Open Data als Innovationsmotor

Experteninterview zur Freigabe bibliothekarischer Daten / Vorteile überwiegen

Open Data, also die Freigabe bibliothekarischer Daten, ist ein spannendes Thema und eröffnet Bibliotheken ganz neue Möglichkeiten. Julia Bergmann von der Zukunftswerkstatt hat beim BibCamp 4 in Hamburg mit den IT-Experten Patrick Danowski (IST Austria), Adrian Pohl (hbz) und Kai Eckert (UB Mannheim) über die Vor- und Nachteile diskutiert.

Julia Bergmann: Was versteht man eigentlich genau unter Open Data?

Adrian Pohl: Wir haben gerade in der »Working Group on Open Bibliographic Data« bei der »Open Knowledge Foundation« Prinzipien zu offenen bibliografischen Daten erarbeitet, um genau diese Frage zu beantworten. In diesen Prinzipien wird klar gemacht, was Open Data ist. Das Ganze basiert auf der allgemeinen Open-Definition, die wiederum grundlegend definiert, was »Open« bedeutet, damit nicht jeder etwas anderes darunter versteht. Open Data heißt letztlich, dass Daten dermaßen frei zugänglich sind, dass ihre Nutzung, Bearbeitung und Wiederveröffentlichung erlaubt ist. Dazu eignen sich bestimmte Lizenzen für offene Daten, die auch in den Prinzipien genannt werden. Offenheit ist erstens eine Frage des Zugangs, zweitens eine Frage des rechtlichen Status dieser Daten – der Lizenzierung – und drittens auch eine Frage der Standards: Es müssen offene Standards verwendet werden, damit jeder mit den Daten etwas anfangen kann. Werden proprietäre Datenformate verwendet, ist das problematisch.

Patrick Danowksi: Um es etwas verkürzt darzustellen: Es meint einfach Open Access zu den bibliografischen Daten, also den Daten, die Bibliotheken herstellen.

Julia Bergmann: Warum benötigen wir Open Data?

Kai Eckert: Eine grundsätzliche Idee dabei kann auf jeden Fall sein, die Hürden zu senken, um mit den Daten arbeiten zu können, sowohl innerhalb der Bibliothekswelt als auch außerhalb der Einrichtungen, die derzeit die Daten verwalten und vorhalten. Es geht einerseits darum, dass neue Anwendungen möglich sind, andererseits aber auch darum, an die Welt außerhalb der Bibliotheken anzuschließen, Verknüpfungen mit anderen Daten herzustellen und so am Ende auch Anwendungen zu erstellen, die tatsächlich über das hinausgehen, was heute möglich ist.

Patrick Danowski: Ich würde auch gern beim Open Access-Begriff anknüpfen. Klassisch wird ja gefordert: Da Wissenschaftler aus öffentlichen Mitteln finanziert werden, sollten ihre Publikationen nicht noch einmal mit öffentlichen Mitteln zurückgekauft werden. Was also öffentlich finanziert wird, soll allen zur Verfügung stehen. Und genau diese Argumentation lässt sich auch auf die bibliothekarischen Daten anwenden: Da die Bibliotheken aus öffentlichen Mitteln finanziert werden, sollten auch die von ihnen produzierten bibliographischen Daten frei zur Verfügung stehen.

Julia Bergmann: Kommen wir nun zu den Lizenzen, die ja eben schon erwähnt wurden. Man stolpert in diesem Zusammenhang über Begriffe wie die Creative Commons License, Waiver, Open Definition, PDDL, CC0. Was darf ich mir darunter vorstellen? Welche Lizenz ist hier nun die gewünschte und welche Modelle sind die favorisierten?

Patrick Danowski: Um mit der Creative Commons License zu beginnen: Der Begriff ist ein bisschen irreführend, weil man von einer Mehrzahl von Creative Commons-Lizenzen sprechen muss. Das ist ein Modell, das erarbeitet wurde, um bestimmte Rechte freizugeben oder auch vorzubehalten. Bei Bibliotheken ist es sehr beliebt, die kommerzielle Nutzung auszuschließen. Was bei Content vielleicht noch einigermaßen funktioniert, aber selbst da schon problematisch ist, ist bei Daten noch viel problematischer. Wenn man beispielsweise Daten aus verschiedenen Bereichen zusammenführen möchte, dann kommt man aufgrund verschiedener Lizenzen ganz schnell in einen Konflikt. Eine Lösung sind Verträge, die praktisch unbeschränkte Nutzungsrechte einräumen. Dies sind die sogenannten Waiver. Unter diesem Oberbegriff versteht man verschiedene Ansätze wie die Public Domain Dedication and License, kurz PDDL, oder auch die Creative Commons Zero (CC0)-Lizenz. Mit den

Waivern erklärt man: Ich gebe es annähernd in die Public Domain, ich erhebe keine Ansprüche mehr auf diese Daten. In Europa kann man ja das Urheberrecht für diese Daten eben nicht aufgeben, man kann sich nur verpflichten, dass man dieses Urheberrecht auf keinen Fall in Anspruch nehmen wird. Und genau dies erklärt man mit dem Creative Commons Zero Waiver.

Adrian Pohl: Ich würde das gerne noch ergänzen. Wir befinden uns mit dem Internet in einer vernetzten Welt, und es geht darum, über die Vernetzung und Kombination von Ressourcen, Daten und Inhalten aus verschiedenen Quellen Mehrwert zu generieren. Die technische Infrastruktur wird durch Standards ermöglicht – das sind vor allem HTTP, URIs und HTML, je nachdem kommen noch Linked-Data-Standards hinzu. Aber wir benötigen auch eine rechtliche Infrastruktur, um die Kompatibilität der verschiedenen Daten zu gewährleisten. Und deswegen gibt es die Open Definition. Das ist selbst keine Lizenz, man könnte sie aber vielleicht als Meta-Lizenz bezeichnen, die sicherstellt, dass alle Lizenzen, die der Open Definition entsprechen, miteinander kompatibel sind. Das heißt, dass man sämtliche Daten, die offen lizenziert sind, zusammenführen und gemeinsam nutzen kann, mischen kann, Mash-Ups erstellen kann und so weiter, ohne dass man sich über rechtliche Fragen Gedanken machen muss. So kann man Wissen problemlos zusammenführen, ohne sich von den Produzenten eine zusätzliche Erlaubnis holen zu müssen. Deswegen wurde die Open Definition als eine Art Meta-Lizenz entwickelt.

Julia Bergmann: Steckt hier auch die Hoffnung dahinter, dass beispielsweise kommerzielle Anbieter daraus vielleicht Dienste entwickeln, zum Beispiel für Smartphones, für andere Umgebungen. Dass also die Sichtbarkeit der bibliografischen Daten, die wir produzieren, erhöht wird?

Adrian Pohl: Im Kontext von Bibliotheken ist das sicher eine Hoffnung, die man hat. Die finanziellen Ressourcen in der Bibliothekswelt werden eher geringer, als dass es mehr werden. Wir können nicht alle Dienste, die vielleicht wünschenswert sind, selbst anbieten. Deswegen ist es auf jeden Fall auch eine strategische Entscheidung zu sagen: Wir geben diese Daten frei für alle, damit sie damit nützliche Dinge machen können.

Kai Eckert: Das möchte ich noch mal bekräftigen. Es geht ganz grundsätzlich darum, die Nutzung der Daten zu ermöglichen. Und das kann auch explizit kommerzielle Nutzung sein. Man muss sich vor Augen führen: Wenn die Daten offen sind, werden sie sozusagen zu einer Art Gemeingut, dann kann auch der reine Verkauf der Daten in dem Sinne kein Geschäftsmodell mehr sein, denn sie stehen ja jedem zur Verfügung. Aber die Möglichkeit zur Innovation, also zum Beispiel eine Dienstleistung anzubieten, die auf den Daten aufsetzt, ist jedem unbenommen. Diese Innovation ist auch ganz bewusst erwünscht, denn da wollen wir ja hin: Wir wollen Innovation, die letztlich wiederum allen nützt. Da sind auch die kommerziellen Anbieter bewusst nicht ausgeschlossen.

Patrick Danowksi: Die klassische Angst ist ja: Der kommerzielle Anbieter könnte auf Basis meiner Daten etwas entwickeln, ich habe da Arbeit reingesteckt und verdiene nichts daran mit. Die Frage ist: Was wäre die Alternative? Würde das Produkt überhaupt entwickelt, wenn diese Daten gar nicht frei zur Verfügung stehen? Würde da nicht einfach nur die Innovation gebremst? Oder sind die freien Daten nicht gerade ein Innovationsmotor dafür, dass neue Dienste entstehen, wovon im Endeffekt Bibliotheken profitieren. Es heißt ja, die Anbieter zahlen nichts dafür. Wenn sie das Produkt verkaufen, zahlen sie aber, zumindest wenn es ein deutscher Hersteller ist, auch wieder Steuern, sodass zumindest ein Teil an den Staat zurückfließt. Die Bibliothek profitieren zwar nicht unmittelbar davon, es hat aber zumindest volkswirtschaftlich positive Auswirkungen.

Kai Eckert: Ich denke, da muss man sich sehr genau anschauen, wo schon Geld verdient wird. Das ist meiner Meinung nach zweigeteilt. Wo Daten derzeit verkauft werden, muss man sich fragen, ob mit einer Datenfreigabe vielleicht ein Geschäftsmodell zusammenbricht. In dem Fall kann man natürlich nicht einfach sagen: Na ja, das Geschäftsmodell ist einfach nicht mehr zeitgemäß. Das mag so sein, ist aber ein schwieriges Thema. Wir haben aber in der Realität auch oft den Fall, dass eigentlich kein Geld mit den Daten verdient wird – gerade auch im Bibliotheksbereich und in den Verbünden – und dann eher die Angst mitschwingt, jetzt könnte ja jemand anderes Geld mit den Daten verdienen. Ich halte diese

Angst für unbegründet. Wenn es jemand schafft, mit den freien Daten tatsächlich Geld zu verdienen, dann muss eine sehr erstrebenswerte Innovation dahinterstecken. Ansonsten kann man durchaus darauf vertrauen, dass keiner einfach dadurch Geld verdienen kann, dass er sich die Daten nimmt, die jemand anderes erstellt hat – denn das wurde bisher eben auch nicht geschafft.

Julia Bergmann: Jetzt noch mal zu den Daten selber: In welcher Form liegen diese Daten im Web? Kann dieses Format jeder Nutzer einfach verwenden?

Adrian Pohl: Das ist unterschiedlich. Sowohl in der CERN-Bibliothek, die als erste ihre Daten freigegeben hat, als auch im hbz waren es eher opake, MARC-/MAB-basierte Formate, mit denen eigentlich niemand außerhalb der Bibliothekswelt etwas anfangen kann. Das kann der Bibliothekswelt natürlich auch schon Nutzen bringen – man kann mit Daten aus anderen Bibliotheken und Verbünden ja wechselseitig Daten anreichern. Aber trotzdem hat sich eine Strömung entwickelt: Open Data heißt ja erst mal nur offene Lizenzierung und Freigabe der Daten. Aber die meisten Akteure haben sich darüber hinaus zum Ziel gesetzt, Linked Data aus diesen Daten zu machen, die Daten in einem Linked-Data-kompatiblen Format herauszugeben, aufbauend auf Web-Standards wie HTTP, URIs und dem RDF-Datenmodell, um diese Daten auch miteinander zu verknüpfen und dadurch Mehrwerte zu schaffen.

Patrick Danowski: Das CERN hat deshalb erstmal ein bibliografisches Format gewählt, weil man gesehen hat: Wir haben derzeit nicht die Ressourcen, um Linked Data zu realisieren, aber vielleicht hat diese irgendjemand anderes und möchte nur ein Projekt machen: Wie transformiere ich das Format in ein Linked-Data-Format, das allgemein lesbar ist? Er kann unsere Daten direkt mit verwenden, weil diese schon da sind und kann dann seinen Algorithmus, den er entwickelt hat, darauf auch verfeinern. Das ist die Idee: Auf den Schultern von Giganten. Man macht erstmal einen kleinen Schritt, und ein Anderer kann dann vielleicht schon den nächsten Schritt machen. Das ist aber nur möglich, wenn man möglichst früh anfängt, diese Dinge zu veröffentlichen.

Kai Eckert: Genau so ein Mitspieler war die Universitätsbibliothek Mannheim, weil uns tatsächlich erstmal die technische Sicht interessiert hat. Das liegt natürlich auch in meiner Person begründet, ich bin erst seit einem Jahr an der UB und komme aus dem Semantic Web/Linked Data-Bereich. Mich haben die Daten interessiert und deswegen haben wir den Schritt gewählt, erst mal eine technische Umsetzung auszuprobieren, aus purem Eigeninteresse. Wir sind jetzt den konsequenten und richtigen Schritt gegangen, auch die Daten lizenzrechtlich freizugeben und gerade diese Arbeitsteilung macht die Sache sehr spannend: Wenn jemand sagt, ich möchte meine Daten gerne beitragen, dann besteht keine Pflicht, sie sofort in einer nutzbaren Form mit viel technischem Sachverstand aufzubereiten. Denn hinter der Aufbereitung steckt eine ganze Menge an zu klärenden Details, über die man auch noch mal lange diskutieren kann. In vielen Fällen ist auch noch sehr unklar, wie das letztlich am besten zu bewerkstelligen ist. Deswegen ist Linked Data kein zwingender Schritt für Open Data.

Patrick Danowksi: Ich möchte eine Illusion zerstören: Wenn wir Linked Data haben, können wir nicht davon ausgehen, dass sie sofort für jedermann nutzbar sein werden, sondern wir benötigen erst noch einen weiteren Entwicklungsschritt. Wir brauchen entsprechende Werkzeuge, die diese Daten aufbereiten und auf ihnen arbeiten können. Der große Vorteil ist, dass mehr Entwickler in der Lage sein werden, diese Werkzeuge zu entwickeln, weil Linked Data ein schon etablierter Standard ist, der in der Informatik breit verwendet wird. Wir haben ein größeres Entwicklerpotenzial, diese Werkzeuge zu verwenden, und unsere Daten werden auch leichter einspielbar in Werkzeuge, die schon da sind und die genau auf diesem Standard basieren.

Julia Bergmann: Lässt sich ganz kurz in einem Satz der Unterschied zwischen Open Data und Linked Data zusammenfassen?

Adrian Pohl: Es geht grundsätzlich darum, Daten im Web zu veröffentlichen. Open Data ist dazu da, rechtliche Kompatibilität zwischen Daten aus unterschiedlichen Quellen herzustellen, rechtlich die Bedingungen zu schaffen, dass sie gemeinsam genutzt und kombi-

niert werden können. Linked Data wiederum ist der technische Standard, um diese Kombination zu ermöglichen. Er legt fest, in welchem Datenmodell die Daten bereitgestellt und wie zum Beispiel Identifikatoren für verschiedene Ressourcen gebildet werden. Natürlich müssen auf einer Ebene darüber auch Vokabulare entwickelt werden, wie man die Daten konkret publiziert, was noch weitestgehend ungeklärt ist. In diesem Bereich findet noch eine Menge Diskussion und Entwicklung statt; da muss sich die Bibliothekswelt in Vielem noch einig werden.

Julia Bermgann: Wir tun also etwas in der Hoffnung, dass damit etwas passiert. Beim CERN ist es jetzt, glaube ich, schon fast ein Jahr her, dass die Daten freigegeben wurden. Ist denn schon eine erste Entwicklung da? Gibt es erste Anwendungen, die daraus entstanden sind? Oder ist die erste Entwicklung einfach nur, dass erstmal andere mitziehen?

Patrick Danowski: Ja, nach der Freigabe der CERN-Daten kam relativ schnell ein erstes Feedback von Dan Brickley, der diese Daten für eine Visualisierung benutzt hat. Es kam raus, dass in den Daten eine UDC-Stelle benutzt wurde, die in der Klassifikation nicht existiert. Durch die Visualisierung fiel der Fehler ziemlich schnell auf. Das war eine interessante Feedback-Schleife. Das war schon gut, aber ich glaube, die großen Entwicklungen kommen erst so langsam: Wir haben jetzt davon gehört, dass bei einem semantischen Projekt – dem Contentus-Projekt – durchaus Interesse besteht, diese Daten auch zu nutzen.

Kai Eckert: Ich denke auch, dass die ersten Anwendungen im Moment erstmal Basis-Anwendungen sind. Ich finde hier das Stichwort der Entzauberung ganz gut. Man muss wirklich aufpassen, dass die Anwender jetzt nicht die Erwartung haben, dass aus Linked Data sozusagen automatisch die nächste tolle Suchanwendung wird, die womöglich Google ablöst, oder was auch immer für Erwartungen mit Linked Data verbunden werden.

Wir haben an der UB Mannheim ganz konkrete Anwendungen: Aus dem Informatik-Studiengang sind Studenten an uns herangetreten und wollten mit Bibliotheksdaten arbeiten. Allein aus Performance-Gründen wollten wir keinen direkten Zugriff auf unseren Opac geben, den wir stabil betreiben müssen, aber wir konnten auf unseren Linked Data-Service verweisen. Für die Studenten hat das die Barriere enorm verringert, denn Linked Data und RDF sind ihnen aus dem Studium ein Begriff. Sie konnten dann anfangen, mit diesen Daten in einem eigenen Projekt zu arbeiten. Die andere Form der Nachnutzung ist, dass wir mit den Daten selbst in eigenen Projekten arbeiten. Die Projektergebnisse stellen wir als Linked Data zur Verfügung, wir kombinieren zum Beispiel Titeldaten miteinander oder erschließen Titel automatisch. Das sind immer relativ kleine Datenmengen, die in durchaus spannenden Projekten erstellt werden. Dabei stellt sich stets die Frage, wie man diese Ergebnisse letztlich publiziert. Man kann diese Daten nicht ohne Weiteres in eine Verbunddatenbank übernehmen, aber man kann sie als Linked Data bereitstellen, denn genau dafür ist Linked Data gemacht. Das Linked-Data-Netz ist der richtige Platz für alle möglichen Daten, sodass sie dann von jedem Interessenten nachgenutzt werden können. Solche Daten können die Basis für neue Anwendungen bilden, und an dieser Basis arbeiten wir im Moment mit unserem eigenen Linked-Data-Dienst.

Adrian Pohl: Es gibt sicher noch keine großartigen Endnutzeranwendungen. Das wird auch noch eine Weile dauern. Ein Beispiel aus der USB Köln, die auch ihre Lokaldaten freigegeben hat: Diese Daten werden in Wikipedia für eine Personensuche nachgenutzt. Dabei werden zu einer Person entsprechende Literatureinträge ergänzt. Auf Basis dieser Personensuche ist ein Nutzer darauf gestoßen, dass ein Buch mit ihm verknüpft ist, was er nicht geschrieben hat, und mit dieser Information konnte die USB Köln ihre Daten korrigieren. Das ist ein kleines Beispiel, aber es zeigt, was es bringen kann, Daten freizugeben, sie mit anderen Diensten zu vernetzen, sie zu verbreiten und dadurch auch die Daten zu verbessern. Wenn die Daten nur im Opac liegen, dann ist die Anzahl der Nutzer eher gering, geringer, als wenn sie weit verbreitet und stark mit anderen Diensten verlinkt sind.

Julia Bergmann: Zum Abschluss ein Blick in die Zukunft. Also wenn wir jetzt ein Jahr vorausblicken und annehmen, wir treffen uns beim BibCamp 5 wieder: Was hat sich bis dahin getan? Was wäre wünschenswert?

Kai Eckert: Von der Zukunft erwarte ich, dass die Daten dadurch, dass sie zum Allgemeingut werden, Innovatoren anlocken, die anfangen, spannende Anwendungen damit zu entwickeln und dass daraus Anwendungen entstehen, die wir uns im Moment tatsächlich noch nicht vorstellen können. Ich befürchte allerdings, dass das Jahr nicht ausreicht.

Adrian Pohl: Ich glaube, dass Katalogdaten verstärkt als Linked Data veröffentlicht werden, dass die dahinter liegenden Vokabulare und die Form der Daten verbessert werden. Außerdem fangen auch Wissenschaftler an, ihre Daten, wie zum Beispiel Literaturangaben und Verweise auf andere Texte, in RDF abzubilden. Ich kann mir vorstellen, dass es Links aus wissenschaftlichen Texten in Kataloge geben wird und dass man unter Umständen in dem wissenschaftlichen Text anzeigen kann, wo die referenzierte Stelle zu finden ist oder Ähnliches. Das wird sicher noch einige Zeit benötigen, aber erste Anwendungen in diese Richtung halte ich in nächster Zeit durchaus für wahrscheinlich.

Patrick Danowski: Ich würde mir wünschen, dass noch mehr Bibliotheken beim BibCamp 5 vertreten sind, die ihre Daten bereits veröffentlicht haben, und vielleicht auch noch ein paar externe Leute, die ersten Innovativen, die an den Daten interessiert sind. Dann könnten wir während des BibCamps 5 vielleicht schon über die Planung der ersten praktischen Anwendungen sprechen oder sogar schon die ersten vorhandenen Anwendungen diskutieren.

»Es geht einerseits darum, dass neue Anwendungen möglich sind, andererseits aber auch darum, an die Welt außerhalb der Bibliotheken anzuschließen.« (Kai Eckert)

»Da die Bibliotheken aus öffentlichen Mitteln finanziert werden, sollten auch die von ihnen produzierten bibliographischen Daten frei zur Verfügung stehen.« (Patrick Danowski)

»Open Data ist dazu da, rechtliche Kompatibilität zwischen Daten aus unterschiedlichen Quellen herzustellen.« (Adrian Pohl)

»Von der Zukunft erwarte ich, dass die Daten dadurch, dass sie zum Allgemeingut werden, Innovatoren anlocken, die anfangen, spannende Anwendungen damit zu entwickeln.« (Kai Eckert)

»Ich glaube, dass Katalogdaten verstärkt als Linked Data veröffentlicht werden, dass die dahinter liegenden Vokabulare und die Form der Daten verbessert werden.« (Adrian Pohl)

»Ich würde mir wünschen, dass noch mehr Bibliotheken beim BibCamp 5 vertreten sind, die ihre Daten bereits veröffentlicht haben.« (Patrick Danowski)

Info-Kasten, einspaltig

Patrick Danowski arbeitet zurzeit am Institute of Science and Technology Austria (IST Austria) in Klosterneuburg, einer Spitzenforschungseinrichtung in Österreich, er war vorher bei der Europäischen Organisation für Kernforschung CERN (Kanton Genf, Schweiz) tätig und hat dort relativ früh Daten als »Open Data« veröffentlicht.

Adrian Pohl arbeitet am Hochschulbibliothekszentrum des Landes Nordrhein-Westfalen (hbz) im Bereich »Open Data und Linked Data« und ist zudem Koordinator der »Working Group on Open Bibliographic Data« der »Open Knowledge Foundation«.

Kai Eckert ist in der Universitätsbibliothek Mannheim tätig und koordiniert dort die Linked- und mittlerweile auch Open Data-Aktivitäten.

Weiterführende Links zum Thema und den Podcast zu diesem Interview gibt es unter www.zukunftswerkstatt.org.