ersten Aufsatz offenbar nicht der Fall war. Bei einer (nicht repräsentativen) Stichprobe von 2501 über die Crossref-API verfügbaren Zitationen, die mit dem Stichwort „social" gefunden werden, enthalten nur 33% eine DOI. Bei immerhin 14% der Zitationen fehlt eine so grundlegende Angabe wie der Titel des zitierten Werks.

16 Das gezeigte Beispiel ist ein Ausschnitt aus einem Artikel in der Springer-Zeitschrift *Nature*, vgl. https://api.crossref.org/works/10.1038/227680a0 (zuletzt geprüft am 28.11.2017).
17 „Reference Linking," Cross Ref, zuletzt geprüft am 28.11.2017, https://www.crossref.org/services/reference-linking.

Einen Eindruck, inwieweit die vorhandenen Zitationsdaten in Crossref mit ihrer aktuellen Qualität und Abdeckung in einem bestehenden Recherchedienst für wissenschaftliche Nutzerinnen und Nutzer bereits hilfreich sein können, vermittelt das Discovery System Primo mit dem Index Primo Central. Hier sind seit Mai 2016 auch Zitationsdaten eingebunden, die von Crossref bezogen werden; aktuell (Stand 19.07.2017) handelt es nach Aussage der Firma ExLibris um 124.341.120 Verknüpfungen, die wöchentlich aktualisiert werden. An der UB Mannheim wird Primo Central zusammen mit den lokalen Mannheimer Katalogdaten für eine breite Recherche über „Aufsätze und UB-Bestand“[18] eingesetzt, wobei standardmäßig nur Bestände angezeigt werden, die in Mannheim vorhanden sind. Gibt man in dieser Suche das Stichwort „biology“ ein, so werden bei 45 der 100 ersten Einträge in der Ergebnisliste Zitationen angezeigt. Bei der Eingabe „social“ ist dies nur bei 17 von 100 Einträgen der Fall. Schränkt man die Suche auf den Ressourcentyp „Bücher“ oder auf das Format „Printmedien“ ein, so findet man bei beiden Anfragen in den ersten 100 Ergebnissen keinen einzigen Eintrag mit Zitaten.

Dieses Ergebnis zeigt sehr anschaulich die systematischen Lücken, die noch in den verfügbaren Zitationsdaten bestehen: Während elektronische Zeitschriftenartikel bereits recht gut abgedeckt sind und hier auch eine realistische Chance besteht, dass sich die Situation mit dem Voranschreiten der I4OC-Initiative noch deutlich verbessert, sind Bücher und jede Art von gedruckter Literatur bisher kaum erfasst. Inwieweit dies problematisch ist, hängt vermutlich vom Rechercheziel und vom Fachgebiet ab. Denkt man an das Schneeballsystem als Recherchemethode zurück, so wird dort allerdings in der Regel empfohlen, mit einem einschlägigen Überblickswerk oder einer inhaltlich gut passenden Monographie zu beginnen und sich über die dort zitierten Werke zu den spezielleren Abhandlungen vorzuarbeiten. Für die technische Umsetzung einer solchen Strategie in unseren Recherchesystemen fehlt derzeit die Datengrundlage.

5. Perspektiven für Bibliotheken: Das Projekt Linked Open Citation Database (LOC-DB)

An diesem Punkt setzt das DFG-Projekt Linked Open Citation Database (LOC-DB) an.[19] Das Projekt ist im Oktober 2016 gestartet und läuft 24 Monate. Beteiligte Partner sind das Deutsche Forschungsinstitut für Künstliche Intelligenz in Kaiserslautern (Prof. Andreas Dengel), die Hochschule der Medien in Stuttgart (Prof. Kai Eckert), die Deutsche Zentralbibliothek für Wirtschaftswissenschaften (ZBW) – Leibniz-Informationszentrum Wirtschaft in Kiel/Hamburg (Prof. Ansgar Scherp) und die Universitätsbibliothek Mannheim.

Ziel des Projekts ist es, nachzuweisen, dass Bibliotheken einen effizienten und nachhaltigen Beitrag zur Erschließung von Zitationsdaten liefern können. Gerade, weil bestehende Dienste bestimmte Teilbereiche bereits gut abdecken, können sich Bibliotheken darauf konzentrieren, das Vorhandene zu ergänzen und zu optimieren. Die Stärke von Bibliotheken liegt darin, dass sie umfangreiche Erfahrungen mit Erschließungsprozessen und bibliographischen Daten besitzen. Geschultes Personal ist

18 Vgl. „Primo,“ Universitätsbibliothek Mannheim, zuletzt geprüft am 28.11.2017, http://primo.bib.uni-mannheim.de/primo_library/libweb/action/search.do?mode=Basic&vid=MAN_UB&tab=man_all.
19 *LOC-DB*, https://locdb.bib.uni-mannheim.de/.

vorhanden und wird ohnehin zur Erschließung der verschiedenen Medien, die von der Bibliothek angeboten werden, eingesetzt. Es ist also naheliegend, in diese bestehenden Prozesse auch die Erfassung von Zitationsdaten einzubetten, sofern dies durch eine weitgehende Automatisierung mit vertretbarem Aufwand möglich ist.

Um dies zu erreichen, sollen, soweit möglich, bereits vorhandene Daten und automatische Methoden genutzt werden. Darüber hinaus soll jedoch eine intellektuelle Kontrolle und Korrektur erfolgen, so dass Daten mit einem zuverlässig hohen Qualitätsniveau produziert werden, die wiederum als „Goldstandard" zur Verbesserung der automatischen Erschließungsverfahren verwendet werden können. Um eine systematische Lücke in den bisher verfügbaren freien Zitationsdaten zu schließen, soll in jedem Fall auch Printliteratur einbezogen werden; letztlich muss es jedoch möglich sein, alle Publikationstypen und Medienarten in praxisgerechten Workflows adäquat zu bearbeiten.

Zu diesem Zweck wird ein Redaktionssystem entwickelt, das die verschiedenen Prozessschritte integriert und den Ablauf so effizient wie möglich unterstützt. Bei Printliteratur ist es zunächst notwendig, die enthaltenen Literaturverzeichnisse zu scannen – ähnlich wie dies bei Inhaltsverzeichnissen in einer Reihe von Bibliotheken bereits praktiziert wird. Diese Scans oder alternativ Literaturlisten, die bereits in elektronischer Form vorliegen, werden im Redaktionssystem mit den Metadaten der Publikation, aus der sie stammen, verknüpft. Anschließend werden sie mit Methoden der automatischen Texterkennung (OCR) aufbereitet, und einzelne Zitationen werden extrahiert.[20] Das System generiert auf Wunsch Vorschläge zur Verknüpfung der erkannten Daten mit bereits vorhandenen bibliographischen Daten, vorzugsweise aus Datenquellen mit hoher Datenqualität wie z.B. den deutschen Bibliotheksverbünden oder den OLC-Datenbanken des GBV, aber auch aus Crossref oder Google Scholar. Kann eine Verknüpfung mit einem hochwertigen Datensatz hergestellt werden, brauchen die Ausgangsdaten nicht mehr weiter bearbeitet zu werden, und der Prozess kann rasch abgeschlossen werden. Ist dies nicht möglich, erfolgt die Korrektur und Ergänzung der erkannten Daten manuell. Sofern bereits strukturierte elektronische Zitationsdaten aus Quellen wie Crossref oder OpenCitations vorliegen, kann die Texterkennungskomponente übersprungen werden. Eine Überprüfung, Verlinkung und ggf. Ergänzung der vorhandenen Daten ist aber aufgrund der dargestellten Qualitätsprobleme auch in diesem Fall durchaus sinnvoll.

Abbildung 4 zeigt den zentralen Bildschirm des Redaktionssystems, in dem links ein zu bearbeitender Scan angezeigt wird und rechts die durchzuführenden Prozessschritte. Die produzierten Daten aus dem LOC-DB Projekt werden als Linked Open Data unter der Lizenz CC0 bereitgestellt. Das Datenmodell folgt demjenigen von OpenCitations, mit einer geringfügigen Erweiterung, die zur Verknüpfung gescannter Literaturverzeichnisse erforderlich ist. Mit den Direktoren von OpenCitations wurde vereinbart, dass die Weiterentwicklung des Datenmodells künftig abgestimmt wird, so dass die produzierten Daten auch in Zukunft problemlos verknüpft und zusammengespielt werden können.

20 Einen Eindruck davon vermittelt Kai Eckert, Anne Lauscher und Akansha Bhardwaj, „LOC-DB: A Linked Open Citation Database Provided by Libraries: Motivation and Challenges" (Vortrag auf dem EXCITE Workshop 2017, 30.-31.03.2017), Vortragsfolien, zuletzt geprüft am 17.07.2017, https://locdb.bib.uni-mannheim.de/wordpress/wp-content/uploads/2016/11/LOC-DB@EXCITE.pdf, 25–27. Im Unterschied zu bestehenden Verfahren werden für die Extraktion der Referenzen Deep-Learning-Methoden angewendet. Eine wissenschaftliche Publikation hierzu ist in Vorbereitung.

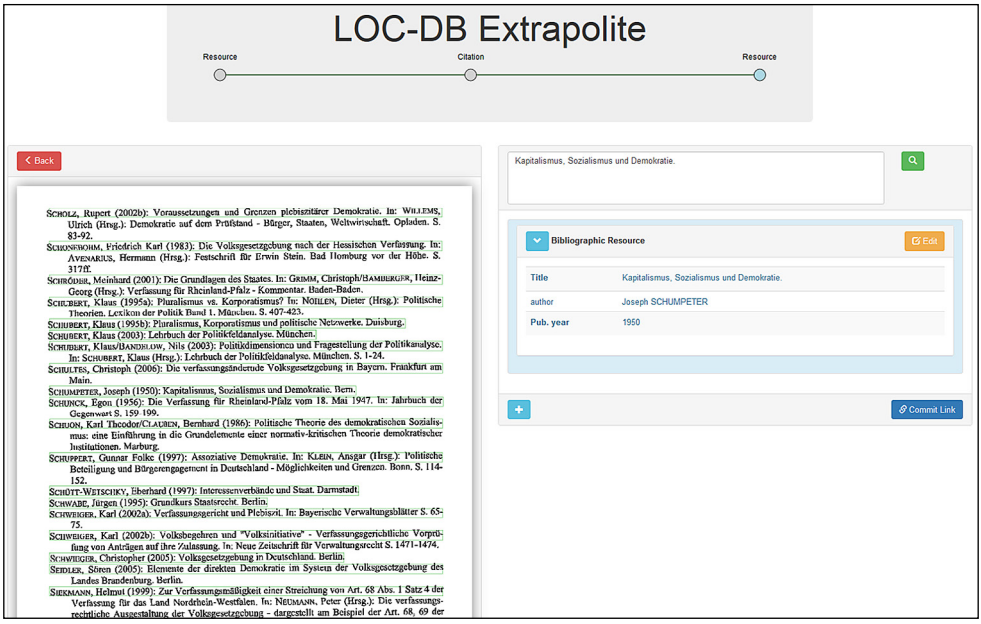

Abb. 4: Redaktionssystem von LOC-DB (erster Prototyp)

Im Rahmen eines zweijährigen Projekts, in dem zunächst eine Software entwickelt wird, ist die Anzahl der Daten, die tatsächlich produziert werden können, begrenzt. Da die Universitätsbibliothek Mannheim einen Schwerpunkt im Bereich Sozialwissenschaften hat und die Sozialwissenschaften zugleich bei den bisher vorhandenen Zitationsdaten unterrepräsentiert sind, wurde als Testkorpus ein Teilbestand dieses Fachbereichs ausgewählt: Ca. 500 Bücher im Präsenzbestand der UB Mannheim mit Erscheinungsjahr 2011 sowie der gleiche Jahrgang von 104 soziologischen Zeitschriften werden retrospektiv erfasst. Außerdem wird seit Juli 2017 der Neuzugang im Präsenzbestand Sozialwissenschaften regelmäßig bearbeitet. Im Laufe des Jahres sollen dann auch die neu erscheinenden Artikel der 104 ausgewählten Zeitschriften in die kontinuierliche Bearbeitung einbezogen werden.

Diese Stichprobe sollte ausreichend sein, um am Projektende eine belastbare Kosten-Nutzen-Analyse vorlegen zu können. Sicher ist: Keine Bibliothek wird in der Lage sein, allein alle für die Wissenschaft relevanten Zitationsdaten zu erschließen – dies ist nur kooperativ mit einer größeren Anzahl beteiligter Partner möglich. Naheliegend wäre z.B. eine Einbindung in die Fachinformationsdienste, aber auch an „normalen“ wissenschaftlichen Bibliotheken könnten Teilbestände von besonderem Interesse (z.B. die Publikationen der eigenen Hochschulangehörigen oder Spezialsammlungen und Schwerpunktfachgebiete) gezielt erschlossen werden. Diese Entscheidung kann jedoch nur getroffen werden, wenn klar ist, welche Ressourcen erforderlich sind, um eine bestimmte Menge von Daten in einer bestimmten Qualität zu produzieren.

Das Projekt LOC-DB wird seine ersten Ergebnisse in einem Workshop im Herbst 2017 präsentieren. Interessierte aus Bibliotheken oder verwandten Projekten und Entwickler, die an den eingesetzten Methoden interessiert sind, können sich bei dieser Gelegenheit einen Eindruck von der Funktionalität des Systems verschaffen.

6. Fazit

Durch die große Zahl von bereits frei verfügbaren Zitationsdaten und die dynamische Entwicklung in diesem Bereich ist eine annähernd vollständige Erschließung der Zitationen aller wissenschaftlich relevanten Publikationen in den Bereich des Möglichen gerückt. Bibliotheken könnten dabei eine wichtige Rolle spielen, und sie könnten davon erheblich profitieren: In Bibliothekskatalogen und anderen bibliothekarischen Recherchesystemen könnte durch das Einbinden von Zitationsdaten ein Mehrwert für die Nutzerinnen und Nutzer geschaffen werden. Darüber hinaus könnten freie Zitationsdaten eine transparente und erweiterbare Datengrundlage für bibliometrische Auswertungen bilden. Auch wenn die Qualität freier Zitationsdaten zum jetzigen Zeitpunkt noch nicht mit derjenigen der kommerziellen Anbieter vergleichbar ist, so führt doch jeglicher Aufwand, der in die Verbesserung freier Daten investiert wird, auch zu einem nachhaltigen Nutzen. Das Projekt LOC-DB entwickelt eine Lösung, mit der eine verteilte Infrastruktur für offene Zitationen an Bibliotheken effizient und nachhaltig realisiert werden könnte.

Literaturverzeichnis

- Eckert, Kai, Anne Lauscher und Akansha Bhardwaj. „LOC-DB: A Linked Open Citation Database Provided by Libraries: Motivation and Challenges.“ Vortrag auf dem EXCITE Workshop 2017, 30.-31.03.2017. Vortragsfolien. Zuletzt geprüft am 17.07.2017. https://locdb.bib.uni-mannheim.de/wordpress/wp-content/uploads/2016/11/LOC-DB@EXCITE.pdf.

- Fochler, Maximilian, Ulrike Felt und Ruth Müller. „Unsustainable Growth, Hyper-Competition, and Worth in Life Science Research: Narrowing Evaluative Repertoires in Doctoral and Postdoctoral Scientists' Work and Lives.“ *Minerva* 54, Nr. 2 (2016): 175–200. http://doi.org/10.1007/s11024-016-9292-y.

- Hall, Wendy. „Linked Data: The Quiet Revolution.” *ERCIM News* 96 (2014): 4. Zuletzt geprüft am 17.07.2017. https://ercim-news.ercim.eu/en96/keynote.

- Taborelli, Dario. „WikiCite: The Journey and the Road Ahead.“ Zuletzt geändert am 23.05.2017. http://doi.org/10.6084/m9.figshare.5032235.v1.

- Walker, Benjamin, Sepand Alavifard, Surain Roberts, Andrea Lanes, Tim Ramsay und Sylvain Boet. „Inter-rater Reliability of H-Index Scores Calculated by Web of Science and Scopus for Clinical Epidemiology Scientists.” *Health Information and Libraries Journal* 33, Nr. 2 (2016): 140–49. http://doi.org/10.1111/hir.12140.

