Bibliotheken: Wir öffnen Daten. Zum Stand der Entwicklung einer offenen Dateninfrastruktur

Adrian Pohl, Hochschulbibliothekszentrum des Landes Nordrhein-Westfalen

Zusammenfassung:
In den letzten Jahren haben sich – gerade in Deutschland – viele Bibliotheken und insbesondere Verbundzentralen entschlossen, bibliothekarische Daten unter einer offenen Lizenz zu publizieren und/oder als Linked Data aufzubereiten. Dieser Beitrag versucht, diese Aktivitäten in einem größeren Kontext zu sehen, indem er Open Data als eines von mehreren Elementen einer „offenen Dateninfrastruktur" versteht. Es werden sieben Elemente einer Dateninfrastruktur skizziert und eingeschätzt, wie offen die einzelnen Elemente gegenwärtig in Deutschland sind. Abschließend wird am Beispiel von (Open) Discovery vorgeführt, wie offene Ansätze bestehende Probleme beheben könnten und welche Handlungsmöglichkeiten es gibt.

Summary:
During the last years many libraries and especially library service centers have decided to publish library data under an open license and/or to expose it as linked data. This paper tries to put these activities in a bigger context by viewing open data as one element of an open data infrastructure. Seven elements of data infrastuctures are outlined and their approximate level of openness in Germany is assessed. Finally, it is demonstrated – using (open) discovery as an example – how open approaches can fix existing problems and which possibilities there are for practical actions.

Zitierfähiger Link (DOI): 10.5282/o-bib/2014H1S45-55
Autorenidentifikation: Pohl, Adrian: GND 14326723X,
ORCID: http://orcid.org/0000-0001-9083-7442

1. (Elektronische) Dienstleistungen in offenen Umgebungen

Jede Bibliothekarin und jeder Bibliothekar, wie auch alle Nutzerinnen und Nutzer von bibliothekarischen Angeboten wünschen sich, dass Bibliotheken ihre gesellschaftliche und kulturelle Relevanz bewahren oder gar ausbauen mögen. In Zeiten des World Wide Web und mit der Entstehung von Google, Amazon und Co. ist die zukünftige Rolle der Bibliotheken allerdings alles andere als klar umrissen. Fest steht: Um weiterhin eine wichtige gesellschaftliche und kulturelle Rolle einnehmen zu können, sind neben der Ausstattung mit ausreichenden Ressourcen mehr denn je folgende Prämissen zu erfüllen:

- kompetente, freundliche und hilfsbereite Mitarbeiterinnen und Mitarbeiter
- die Schaffung einer Arbeitsumgebung, die Motivation, Neugierde, das Interesse am Ausprobieren von Neuem und damit auch Innovation fördert
- das Angebot von nützlichen langfristigen Diensten, die einfach zu nutzen und zukunftssicher sind

Wohlgemerkt gelten diese Voraussetzungen nicht nur, aber insbesondere auch im Hinblick auf Dienstleistungen im virtuellen Raum. Um die Vorgaben auch in Zeiten sinkender Budgets erreichen zu können, ist aus Sicht der Bibliotheken besonders wichtig, dass die angebotenen Services möglichst günstig und leicht anpass- und erweiterbar sind, damit zügig auf neue Anforderungen reagiert werden kann.

Sowohl die Ziele von kompetenten, neugierigen und motivierten Mitarbeiterinnen und Mitarbeitern als auch das Ziel intuitiver, kostengünstiger und zukunftssicherer elektronischer Angebote sind freilich nie vollständig und abschließend zu erreichen. Das Hinarbeiten auf diese Ziele ist vielmehr ein immerwährender Zyklus von Analyse-, Entwicklungs-, Test- und Umsetzungsprozessen.

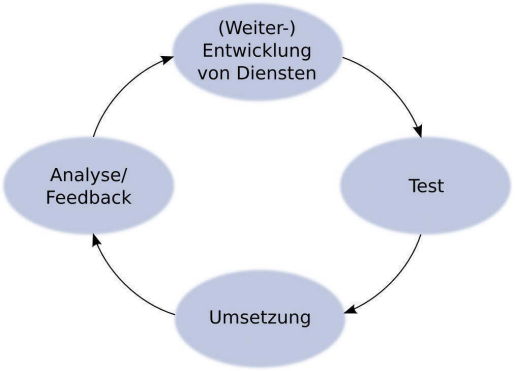

Abb. 1: (Weiter)entwicklungszyklus von Dienstleistungen

Wie schaffen wir eine Umgebung, in der Motivation und Neugierde gedeihen und erfolgreiche Angebote entstehen können? Die diesem Aufsatz zugrundeliegende These ist:

Offene Umgebungen laden ein zum Experimentieren, zum Teilen, zum Mitmachen und Zusammenarbeiten. Sie fördern Effizienz, Transparenz, Innovation und Zukunftssicherheit.

2. Sieben Elemente einer (offenen) Dateninfrastruktur

Das Thema Open Data hat in der Bibliothekswelt bereits einige Aufmerksamkeit erfahren. Eine zentrale These dieses Beitrags ist, dass Daten nur einen Teilaspekt einer jeden Dateninfrastruktur ausmachen. Im Folgenden werden – neben den Daten selbst – sechs weitere Elemente von Dateninfrastrukturen skizziert. Dabei ist es unwesentlich, ob es sich um eine offene oder aber eine geschlossene, proprietäre Infrastruktur handelt.

Daten stehen selbstverständlich im Zentrum der Infrastrukturanstrengungen, denn die möglichst schnelle, zuverlässige und fehlerfreie Übertragung von Daten ist ihr Sinn und Zweck.

Daten werden hier allgemein verstanden als Mengen von Zeichen, die elektronisch gespeichert und kopiert[1] werden können, wobei Original und Kopie nicht voneinander zu unterscheiden sind. Rechtlich betrachtet lassen sich Daten ferner grob als nicht urheberrechtlich geschützte Tatsachen den urheberrechtlich geschützten Inhalten gegenüberstellen.[2] Im Kontext bibliothekarischer

1 Kopien können sowohl innerhalb eines Dateisystems angelegt werden als auch zwischen zwei Rechnern im Internet stattfinden.

2 Auf der unterschiedlichen rechtlichen Behandlung von Daten und Inhalten beruht letztlich auch die Unterscheidung von „Open Data" und „Open Content". Zur urheberrechtlichen Betrachtung von Biblitohekskatalogdaten siehe Kreutzer, Till: Open Data – Freigabe von Daten aus Bibliothekskatalogen. Ein Leitfaden. Hg. v. Hochschulbibliothekszentrum

Dienstleistungen sind demnach insbesondere folgende Daten von Interesse[3]:

- *Titeldaten*: Beschreibungen bibliographischer Ressourcen, d. h. von Monographien und Periodika sowie von Zeitschriftenartikeln und weiteren Medien.
- *Bestandsdaten* geben an, welche Institutionen Exemplare eines Titels in ihrem Bestand haben und wo diese zu finden sind.
- *Verfügbarkeitsdaten* geben an, ob ein bestimmtes Exemplar derzeit ausleihbar, ausgeliehen, vorgemerkt etc. ist.
- *Bibliographien* sind thematische Listen von Titeldaten.
- *Kontrollierte Vokabulare* wie Normdaten, Thesauri und Klassifikationen werden etwa zur Verschlagwortung und Klassifizierung bibliographischer Ressourcen benutzt.
- *Nutzungsdaten* geben z.B. Auskunft darüber, wie häufig auf eine Online-Ressource zugegriffen wurde oder wie häufig und wann Exemplare eines bestimmten Titels ausgeliehen wurden.
- *Forschungsdaten* werden im Zuge der wissenschaftlichen Forschung durch Messung mit entsprechenden Apparaturen oder Beobachtung gewonnen. Sie werden zunehmend auch mit Unterstützung von Bibliotheken verwaltet.

Hardware: Die physische Grundlage einer jeden elektronischen Dateninfrastruktur bildet die Hardware. Für das Internet sind dies die Glasfaser- und Kupferkabelnetze, die Infrastruktur für Mobilgeräte wie UMTS-Basisstationen, außerdem die Server, Router und Switches, die an der Bereitstellung und Übertragung von Daten, der Auflösung von Namen etc. beteiligt sind, sowie die Client-Rechner, d. h. PCs oder mobile Endgeräte, die mit den Servern kommunizieren. Bei automatischer Datenerfassung wird weitere (Mess-)Hardware benötigt.

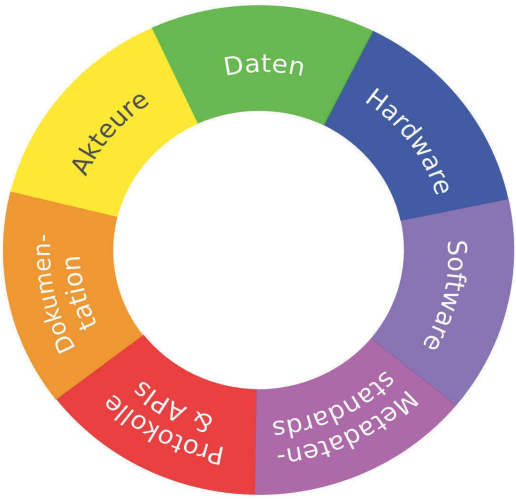

Abb. 2: Sieben Elemente einer Dateninfrastruktur

Software wird in allen Stadien des Datenlebenszyklus benötigt, um Daten überhaupt erst zu generieren, um sie zu präsentieren, zu durchsuchen, zu bearbeiten, zu visualisieren oder sie mit anderen Daten in Beziehung zu setzen. Verschiedene Softwaresysteme spielen in der bibliothekarischen Praxis eine Rolle, u. a.:

des Landes Nordrhein-Westfalen, 2011. http://www.hbz-nrw.de/dokumentencenter/veroeffentlichungen/open-data-leitfaden.pdf (30.10.2014), Abschnitt 2.3: „Der urheberrechtliche Schutz einzelner Daten".

3 Selbstverständlich ist eine Dateninfrastruktur auch nötig für die Publikation und die Nutzung von Inhalten, die als HTML-Seiten, PDF, XML-Dateien etc. daherkommen und letztlich nichts anderes sind als Mengen elektrifizierter Zeichen. Da diese „Inhalte-Infrastruktur" – solange sich Open Access nicht flächendeckend durchgesetzt hat – durch den urheberrechtlichen Rahmen erheblich verkompliziert wird, sei sie aber aus dieser Betrachtung ausgenommen.

- *Bibliothekssysteme* zur Katalogisierung und zur Verwaltung der Erwerbungen, Nutzer/innen und Ausleihen;
- *Verbundsysteme* für die kooperative Katalogisierung;
- *Discovery-Lösungen* für die Informationsrecherche;
- *Open-Access-Repository-Software* zum Aufbau von institutionellen und fachspezifischen Repositorien;
- *Werkzeuge für die Datentransformation*: einfache Skripte zur Überführung von Daten aus einem Ausgangsformat in ein gewünschtes Zielformat oder komplexere ETL-Werkzeuge (ETL= Extract Transform Load).

(Meta)datenstandards spielen eine wichtige Rolle in einer verteilten Infrastruktur mit verschiedenen kooperierenden Akteuren. Standardisierung von (Meta)daten findet auf vier verschiedenen Ebenen statt:

- *Zeichenkodierungen* sind elementare Standards für die elektronische Datenverarbeitung. Sie definieren die numerische Darstellung von Zeichen (Buchstaben, Ziffern und anderen Symbolen). Beispiele sind UTF-8, ASCII, ISO 646, ISO 8859-1.
- *Datenstrukturen* wie MAB, MARC, RDF, JSON oder CSV definieren, in welcher Struktur Daten abgelegt werden können. Häufig kann eine Datenstruktur auf verschiedene Weise serialisiert werden, d. h. eine Datenstruktur kann auf verschiedene sequenzielle Darstellungsformen (*Formate*) abgebildet werden. So gibt es neben MARC auch MARCXML und sogar MARC-in-RDF.[4] Das Resource Description Framework (RDF) wiederum lässt sich in einer Reihe verschiedener Formate wie N-Triples, RDF/XML oder Turtle serialisieren.
- *Datenmodelle, Vokabulare und Schemas* definieren Typen von Ressourcen sowie Metadatenelemente für ihre Beschreibung. Beispiele sind Dublin Core und Bibframe; aber auch Katalogisierungsregelwerke wie RAK-WB und RDA enthalten entsprechende Anteile konzeptueller Datenmodellierung.
- *Erfassungsregeln* bestimmen, wie Ressourcen (Titel, Normdaten etc.) erfasst werden, d.h. woher die Inhalte (Werte) für ein Metadatenelement genommen werden, ob und wie sie übertragen oder normiert werden und wie erfasste Ressourcen miteinander verknüpft werden. Beispiele sind RAK-WB, AACR, RDA.

Protokolle & APIs: Protokolle, oder genauer: Netzwerkprotokolle, sind Standards, die den sicheren, fehlerfreien und zuverlässigen Austausch von Daten zwischen Rechnern eines Netzes garantieren. D. h., sie regeln die Prozesse des lesenden und schreibenden Zugriffs von einen Rechner auf einen anderen. An der Datenübertragung in einem Computernetzwerk sind immer verschiedene Protokollschichten beteiligt.[5] Bekannte Protokolle sind die Interneprotokollfamilie Transmission Control Protocol / Internet Protocol (TCP/IP), das Hypertext Transfer Protocol (HTTP), das Simple Mail Transfer Protocol (SMTP) oder auch Z39.50, das zur Abfrage vieler Bibliothekskataloge genutzt werden kann.

4 Siehe http://www.marc21rdf.info/(30.10.2014).

5 Als Referenzmodell für Netzwerkprotokolle als Schichtenarchitektur hat sich seit den 1980er Jahren das OSI-Modell (Open Systems Interconnection Model) etabliert.

Aufsetzend auf den Protokollschichten können wiederum Schnittstellen zur Anwendungsprogrammierung (Application Programming Interface, API) bereitsgestellt werden. Heutzutage sind dies meist HTTP-basierte Web-APIs. Diese können systemspezifisch sein (z. B. die Linked-Open-Data-API des hbz: lobid[6]) oder auch auf geteilten Standards basieren, wie der DAIA-Spezifikation.[7] Ein API kann auf zwei Ebenen offen oder geschlossen sein: Auf der Ebene des Zugriffs (offen für alle vs. kostenpflichtig oder passwortgeschützt) sowie auf der Ebene der Dokumentation. Benutzt werden kann ein API von allen, die sowohl Zugriff auf das API als auch auf seine Dokumentation haben.

Dokumentation: Dokumentation spielt grundsätzlich nicht nur bei APIs eine essentielle Rolle, sondern auch beim Umgang mit Daten im Allgemeinen. Metadatenmappings, Mailinglisten und -archive, Ticketsysteme, Anwendungsprofile, Projektanträge, -pläne, -berichte, Lessons Learned, Konfigurationsdateien für Datentransformation usw. können eine wichtige Rolle spielen beim Verständnis eines Angebots. Eine transparente Projekt- und Angebotskultur ist ein wichtiges Element einer Dateninfrastruktur.

Akteure: Ohne konkrete institutionelle oder menschliche Akteure lässt sich keine Dateninfrastruktur aufbauen und erhalten. Im Bereich der Informationsversorgung spielen insbesondere folgende Akteure eine wichtige Rolle: Bibliotheken und Verbundzentralen, Verlage, Universitäten und Wissenschaftler/innen, Förderinstitutionen wie die Deutsche Forschungsgemeinschaft (DFG), gemeinnützige Organisationen wie Wikimedia, das Internet Archive oder die Open Knowledge Foundation sowie Unternehmen wie Google, Ex Libris oder OCLC. Selbstverständlich sind als diejenigen Akteure, für die diese gesamte Infrastruktur überhaupt aufgebaut und unterhalten wird, auch die Bibliotheksnutzerinnen und -nutzer zu nennen.

3. Status Quo

Wie offen sind die sieben Elemente der gegenwärtigen bibliothakarischen Dateninfrastruktur? Mangels Erhebungen und belastbarer Daten können wir uns hier nur ein ungefähres Bild machen. Zunächst sei aber zur groben Unterscheidung von „offen" und „geschlossen" eine allgemeine Definition gegeben, die sich sowohl auf Hardware als auch auf Software(-Code) und andere digitale Artefakte bezieht:
Etwas ist offen, wenn es selbst bzw. seine Spezifikation oder sein Bauplan

- frei zugänglich ist,
- in öffentlich dokumentierten Formaten vorliegt und
- von jeder und jedem benutzt, modifiziert und weiterverbreitet werden darf.[8]

6 Siehe http://lobid.org/ (30.10.2014).

7 Siehe http://www.gbv.de/wikis/cls/Verf%C3%BCgbarkeitsrecherche_mit_DAIA (30.10.2014), dort heißt es: „Die Document Availability Information API (DAIA) ist ein Datenmodell und eine Programmierschnittstelle (API) zur Abfrage von aktuellen Verfügbarkeitsinformationen von Dokumenten in Bibliotheken und ähnlichen Einrichtungen." DAIA wurde gemeinsam von der Verbundzentrale des GBV, der HEBIS-Verbundzentrale und dem Beluga-Projekt entwickelt.

8 Vgl. Pohl, Adrian; Danowski, Patrick: Linked Open Data in der Bibliothekswelt: Grundlagen und Überblick. In: Dies. (Hg.): (Open) Linked Data in Bibliotheken. Berlin/Boston: de Gruyter, 2013, S. 1-44, Fußnote 11.

3.1 Offenheit der technischen Basis

Diese allgemeine Definition gibt uns für sechs der sieben Dateninfrastrukturelemente Offenheitskritierien an die Hand. Dies sind die technischen Elemente: Daten, Hardware, Software, Metadatenstandards, APIs, Protokolle und Dokumentation.

Während in Deutschland bereits ein Großteil der Verbund*daten* und damit auch Bestandsdaten sowie die Gemeinsame Normdatei (GND) unter offenen Lizenzen zur freien Wiederverwendung zu Verfügung stehen, sind Metadaten zu Zeitschriftenartikeln – wie sie etwa zum Aufbau eines Discovery-Indizes unentbehrlich sind – kaum frei verfügbar und Zirkulations- sowie Nutzungsdaten im Allgemeinen gar nicht. Auch offene Fachbibliographien finden sich kaum, während es wiederum immer mehr offene Schnittstellen zur Abfrage von Verfügbarkeitsinformationen aus Bibliothekssystemen gibt.

Die in Bibliotheken und verwandten Einrichtungen eingesetzte *Hardware* dürfte zu 100 % proprietär sein, denn im Bereich der Computertechnologie haben sich – bis auf wenige Ausnahmen wie der Arduino-Plattform – kaum Open-Source-Hardware-Produkte etabliert.[9]

Im Bereich *Software* existiert bereits eine große Auswahl freier Angebote wie beispielsweise die Discovery-Software VuFind. Insbesondere im Bereich von Repository-Software gibt es bereits ein großes Angebot an Open-Source-Software, z.B. DSpace, E-Prints, Fedora oder das in Deutschland entwickelte OPUS. Auch ist es gute Praxis, Skripte und Tools zur Überführung von Daten eines Formats in ein anderes offen bereitzustellen. Mit Catmandu und Metafacture sind in den letzten Jahren überdies zwei mächtige freie Softwaretools für vielfältige Datentransformationsprozesse[10] entstanden, die in verschiedenen Institutionen eingesetzt werden. Dazu kommt noch unzählige domänenunspezifische Software wie die Suchmaschinen Elasticsearch und Solr, die ihren Einsatz auch in der Bibliothekswelt finden. Hier ist also ein durchaus positiver Trend zu verzeichnen. Bereits seit zehn Jahren gibt es freie Software für das zentrale Werkzeug einer bibliothekarischen Einrichtung: das Bibliothekssystem. Beispiele sind Evergreen und Koha, wobei letzteres in Deutschland auch in kleineren Bibliotheken eingesetzt wird. Neuerdings ist mit Kuali OLE auch ein Open-Source-System entstanden, das sich auch an größere Hochschulbibliotheken richtet. Allerdings werden hier im deutschsprachigen Raum meist proprietäre Produkte eingesetzt, wobei Angebote der Firmen OCLC und Ex Libris am weitesten verbreitet sind. Eine Tendenz zu quelloffenen Lösungen ist nicht zu erkennen, weil viele Institutionen sehr an ihre langjährigen proprietären Systemanbieter gebunden sind.

Die meisten *(Meta)datenstandards* sind bereits offen. Serialisierungen und Encodings, Metadatenschemata und Standards zur Datenstrukturierung sind meist frei zugänglich im Web definiert – wenn

9 Für den Bereich der Landmaschinen gibt es seit 2003 eine stetig wachsende Bewegung unter dem Etikett „Open Source Ecology". In Bezug auf Computerhardware hat sich eine solche Bewegung bisher nicht etablieren können, was sicher mit den enormen Kosten der Produktionsmittel in diesem Bereich zusammenhängt, die eben beispielsweise bei der Softwareentwicklung einen Bruchteil dessen ausmachen.

10 Diese Art von Softwarewerkzeugen werden auch als „ETL-Software" (für „extract, transform, load") bezeichnet.

auch oft nicht unter einer offenen Lizenz.[11] Allerdings erleben wir gerade in Bezug auf die Erfassungsregeln einen Rückschritt im deutschsprachigen Raum. Während die RAK-WB frei zugänglich im Web vorliegen, sind die zukünftig anzuwendenden Erschließungsregeln RDA (Resource Description and Access) hinter einer Bezahlschranke versteckt und somit nicht für jeden von überall einsehbar.

Auch auf der Ebene der *Protokolle* haben wir es bereits mit einer weitestgehend offenen Infrastruktur zu tun. Zentrale Internet- und Web-Standards werden von der Internet Engineering Task Force (IETF) oder dem World Wide Web Consortium (W3C) publiziert, wichtige bibliothekarische Standards von der NISO (National Information Standards Organization). Auf der anderen Seite finden sich aber auch proprietäre Protokolle wie das Simple Library Network Protocol (SLNP), dessen Dokumentation von OCLC – etwa durch den Einsatz von Geheimhaltungserklärungen – streng behütet wird.

Die *Dokumentation* von Aktivitäten im Bereich Datenbereitstellung und Datenmanagement findet in unterschiedlichem Maße offen statt – je nachdem mit welchen Akteuren man es zu tun hat. Offene Mailinglistenarchive sind im Bibliotheksbereich häufig zu finden, aber auch geschlossene, selbst wenn die Listen offen zum Abonnement durch alle sind. Auch Sitzungsprotokolle von Arbeitsgruppen im Bereich Metadaten werden nicht immer offen im Web veröffentlicht. Projektanträge, -pläne und -berichte werden in der Regel gar nicht im Netz publiziert.

3.2 Offenheit von Institutionen und Individuen

Die oben gegebene Definition von Offenheit lässt sich zwar auf Texte und technische Artefakte anwenden, für die Betrachtung konkreter Akteure (Institutionen, Individuen) taugt sie allerdings wenig. Man könnte zwar behaupten, ein Akteur sei in dem Maße offen, wie er/sie offene Artefakte nutzt und selbst bereitstellt. Zweifellos spielt dieser Aspekt eine wichtige Rolle bei der Einschätzung der Offenheit eines Akteurs, allerdings erschöpft sich seine Bewertung nicht darin.

Im Umgang von Institutionen und Individuen mit- und untereinander sind weitere Aspekte von Offenheit von zentraler Wichtigkeit: die Transparenz einer Institution und ihrer Abläufe; die Bereitschaft zur Zusammenarbeit und zur Einbeziehung von Interessierten in ein Projekt oder einen Entwicklungsprozess; die Bereitschaft aktiv nach bestehenden, anderswo entwickelten wiederverwendbaren Modulen zu suchen und bei deren Weiterentwicklung zu helfen etc.[12] Zweifelsohne gibt es hier in allen bibliothekarischen Institutionen im deutschsprachigen Raum noch erhebliches Entwicklungspotential.

4. Das Beispiel (Open) Discovery

Zur Konkretisierung dieser Überlegungen soll hier nun das Beispiel von Discovery-Services näher betrachtet werden. Zunächst stellt sich die Frage, inwiefern derzeitige Discovery-Ansätze die anfangs

11 Unter Umständen könnte eine offene Lizenz bei manchen Standards sogar kontraproduktiv sei. Zumindest muss sichergestellt sein, dass es eine klar als solche angebotene kanonische Version gibt.

12 Vgl. den Entwurf eines „Libraries Empowerment Manifesto", das versucht, die verschiedenen Aspekte von Offenheit, die zum Aufbau einer zukunftssicheren, innovativen Bibliotheksinfrastruktur nötig sind, kurz und bündig zusammenzufassen: http://etherpad.lobid.org/p/LEM.

genannten Zielsetzungen erreichen. Handelt es sich um nützliche, stabile Services? Sind sie einfach zu nutzen und dazu kostengünstig? Lassen sie sich leicht anpassen und erweitern?

Bei den derzeitigen Ansätzen, die sämtlich von kommerziellen Herstellern angeboten werden[13], lassen sich folgende Probleme ausmachen:

- Daten und Software sind nicht offen und nicht durch jeden Interessierten nachnutzbar.
- Selbst wenn eine Software lizenziert wurde, kann sie nicht – oder nur mit viel Aufwand und unter gewissen Risiken – lokal erweitert oder angepasst werden.
- Die Lizenzierung von Discovery-Indizes ist kostenintensiv.
- Discovery-Services sind intransparent z.B. in Hinblick auf die konkreten Retrieval- und Rankingmechanismen.
- Es ist nicht einmal sichergestellt, dass ein eingekaufter Discovery-Service auch Zugriff auf sämtliche durch die jeweilige Institution gesicherten Inhalte bietet.[14]

Dieser Text plädiert dafür, das Angebot von Discovery-Services nicht auf einige wenige kommerzielle Akteure zu beschränken, sondern mittel- bis langfristig eine offene Dateninfrastruktur aufzubauen, die Zugriff auf relevante Daten und Softwarewerkzeuge für alle ermöglicht. Die Verwendung verteilt vorliegender Daten für die Indexierung in beliebige Discovery-Services setzt zwei Dinge voraus:

- Die Nutzung der Daten ist rechtlich gestattet.
- Die Daten liegen in einem offen dokumentierten, leicht zu verarbeitenden wohlstrukturierten Format vor, so dass sie mit wenig Aufwand indexiert werden können.

Dem ersten Desiderat tragen offen lizenzierte Daten (Open Data) Rechnung, während sich der zweite Punkt mit Linked Data erfüllen lässt.

Effekte von Linked Open Data

Der Sinn offen lizenzierter Daten besteht darin, dass es eine explizite Erlaubnis gibt, Daten zu kopieren, anzupassen, sie mit anderen offenen Daten zu kombinieren und in einer Suchmaschine zu indexieren. Würden Verlage, Universitäten, Wissenschaftler/innen und Bibliotheken bibliographische Daten in großem Umfang offen lizenziert bereitstellen, könnten nützliche Discovery Services aufgebaut werden – sei es fachspezifisch[15] oder -übergreifend. Die Bereitstellung von strukturierten Daten nach Linked-Data-Prinzipien[16] würde eine einfache Verarbeitung wie auch Anreicherung der Daten auf Basis weiterer verlinkter Datenquellen ermöglichen.

13 Die meistgenutzten Discovery-Dienste in Deutschland sind EBSCO Discovery Service (EDS), Ex Libris Primo und Primo Central sowie Summon von Serials Solutions.

14 Die Debatte um EBSCO-Metadaten in Ex Libris' Discovery-Service Primo zeigt dies deutlich, siehe Orbis Cascade Alliance: EBSCO and Ex Libris [Webseite]. https://www.orbiscascade.org/ebsco-ex-libris/ (22.09.2014) und Pohl, Adrian: Discovery silos vs. the open web. In: Open Bibliography and Open Bibliographic Data [Blog], 2013-06-23. http://openbiblio.net/2013/06/23/discovery-silos-vs-the-open-web/ (22.09.2014).

15 Vgl. auch Christoph, Pascal; Pohl, Adrian: Dezentral, offen, vernetzt – Überlegungen zum Aufbau eines LOD-basierten FID-Fachinformationssystems. In: Bibliothek Forschung und Praxis 38, 1 (2014), S. 114-123. http://dx.doi.org/10.1515/bfp-2014-0005.
Open-Access-Preprint: https://wiki1.hbz-nrw.de/x/EYOf (30.10.2014).

16 Vgl. Berners-Lee, Tim: Linked Data - Design Issues. 2006, letzte Änderung 2010. http://www.w3.org/DesignIssues/LinkedData.html (22.09.2014).

Ein positiver „Neben"-Effekt wäre – insofern die Daten unter Benutzung des weit verbreiteten schema.org-Vokabulars publiziert würden –, dass auch Internetsuchmaschinen wie Google, Yahoo! und Bing mit diesen Daten etwas anfangen könnten. Das hieße zum einen, dass die Bibliotheksnutzer/innen da abgeholt werden, wo sie sich ohnehin befinden: bei Google. Zum anderen bieten bibliographische Daten mit schema.org-Markup auch für Bibliotheken einige Möglichkeiten. Etwa gibt es bereits erste Experimente, auf dieser Basis mit wenig Aufwand Verbundkataloge – z. B. mit den Beständen von Bibliotheken aus einer Stadt – aufzubauen oder einen kostengünstigen OPAC auf Basis von Googles Custom Search anzubieten.[17]

5. Was tun?

Was sind konkrete Schritte, die Bibliotheken[18] gehen können, um zum Aufbau einer offenen Discovery-Infrastruktur beizutragen?

Daten freigeben: Für den Aufbau von Discovery-Services sind insbesondere Metadaten für Zeitschriftenartikel wichtig. Wenn auch die Katalogisierung von Aufsätzen in den meisten Bibliotheken nicht zum täglichen Geschäft gehört, gibt es dennoch einige Bibliothekskataloge, die Aufsatzmetadaten in teilweise erheblichem Umfang enthalten – seien dies Kataloge von Spezialbibliotheken wie z. B. Parlaments- und Behördenbibliotheken, Virtuelle Fachbibliotheken, Fachbibliographien oder die Metadaten, die in Open-Access-Repositorien entstehen. Ein wichtiger erster Schritt ist es, diese Daten unter einer offenen Lizenz bereitzustellen – ganz gleich ob als PDF-Datei, in XML, MAB oder RDF.

LOD veröffentlichen: In einem zweiten Schritt können diese Daten dann – von der jeweiligen Bibliothek selbst, einer Verbundzentrale oder anderen Interessierten – nach Linked Open Data überführt werden, indem Uniform Resource Identifier (URIs) für die einzelnen bibliographischen Ressourcen angelegt und sie mit RDF beschrieben werden. Dabei bietet sich – soweit möglich – die Nutzung des schema.org-Vokabulars an, das sich insbesondere seit den Ergänzungen durch die W3C Schema Bib Extend Community Group[19] recht gut zur Angabe der wichtigsten Informationen eignet und das darüber hinaus – wie oben erwähnt – einige Vorteile zur Auffindbarkeit bibliothekarischer Daten über große Internetsuchmaschinen bietet. Beispiele und Hinweise zur Anwendung von schema.org für bibliographische Daten bieten die Wikiseiten „Recipes and Guidelines" der Schema Bib Extend Community Group.[20]

17 Vgl. hierzu Scott, Dan: Tales of a semantic web dropout (or what I meant to say at code4lib 2014). In: Coffe|Code [Blog], 2014-04-02. https://coffeecode.net/archives/286-Tales-of-a-semantic-web-dropout-or-what-I-meant-to-say-at-code-4lib-2014.html (22.09.2014) und Aery, Shawn: Schema.org and Google for Local Discovery: Some Key Takeaways. In: Bitstreams [Blog], 2014-03-27. http://blogs.library.duke.edu/bitstreams/2014/03/27/schema-org-and-google-for-local-discovery-some-key-takeaways/ (22.09.2014).

18 Dies bezieht sich nicht ausschließlich auf Bibliotheken und Bibliotheksverbünde. Lukas Koster hat auf der diesjährigen Konferenz der International Group of Ex Libris Users (IGeLU) dargelegt, dass auch kommerzielle Anbieter wie Ex Libris vom Aufbau einer offenen, unabhängigen, transparenten webbasierten Dateninfrastruktur profitieren könnten, vgl. Koster, Lukas: Relevance redefined [Präsentationsfolien]. Präsentation auf der IGeLU-Konferenz in Oxford, 2014. http://de.slideshare.net/lukask/relevance-redefined.

19 Siehe etwa Wallis, Richard; Scott, Dan: Schema.org Support for Bibliographic Relationships and Periodicals. In: schema blog [Blog], 2014-09-02. http://blog.schema.org/2014/09/schemaorg-support-for-bibliographic_2.html (22.09.2014

20 Siehe https://www.w3.org/community/schemabibex/wiki/Recipes_and_Guidelines (30.10.2014).

(Linked) Open Data einfordern: Es ist fraglich, ob sich ein Discovery-Service theoretisch komplett aus existierenden Bibliothekskatalogdaten befüllen lassen kann. In einer funktionierenden Open-Discovery-Infrastruktur sollten deshalb auch und gerade die Inhalteanbieter selbst Metadaten offen und strukturiert zur Verfügung stellen. Bibliotheken als ihre Kunden sollten sie zu einer solchen Praxis anhalten – schließlich verdanken die Inhalteanbieter den Bibliotheken einen Großteil ihrer teilweise enormen Profite. Bibliotheken könnten etwa bei der Lizenzierung von Inhalten – seien es Zeitschriften, E-Books usw. – die freie Nutzbarkeit der zugehörigen Metadaten vertraglich regeln. Entsprechende Mustervertragsklauseln könnten zur einfachen Wiederverwendung im Web veröffentlicht werden. Darüber hinaus sollten Bibliotheken die Publikation dieser Metadaten im Web in Verbindung mit den eigentlichen (kostenpflichtigen) Inhalten unter Anwendung domänenübergreifender Web-Standards (wie RDF) verlangen.

Offene Suchmaschinenindexe aufbauen: Auf Basis der von vielerlei Akteuren bereitgestellten offen lizensierten Daten können schließlich Akteure wie etwa Fachinformationsdienste oder Verbundzentralen offene Discovery-Indexe aufbauen. Dabei muss die Aktualität und Homogenität der gesammelten Daten sichergestellt werden. Ein solcher offener Index wäre dann die Basis für übergreifende oder fachspezifische Discovery-Dienste und darüber hinausgehende Angebote. Im Sinne der Offenheit sollten dabei die Workflows offen dokumentiert und die verwendete Software und Konfigurationsdateien geteilt werden.

6. Fazit

Offene Ansätze könnten nicht nur das dringend benötigte Alleinstellungsmerkmal von Bibliotheken im „digitalen" Zeitalter werden, sondern auch die Qualität bibliothekarischer Dienstleistungen in Zeiten knapper Ressourcen verbessern. Zwar hat sich in den letzten Jahren eine Praxis der Publikation von Daten unter offenen Lizenzen etabliert, Daten stellen allerdings nur ein – wenn auch zentrales – Element einer Dateninfrastruktur dar. Beim weiteren Ausbau einer offenen Dateninfrastruktur müssen die übrigen Elemente (Software, Hardware, (Meta)datenstandards, Protokolle, APIs und Dokumentation sowie eine offene Organisationskultur ausreichende Berücksichtigung finden.

Literaturverzeichnis:

- Aery, Shawn: Schema.org and Google for Local Discovery: Some Key Takeaways. In: Bitstreams [Blog], 2014-03-27. http://blogs.library.duke.edu/bitstreams/2014/03/27/schema-org-and-google-for-local-discovery-some-key-takeaways/ (22.09.2014).

- Berners-Lee, Tim: Linked Data - Design Issues. 2006, letzte Änderung 2010. http://www.w3.org/DesignIssues/LinkedData.html (22.09.2014).

- Christoph, Pascal; Pohl, Adrian: Dezentral, offen, vernetzt – Überlegungen zum Aufbau eines LOD-basierten FID-Fachinformationssystems. In: Bibliothek Forschung und Praxis 38, 1 (2014), S. 114-123. http://dx.doi.org/10.1515/bfp-2014-0005. Open-Access-Preprint: https://wiki1.hbz-nrw.de/x/EYOf (30.10.2014)

– Koster, Lukas: Relevance redefined [Präsentationsfolien]. Präsentation auf der IGeLU-Konferenz in Oxford, 2014. http://de.slideshare.net/lukask/relevance-redefined (30.10.2014).

– Kreutzer, Till: Open Data – Freigabe von Daten aus Bibliothekskatalogen. Ein Leitfaden. Hg. v. Hochschulbibliothekszentrum des Landes Nordrhein-Westfalen, 2011. http://www.hbz-nrw.de/dokumentencenter/veroeffentlichungen/open-data-leitfaden.pdf (30.10.2014).

– Lohmeier, Felix; Mittelbach, Jens: Offenheit statt Bündniszwang. In: Zeitschrift für Bibliothekswesen und Bibliographie 61, 4–5 (2014), S.209–215 [im Druck]. Preprint: http://jensmittelbach.de/preprints/Offenheit.pdf (22.09.2014).

– Orbis Cascade Alliance: EBSCO and Ex Libris [Webseite]. https://www.orbiscascade.org/ebsco-ex-libris/ (22.09.2014).

– Pohl, Adrian: Discovery silos vs. the open web. In: Open Bibliography and Open Bibliographic Data [Blog], 2013-06-23. http://openbiblio.net/2013/06/23/discovery-silos-vs-the-open-web/ (22.09.2014).

– Pohl, Adrian / Danowski, Patrick: Linked Open Data in der Bibliothekswelt: Grundlagen und Überblick. In: Dies. (Hg.): (Open) Linked Data in Bibliotheken, Berlin/Boston: de Gruyter, 2013, S. 1-44. http://dx.doi.org/10.1515/9783110278736.1.

– Scott, Dan: Tales of a semantic web dropout (or what I meant to say at code4lib 2014). In: Coffe|Code [Blog], 2014-04-02. https://coffeecode.net/archives/286-Tales-of-a-semantic-web-dropout-or-what-I-meant-to-say-at-code4lib-2014.html (22.09.2014).

– Wallis, Richard; Scott, Dan: Schema.org Support for Bibliographic Relationships and Periodicals. In: schema blog [Blog], 2014-09-02. http://blog.schema.org/2014/09/schemaorg-support-for-bibliographic_2.html (22.09.2014).