Abschlussarbeit zum Master of Advanced Studies in Bibliotheks- und Informationswissenschaften, Universität Zürich

Nachnutzung und Anreicherung von Metadaten aus institutionellen Open Access Repositorien

Unter Berücksichtigung von Technologien des Semantic Web und Linked Open Data

Referent: **Tobias Viegener**, Schweizerische Nationalbibliothek, Bern

Koreferent: **Dr. Dirk Verdicchio**, Universitätsbibliothek Bern

Dominique Blaser
Universitätsbibliothek Bern

Feerstrasse 2
5000 Aarau
blaser.dominique@gmail.com

März 2015

Abstract

Institutionelle Open Access Repositorien sind einerseits Evaluationsinstrument wie auch Präsentationsplattform für die Trägerinstitution, zum anderen sind sie dem Open Access Gedanken verpflichtet. Somit sind Offenheit und Sichtbarkeit von Grund auf zentrale Punkte in der Bereitstellung von Repositorien-Inhalten. Dennoch kommen Repositorien nicht umhin, sich weiterzuentwickeln. Dabei werden die verbreiteten Standards OAI-PMH und Dublin Core durch ein vermehrt semantisches World Wide Web herausgefordert. Techniken des Semantic Web und Linked Open Data bieten sich für die Anreicherung und bessere Verbreitung der Metadaten an.

Ausgehend von aktuellen Arbeiten zur Integration von Inhalten des Berner Universitäts-repositoriums in den Resource Discovery Service der Universitätsbibliothek skizziert diese Arbeit verschiedene Nachnutzungsszenarios, wie sie sich auch für institutionelle Repositorien andernorts präsentieren können. Punktuell Bezug nehmend auf die konkreten Gegebenheiten am Berner Repositorium werden Möglichkeiten zur Steigerung der Sichtbarkeit von Repositorieninhalten in verschiedenen Rechercheumgebungen diskutiert, wobei die Reichhaltigkeit und die Attraktivität der Metadaten für nachnutzende Systeme im Zentrum der Überlegungen stehen.

1. Einleitung

Ausgehend von aktuellen Arbeiten zur Integration von Inhalten des Berner Universitäts-repositoriums in den Resource Discovery Service der Universitätsbibliothek skizziert diese Arbeit verschiedene Nachnutzungsszenarios, wie sie sich auch für institutionelle Repositorien andernorts präsentieren können. Punktuell Bezug nehmend auf die konkreten Gegebenheiten am Berner Repositorium werden Möglichkeiten zur Steigerung der Sichtbarkeit von Repositorien-Inhalten in verschiedenen Rechercheumgebungen diskutiert, wobei die Reichhaltigkeit und die Attraktivität der Metadaten für nachnutzende Systeme im Zentrum der Überlegungen stehen. Im Hinblick auf eine möglichst breite Nachnutzung kommen insbesondere auch Technologien des Semantic Web und Linked Open Data in den Fokus.

2. Hintergrund, Ausgangssituation

Die Universität Bern hat Open Access zu ihrem strategischen Ziel erklärt und 2007 die Berliner Deklaration[1] unterzeichnet. Sie unterstützt die Umsetzung ihrer Open Access-Strategie u.a. indem sie ein institutionelles Repositorium (IR) betreibt und ihre Forschenden verpflichtet, ihre wissenschaftlichen Arbeiten darin zu hinterlegen.[2] Seit Oktober 2013 besteht das Bern Open Repository and Information System (BORIS) als online und offen zugängliches Repositorium, das die Metadaten zu allen von Angehörigen der Universität sowie des Inselspitals verfassten Publikationen unter der Creative Commons Lizenz CC0 frei zugänglich bereitstellt.[3] Die Volltexte der Publikationen selber werden ebenfalls in BORIS abgelegt und wo immer möglich – unter Wahrung der lizenzrechtlichen Vorgaben der Verlage – für die Öffentlichkeit frei zugänglich gemacht.

In erster Linie dient BORIS also dem Ziel des möglichst offenen Zugangs zu den Forschungsergebnissen der Universität Bern. Das OA IR der Universität fördert nicht nur den freien Zugang, sondern auch die weltweite Sichtbarkeit und somit auch Zitierhäufigkeit der Berner Forschungsleistungen. Gleichzeitig dient BORIS aber dem Vizerektorat Forschung auch intern als Instrument zur Evaluation der Gliederungseinheiten und Universitätsangehörigen. Die Metadaten

[1] „Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities", zugegriffen 1. Februar 2015, http://openaccess.mpg.de/Berlin-Declaration.

[2] „Open Access-Policy der Universität Bern", 29. April 2013, http://www.ub.unibe.ch/openaccess/content/open_access_policy/index_ger.html.

[3] Vgl. die unter der OAI-Base-URL von BORIS einsehbare Metadata Policy: „OAI 2.0 Request Results", zugegriffen 23. Februar 2015, http://boris.unibe.ch/cgi/oai2?verb=Identify.

der an der Universität erstellten, begutachteten und publizierten Arbeiten werden ab Berichtsjahr 2014 aus BORIS in die Personaldatenbank übernommen und bilden die Grundlage für die akademischen Berichte sowie die Erfassung des Leistungsparameters Publikationsoutput für die einzelnen Institute, Fakultäten und Kliniken.

Der Betrieb und die Weiterentwicklung von BORIS sind mit der Koordinationsstelle Open Access in der Universitätsbibliothek angesiedelt. Zwar geschieht die Datenerfassung zu den Publikationen dezentral und wird meist durch die Forschenden selbst oder durch weitere Berechtigte vorgenommen. Vor der Freischaltung der Einträge prüft die UB Bern die bibliographischen Metadaten aber nochmals auf Vollständigkeit und Qualität und übernimmt auch die Abklärung der Zweitveröffentlichungsrechte für die abgelegten Volltexte. Dabei kommen zum einen die klassischen bibliothekarischen Kompetenzen in den Bereichen Metadatenpflege und Autoritätskontrolle sowie Erwerbung und Erhaltung elektronischer Medien, aber auch im Zeitschriften- und Lizenzwesen zum Tragen. Zum anderen nimmt die UB Bern mit dem Betrieb des IR und den damit verbundenen Schulungs- und Beratungsangeboten auch eine Gelegenheit wahr, sich innerhalb der Universität als wichtige Dienstleisterin zu positionieren.

Dass die UB Bern diese Dienstleistung nicht über ihr angestammtes Bibliotheksverwaltungssystem abwickelt, erklärt sich hauptsächlich aus dem Open Access Gedanken: die Forschungsergebnisse der Universität sollen für alle im WWW möglichst sichtbar und einfach zugänglich sein. Genau dieser Forderung lässt sich aber mit dem alleinigen Verzeichnen im Bibliothekskatalog nicht optimal nachkommen, da dieser die Metadaten im Deep Web versteckt, wo sie von Websuchmaschinen nicht indexiert und somit von deren Anwendern nicht gefunden werden können. Schliesslich ist die erhöhte Sichtbarkeit und resultierende Zitierhäufigkeit auch ein Hauptargument und Versprechen gegenüber Forschenden, um sie zur Unterstützung der OA-Strategie zu motivieren, also ihre Arbeiten möglichst OA zu publizieren und im IR abzulegen. Im Vergleich zum klassischen OPAC haben offene institutionelle Repositorien die besseren Voraussetzungen, dieses Versprechen zu erfüllen. Sie halten ihre Metadaten auf HTML codierten Webseiten vor, die von Suchmaschinen wie Google Scholar gecrawlt und indexiert werden.[4] Zudem unterstützen OA Repositorien standardmässig das Austauschprotokoll OAI-PMH[5], wodurch die Anbindung an wichtige wissenschaftliche Service Provider wie die Bielefeld Academic Search Engine (BASE)[6] sichergestellt wird.

[4] Die Zugriffsstatistik über die Betriebszeit von Oktober 2013 bis Januar 2015 belegt, dass der grösste Anteil (43.1 %) der Zugriffe auf BORIS von Google ausgeht. Dass dabei die internationale Sichtbarkeit auch erfüllt ist, zeigt die Auflistung der zahlenmässig wichtigsten Herkunftsländer der Anfragen: Schweiz (34'422), Deutschland (7'380), USA (2'077), Grossbritannien (974). Vgl. dazu: openaccess-unibe.atlassian.net/wiki/x/AgCuAg ; zugegriffen 8. Februar 2015)

[5] „Open Archives Initiative Protocol for Metadata Harvesting", zugegriffen 8. Februar 2015, http://www.openarchives.org/pmh/.

[6] https://www.base-search.net ; BASE ist eine der weltweit grössten wissenschaftlichen Suchmaschinen und harvestet mehr als 3000 Quellen – hauptsächlich über die Schnittstelle OAI-PMH. Vgl.: „BASE Weblog: 10 Jahre BASE", zugegriffen 8. Februar 2015, http://ekvv.uni-bielefeld.de/blog/base/entry/10_jahre_suchmaschine_base.

Obwohl nun mit der Infrastruktur von BORIS die Basis zur Erfüllung der Forderung nach offenem Zugang, weltweiter Sichtbarkeit und erhöhter Zitierhäufigkeit gelegt ist, gibt es gute Gründe, die Inhalte von BORIS auch in bibliothekarischen Resources Discovery Services (RDS) nachzuweisen. Denn indem die BORIS-Metadaten in den Metakatalog der Schweizer Hochschulbibliotheken und der Schweizerischen Nationalbibliothek swissbib.ch integriert werden, weitet sich die Sichtbarkeit von Berner Forschungspublikationen nochmals auf ein zentrales nationales Rechercheportal aus. Zudem sind diesen Weg schon zwei universitäre IR vorausgegangen. ZORA und SERVAL haben ihre OA-Volltexte bereits in swissbib.ch recherchierbar gemacht, was durchaus weitere Repositorienbetreiber zu diesem Schritt animieren mag.[7] Ausserdem besteht seit der Einbindung von Indices unselbständiger Literatur im neuen RDS des IDS Basel Bern (baselbernswissbib.ch, 2. Suchtab „Artikel & mehr") ein gewisser Widerspruch darin, dass Artikelpublikationen der Berner Forschenden nicht auch systematisch im RDS der eigenen Universitätsbibliothek nachgewiesen werden.

Die Nachnutzung der BORIS-Inhalte in den verschiedenen Indices des RDS swissbib Basel Bern[8] wird am besten unterstützt durch möglichst reiche Metadaten, die gut auf den jeweiligen Ziel-Index gemappt und möglichst nahtlos in die jeweilige Facettierung eingebunden werden können. Der Wert von bibliographischen Metadaten für die Nachnutzung erhöht sich, wenn sie möglichst alle kontrollierten Erschliessungsvokabulare aufweisen, die auch im nachnutzenden Index verwendet werden. Denn nur wenn Fremddatensätze nicht nur Abstracts und andere freitextliche Erschliessungsdaten, sondern auch Beschlagwortungen (GND, MeSH, LCSH), Klassifikationen (DDC, LCC, RVK) und normierte Autoren-Eintragungen enthalten, können sie überhaupt so eingebunden werden, dass im Zielsystem einheitlich über den gesamten Datenbestand gefiltert werden kann.

Im Hinblick auf Nachnutzung kann also die Grundannahme gesetzt werden: je höher die Qualität und Reichhaltigkeit der Metadaten, desto besser sind die Voraussetzungen für die Integration in nachnutzende Systeme, und umso attraktiver sind solche Metadaten für potentielle Nachnutzer grundsätzlich. Denn nicht nur die eine hohe Sichtbarkeit durch offenen Zugang, sondern auch eine hohe Attraktivität der angebotenen Inhalte und Metadaten per se sind für eine weite Verbreitung über möglichst viele Kanäle im Sinne des Open Access Gedankens wichtig. Je besser Metadaten strukturiert sind, je mehr eindeutig referenzierte Identifikatoren und kontrolliertes Vokabular sie beinhalten, desto anschlussfähiger und desto besser maschinell zu verarbeiten sind sie in nachnutzenden Systemen. Dies gilt insbesondere auch für Szenarien, wo Quelle und Ziel der Datennachnutzung nicht unter dem selben institutionellen Dach beheimatet und

[7] Zurich Open Repository and Archive (http://www.zora.uzh.ch) ; Serveur académique lausannois (http://serval.unil.ch/).

[8] Zum einen den swissbib.ch zugrundeliegenden Index, zum anderen den Summon Index des kommerziellen Anbieters ProQuest.

dementsprechend die Bedürfnisse des Zielsystems weniger einfach im direkten Dialog mit der Datenquelle abgestimmt werden können.[9]

Vor diesem Hintergrund lässt sich die Leitfrage für die nachfolgenden Überlegungen formulieren: Wie ist das Ziel möglichst reicher, qualitativ hochwertiger Metadaten in einem institutionellen Open Access Repositorium (im konkreten Fall von BORIS) zu erreichen, und wie kann dabei deren Erfassung und Anreicherung technologisch unterstützt werden? Das Ziel möglichst einfacher und automatisierter Erfassung und Anreicherung der Metadaten im IR unterstützt nicht nur die Funktion des IR als Evaluations-, Nachweis- und Verbreitungsinstrument für die universitären Publikationen, sondern dient vor allem auch dem Bestreben, die Metadaten zur Ausweitung der Sichtbarkeit und Interoperabilität für weitere Resources Discovery Services und Suchmaschinen ausserhalb der eigenen Institution attraktiver zu machen.

3. Nachnutzungsszenarien für Metadaten aus einem Institutional Repository (IR)

Im folgenden sollen drei grundlegende Szenarien der Integration von IR-Metadaten in ein nachnutzendes System exemplarisch dargestellt werden. Diese Szenarien kommen bei BORIS teilweise zum Tragen, zum Teil sind sie zurzeit aber noch nicht umgesetzt, respektive aufgrund aktueller technischer Voraussetzungen noch nicht realisierbar. Vorteile und Nachteile der verschiedenen Settings werden diskutiert.

3.1. IR zu institutionellem RDS

In einer privilegierten Situation für die Integration ihrer IR-Metadaten befindet sich eine Bibliothek, die ein eigenes RDS aufbaut und weiterentwickeln kann. Wenn nämlich die Rolle der Repositoriumsbetreiberin mit der Rolle der Katalogentwicklerin in eins fallen, lassen sich alle Aspekte der Datenintegration weitestgehend durch die Bibliothek selber bestimmen. Zudem kommen keine kommerzielle Interessen und Ansprüche ins Spiel und es müssen keine speziellen Anreize für eine Nachnutzung geschaffen werden. Denn grundsätzlich ist das Interesse, die Publikationen der eigenen Trägerinstitution im institutionellen RDS nachzuweisen gegeben. Die beiden UB Basel und Bern pflegen nicht nur eine gemeinsamen Katalogdatenbank in Aleph,

[9] Das Szenario „BORIS Integration in swissbib.ch" ist insofern ein Spezialfall, als die Kontrolle über die Metadaten das Haus IDS Basel Bern nicht verlässt und die diesbezügliche Kommunikation zwischen den Betreibern von Ziel- und Quellsystem optimal möglich ist.

sondern entwickeln auch ihre RDS-Instanz baselbern.swissbib.ch in Kooperation, womit sich also das Ziel-RDS des Metadaten-Harvesting von BORIS unter dem selben Dach befinden.[10]

Das Harvesting der Metadaten aus dem Repositorium wird über das Austauschprotokoll der Open Archives Initiative (OAI-PMH) abgewickelt. OAI-PMH – die erste offizielle Version datiert von 2001 – ist entwickelt worden, um die Interoperabilität zwischen elektronischen Dokumentenservern und sogenannten Service Providern (Aggregatoren und Suchmaschinen wie z.B. base-search.net) zu gewährleisten. Es ist im IR-Umfeld inzwischen der etablierte Standard, um Metadaten einfach und niederschwellig über HTTP auszutauschen.[11] Das Repositorium definiert selber, welche Inhalte geharvestet werden (beispielsweise nur die OA Volltexte), welche Repository-internen Felder in Dublin Core für das Harvesting ausgegeben werden, und – in Absprache mit swissbib.ch – wie diese ins Zielformat (MARC) transformiert werden sollen. Das Austauschprotokoll OAI-PMH erlaubt die Parametrisierung von verschiedenen Metadaten-Sets, in denen einmalig festgelegt wird, welche Typen von Datensätzen in zyklischer Abfrage fortan vom Zielsystem automatisch integriert werden sollen. Dabei macht es ein Datestamp in jedem Datensatz möglich, jeweils nur die seit dem letzten Harvesting veränderten, gelöschten oder neu hinzugekommenen Datensätze laden zu müssen, um einen kompletten Datenabgleich mit dem Data Provider (IR) zu erreichen.

Von traditioneller bibliothekarischer Warte gesehen sind allerdings solche Integrationsbestrebungen manchmal mit einem gewissen Unbehagen besetzt. Nicht selten besteht eine gewisse Angst vor „Katalogverwässerung" durch immer mehr Zeitschriftenartikel im „Katalog", der doch primär keine unselbständigen Publikationen beinhalten und die Benutzer nicht mit einem wirren Medien-Mix konfrontieren soll. Technisch wäre eine Beschränkung des IR-Harvestings auf die Monographien kein Problem, allerdings werden dann die Inhalte unter Umständen markant weniger, da Open Access Monographien in IR noch eher selten vertreten sind. Demgegenüber bieten aber einerseits heutige RDS-Suchoberflächen Filtermöglichkeiten, die den Medien-Mix mit einem Klick sortieren lassen. Zum anderen entstammt die „Katalogverwässerung" einer „traditionellen" bibliothekarischen Sicht. Eine kundenzentrierte Sichtweise wird diesen Aspekt gerade umgekehrt bewerten, ein One-Stop-Shop mit möglichst wenigen Brüchen in der Plattformnavigation erscheint mithin wünschenswert. Denn spätestens seit Google-Scholar kommen Bibliotheken vermehrt unter Druck, die Services von Zeitschriftendatenbanken in ihre RDS zu integrieren: „With the massive changes that led to the direct online delivery of articles, and

[10] Genau genommen beruhen swissbib.ch und baselbern.swissbib.ch auf zwei verschiedene Suchindices, diese werden aber beide an der UB Basel aus einer gemeinsamen Datenaufbereitungskomponente (Central Library System, OCLC) heraus aufgebaut. Ob Inhalte nur in swissbib.ch oder auch in baselbern.swissbib.ch erscheinen, ist also primär eine einfache Entscheidungsfrage. Die Option einer Integration im 1. Such-Tab von baselbern.swissbib.ch unterscheidet sich also nicht von der Integration in swissbib.ch, sie bietet die Sichtbarkeit im RDS der Heiminstitution UB Bern ohne zusätzlichen Aufwand und ist vom Workflow her gleich zu behandeln wie die Integration in swissbib.ch.

[11] Vgl. für eine Einführung zu OAI-PMH: Richard E. Jones, Theo Andrew, und John MacColl, The Institutional Repository (Oxford: Chandos, 2006), S. 67–71.

the high user demand for this content, it became critical for the catalogue to somehow integrate and surface article-level content."[12]

Unter dem Aspekt der Verbreitung der IR-Inhalte ist auf jeden Fall als Vorteil zu verbuchen, dass mit swissbib.ch eine mindestens schweizweite Sichtbarkeit in bibliothekarischen Findemitteln erreicht wird. Dabei geht die Datenhoheit aber nicht in fremde, kommerzielle Hände über, was grundsätzlich bedeutet, dass die Metadaten besser auf den Suchindex und die Suchoberfläche des RDS abgestimmt werden können, um dessen Facettierung möglichst gut zu unterstützen. Gleichzeitig bleibt auch der Suchalgorithmus flexibel bestimmbar, wodurch beispielsweise die Gewichtung der IR-Inhalte für das Ranking der Suchresultate nicht Dritten überlassen werden muss.

3.2. IR zu kommerziellem RDS

Für jede Grossbibliothek, die für den Nachweis und die Verwaltung ihres E-Medienangebots auf einen externen Anbieter zurückgreift, bietet es sich an, die Metadaten ihres Repositoriums in dessen Index zu integrieren. Für die grossen kommerziellen RDS-Anbieter wie Primo Central (ExLibris), EBSCO Discovery Service, WorldCat Discovery Services (ehemals WorldCat local), Summon (ProQuest), usw. ist es ein Leichtes, zusätzliche digitale Sammlungen in ihre Suchindices aufzunehmen, zumal Aggregation heterogener Quellen ihr Kerngeschäft darstellt. Ein Harvesting über OAI-PMH ist natürlich auch für kommerzielle Service Provider offen und leicht zu handhaben, womit die Anbindung des IR technisch kein Problem darstellt.

Allerdings ist die Grundlage für die Integration in einen kommerziellen Index eine Geschäftsbeziehung der IR-betreibenden Bibliothek mit dem Anbieter, die eine Lizenzierung des betreffenden Indexes voraussetzt. Schlussendlich liegt es in der Macht des Anbieters, die IR-Inhalte in den Index aufzunehmen und gegebenenfalls auch wieder zu löschen, die finanziellen Konditionen können auch von ihm bestimmt werden. Das Harvesting steht und fällt mit der Geschäftsbeziehung, bei einem Anbieterwechsel müssen die Abläufe für die Datenintegration wieder neu aufgesetzt werden.

Auf der Ebene der Metadaten muss mit einer Aufgabe der institutionellen Hoheit gerechnet werden. Die Abstimmung zwischen IR-Metadaten und Suchindex und dessen Facettierung kann schwierig sein, denn beispielsweise ist das Metadatenformat des Zielindexes nicht notwendig offen dokumentiert, wodurch die Kontrolle über die Feldertransformation durch die Bibliothek erschwert wird. Möglicherweise sind die Vokabularien, die für die Indexierung der Dokumente

[12] Karen Calhoun, „Supporting Digital Scholarship: Bibliographic Control, Library Cooperatives and Open Access Repositories", in Catalogue 2.0 : the future of the library catalogue (London: Facet Publishing, 2013), S. 147.

verwendet werden, nicht ausgewiesen. Auch muss bei der Integration in einen riesigen, aus heterogenen Quellen zusammengesetzten Datenbestand für die Erschliessung eine einheitliche Lösung gewählt werden, die nicht allen Teilbeständen gerecht wird. Mutmasslich hat die Dedublierung keine Priorität für Anbieter, die aus vermarktungstechnischen Gründen primär an einer grossen Zahl von Indexeinträgen in ihrem Angebot interessiert sind. Dadurch vermehren sich Eintragsdubletten im Index, deren Ranking in den Suchresultaten kaum durch die Bibliothek zu bestimmen sein wird.

Wie die in einen kommerziellen Discovery Service integrierten IR-Inhalte schlussendlich in konkreten Suchanfragen ausgegeben werden, hängt also nicht nur von der Metadaten-transformation, sondern auch vom eingesetzten Suchalgorithmus ab. Es ist anzunehmen, dass dieser bei kommerziellen Indices immer eine Black Box sein wird, auf deren Konfiguration die Bibliothek keine Einflussmöglichkeiten hat. Somit besteht die Gefahr, dass sich die Rangfolge innerhalb der Suchergebnisse zum Nachteil für die integrierten IR-Inhalte auswirken könnte. Die für den offenen Zugang zu freien Inhalten so wichtige Findbarkeit und Sichtbarkeit wären in diesem Fall eingeschränkt. Welcher Eintrag eines Artikels erscheint zuoberst in der Trefferliste, welcher wird auf der ersten Resultateseite, welcher erst auf der zweiten ausgegeben: jener für die Verlagsversion oder jener für die Version des Open Access Institutional Repository? Wie werden OA-Inhalte generell ausgegeben, wie werden sie gegenüber lizenzierten Inhalten in der Rangfolge dargestellt?[13]

Im konkreten Integrationssetting mit einem kommerziellen Index wird BORIS wöchentlich via OAI-PMH durch Summon geharvestet, wobei nur die Einträge mit einem OA Volltext übernommen werden. Obwohl die BORIS-Metadaten Angaben zu Sprache, Verlag, Verlagsort, Publikationstyp oder Peerreview-Status eines Dokuments enthalten, fallen diese momentan der Transformation in die Metadatenfelder des Summon-Indexes zum Opfer, wodurch die entsprechenden Filtermöglichkeiten nutzlos werden. Es scheint, dass ein eher starres Mapping-Schema zur Anwendung kommt, über das sich nicht so leicht eine Lösung für die spezifischen Anforderungen einer bestimmten Institution einrichten lässt. Im Falle der Sprachangabe zu einem Dokument führt aber nicht nur das Mapping an sich zu Problemen. Aufgrund oft mangelnder Spracherschliessung auf Artikelebene haben die Produzenten von Summon entschieden, die Sprache eines Artikels über die auf Zeitschriftenebene hinterlegte Sprachcodierung zu indexieren. Diese Vererbung führt allerdings für in mehrsprachigen Zeitschriften erschienene Artikel zu falschen Zuschreibungen, wodurch die Qualität der Sprachfacettierung in Summon so stark beeinträchtigt wird, dass sich die Entwickler von baselbern.swissbib.ch gegen ihre Anwendung

[13] Obwohl die UB Bern im aktuellen Setting keinerlei Einfluss auf den Suchalgorithmus ihres kommerziellen Partners hat, zeigen einige Stichproben allerdings, dass BORIS-Inhalte offenbar durch das Relevanz-Ranking nicht benachteiligt und sogar vor den betreffenden Verlags-Dubletten ausgegeben werden. Wie diese Inhalte aber in einer Installation desselben Indexes an einer anderen Bibliothek gerankt werden, ist eine weiterer Punkt, der sich der Kontrolle der UB Bern entzieht.

entschieden haben. So illustriert dieses Beispiel exemplarisch, wie die Aufgabe der Datenhoheit die Recherche auf die integrierten OA-IR-Inhalte – und schlussendlich auch den Zugang dazu – einschneidend beeinflussen kann.

Dennoch lassen sich durch die Integration von OA-IR-Inhalten in kommerzielle Discovery Services Vorteile ausmachen. Das primäre Ziel der Integration im eigenen RDS wird über die Dienstleistung des Anbieters erreicht; die Bibliothek muss sich – nach Etablierung eines den Qualitätsansprüchen genügenden Integrationsprozesses – nicht mehr um die regelmässige Datenintegration kümmern. Zudem eröffnen sich durch die Aufnahme in kommerzielle Indices durchaus weitere Kanäle zur Verbreitung von OA-Repositorieninhalten, kommen doch Dienste wie Summon global bei einer zahlreichen Kundschaft zum Einsatz. Generell erscheint dieses Integrationsszenario als gute Möglichkeit, OA-Inhalte auch in einem kommerziellen Umfeld einzubinden und sichtbar zu machen, obgleich diese Sichtbarkeit von den Lizenznehmern des jeweiligen Produkts selber erkauft werden muss und daher dem eigentlichen OA-Gedanken zuwiderläuft. Dass auf dem Markt der Resources Discovery Services nicht nach den Regeln von Open Access gespielt wird, erklärt sich von selbst – der Modus Operandi ist hier Closed Access. Daher überrascht es beispielsweise auch nicht, wenn OA-Repositorieninhalte bei der Integration in den Summon-Index nicht selbstverständlich von Anfang an so aufbereitet werden, dass sie auch bei Recherchen von ausserhalb der IP-Range der lizenznehmenden Institution angezeigt werden.

3.3. IR zum WWW: webbasierte Nachnutzung via Suchportale und -maschinen

Die unter 3.1. und 3.2. beschriebenen Szenarios beziehen sich primär auf die Nachnutzung der IR-Metadaten mittels Integration in ein RDS, das an derselben Institution wie das Repositorium beheimatet oder lizenziert ist. Darüber hinaus stehen aber auch Wege für die Nachnutzung offen, die eine noch weiterreichende Verbreitung und Sichtbarkeit ermöglichen. Wie in Kapitel 2 schon kurz erwähnt, unterstützen die gängigen Repositorien-Softwarelösungen (z.B. EPrints, DSpace) bereits standardmässig zwei wichtige webbasierte Distributionskanäle für die im Repositorium vorgehaltenen Metadaten. Sowohl über die OAI-PMH-Schnittstelle, wie auch über die HTML-Repräsentation eines OA-Repositoriums gelangen dessen Inhalte in nachnutzende Systeme mit viel grösserer Reichweite als sie bibliothekarische RDS bieten.

3.3.1. OAI-PMH

Als Paradebeispiel für den auf OAI-PMH aufsetzenden Verbreitungskanal kann die Bielefeld Academic Search Engine gelten, eine der grössten nichtkommerziellen, offenen, wissenschaftlichen Suchmaschinen. [14] Bei diesem Service Provider, der weit über 2000 OA Repositorien aus mehr als 70 Ländern harvestet, deren Metadaten normalisiert und aggregiert, ist von einer grossen Resonanz in der Wissenschaftsgemeinschaft auszugehen.[15] Denn nicht nur bietet BASE eine einfach zu bedienende Suchoberfläche mit gut ausgebauten Recherchefunktionalitäten, sie wird auch in zahlreichen Datenbanken, Katalogen, Metasuchmaschinen und Fachportalen eingebunden. Erfüllt ein OA Repositorium die Anforderungen, die sich aus der verlangten OAI-PMH-Kompatibilität ergeben, wird es nach intellektueller Prüfung als Quelle für BASE registriert und fortan wöchentlich geharvestet. Wie die Metadaten in BASE aufbereitet und bearbeitet werden, entzieht sich aber letztendlich der Kontrolle der Ursprungsinsitution, trotz aller Offenheit der Dienstleistung. Wie wird dedubliert? Welchem Ursprungs-Repositorium wird bei der Dedublierung der Einträge Vorrang gegeben? In welcher Rangfolge werden die Suchresultate gelistet? Werden dabei institutionelle Repositorien gegenüber fachlichen bevorzugt? Es ist zu erwarten, dass all diese Fragen grundsätzlich von den Betreibern von BASE geregelt werden und die Betreiber der Quellsysteme – aus nachvollziehbaren Gründen – kein Mitspracherecht haben. Denn BASE soll ja in erster Linie eine offene wissenschaftliche Suchmaschine sein, nicht ein Publicity-Instrument für die geharvesteten Institutionen.

Da aber OAI-PMH von Grund auf als Austausch- und nicht als Abfrageprotokoll konzipiert wurde, ist als Zwischenhändler immer ein Service Provider nötig, damit die IR-Inhalte über diesen Nachnutzungskanal aus dem Repositorium zum Endnutzer gelangen, denn über die OAI-PMH-Schnittstelle direkt lassen sich keine freien Suchen absetzen. Erst durch die Datenaufbereitung bei Service Providern wie BASE, lassen sich in der Folge spezifische inhaltliche Abfragen über die aggregierten Daten machen. OAI-PMH ist also eine speziell auf den Kontext von Repositorien abgestimmte Schnittstelle, die für eine direkte Einbindung von Repositorienmetadaten in die weitere WWW-Umgebung wenig Bedeutung hat. Dies wird auch durch die Tatsache bekräftigt, dass Google seit 2008 für die Indexierung OAI-PMH nicht mehr unterstützt, und nur noch das XML-Sitemaps-Protokoll sowie sogenannte Meta-Elemente im HTML Quelltext als Basis für das Crawling verwendet.[16]

---

[14] Vgl. „BASE Weblog: 10 Jahre BASE"; und „UB Wiki - Oeffentlich/BASE", zugegriffen 21. Februar 2015, http://www.ub.uni-bielefeld.de/wiki/BASE%20; Die Open Archives Initiative listet weitere bei ihr registrierte Service Provider, deren Dienst auf OAI-PMH beruhen: „Open Archives Initiative Service Providers", zugegriffen 21. Februar 2015, http://www.openarchives.org/service/listproviders.html.

[15] BASE ist nach Google der zweite Top Referrer für BORIS-Inhalte (vgl.: openaccess-unibe.atlassian.net/wiki/x/AgCuAg ; zugegriffen 21. Februar 2015).

[16] Vgl. Pascal-Nicolas Becker, „Repositorien und das Semantic Web – Repositorieninhalte als Linked Data bereitstellen", 2014, S. 17–20, http://www.pnjb.de/uni/diplomarbeit/repositorien_und_das_semantic_web.pdf.

OAI-PMH hat aber auch einen beträchtlichen Einfluss auf die beschreibenden Metadaten zu den Repositorieninhalten, da der standardmässige Einsatz dieses Austauschprotokolls eng mit einem ganz bestimmten Metadatenformat verbunden ist. Bis heute ist Simple Dublin Core die Minimalanforderung für das über OAI-PMH Version 2.0 übertragene Metadatenformat. Die Minimalanforderung schreibt Simple Dublin Core in seiner Reinform vor, die nur die 15 wenig granularen Basiselemente zur Erfassung von Metadaten zulässt. Dadurch fallen die sogenannten Qualifier weg, also alle über die 15 Grundelemente hinausgehenden Terms, die diese verfeinern können und eine granularere Strukturierung der Metadaten erlauben, obwohl technisch jedes in XML codierbare Format zur Übertragung via OAI-PMH tauglich ist. Diese von der Open Archives Initiative festgelegte Minimalanforderung an Repositorien führte aber dazu, dass dieses niedriggranulare Format zum De-facto-Standard geworden ist, der als einfachster gemeinsamer Nenner eine möglichst weitreichende Interoperabilität zwischen Repositorien gewährleisten soll.[17] Die Erfüllung dieses Standards wird beispielsweise auch von BASE erwartet. Sobald ein Repositorium über die Basis von Simple DC hinausgehende Elemente verwendet, gibt der BASE OAI-Validator eine Fehlermeldung für die XML-Validierung zurück.

Dass die Verwendung eines wenig granularen Metadatenformats bei der Integration und Nachnutzung zu Nachteilen führen kann, zeigt sich beispielsweise an den Auswirkungen, die eine ausschliessliche Verwendung von Simple DC auf die Inhaltserschliessung und in der Folge auf die Recherchemöglichkeiten hat. Das für die Erfassung des Themas eines Dokuments verwendete DC-Element „subject" kann mit Stichworten, Phrasen und Klassifizierungscodes jeglicher Art gefüllt werden, wobei kontrollierte Vokabulare empfohlen werden.[18] Allerdings lässt sich mit diesem Element allein keine Aussage über das darin im konkreten Fall tatsächlich verwendete Vokabular machen. Dies führt insbesondere bei der maschinellen Verarbeitung von DC-Datensätzen, wie sie bei OAI Service Providern zur Aggregation von enormen Datenmengen nötig sind, zu erheblichen Schwierigkeiten. Im Hinblick auf eine einheitliche Browsing- oder Facettierungsfunktion in der Suchoberfläche werden die in den subject-Feldern gelieferten Daten nutzlos, da sie sich nur mit grossem Aufwand automatisch einem kontrollierten Vokabular zuschreiben lassen.[19] Bei BASE wird u.a. aus diesem Grund für das klassifikatorisch unterstützte Browsing nicht auf die mitgelieferten subject-Felder zurückgegriffen.

Um dennoch ein Browsing nach DDC anbieten zu können, haben die Betreiber von BASE eine Anwendung zur automatisierten Klassifikation von Datensätzen entwickelt. In einem texttechnologischen Verfahren werden die aggregierten Metadaten, sofern sie genügend

[17] Vgl. „Open Archives Initiative - Protocol for Metadata Harvesting - v.2.0", zugegriffen 22. Februar 2015, http://www.openarchives.org/OAI/openarchivesprotocol.html#Record.

[18] Vgl. „DCMI Metadata Terms", zugegriffen 22. Februar 2015, http://dublincore.org/documents/2012/06/14/dcmi-terms/?v=elements#subject.

[19] Vgl. Timo Borst, „Repositorien auf ihrem Weg in das Semantic Web: Semantisch hergeleitete Interoperabilität als Zielstellung für künftige Repository-Entwicklungen", Bibliothek Forschung und Praxis 38, Nr. 2 (2014): S. 259, doi:10.1515/bfp-2014-0034.

umfangreiche Abstracts in deutscher oder englischer Sprache enthalten, mit einer automatisch berechneten DDC-Nummer und entsprechender Klassenbezeichnung versehen.[20] Wie Stichproben zeigen, wird dieses Verfahren selbst auf Datensätze mit schon vorhandener DDC-Klassifizierung angewendet, da sich diese eben nicht ohne weiteres maschinell auswerten und für das BASE-Browsing nachnutzen lassen. Zudem würde sich der dazu nötige zusätzliche Aufwand auch in Anbetracht der sehr heterogenen, nur teilweise mit DDC inhaltlich erschlossenen Datenbasis nicht lohnen.

Ein möglicher Ausweg aus diesem verlustbehafteten, standardmässigen Austauschverfahren über OAI-PMH wäre sicherlich die Anhebung der Minimalanforderung an das Metadatenschema durch die Open Archives Initiative auf ein um die nötigen Elemente des aktuellen Sets der Dublin Core Metadata Initiative (DCMI) Metadata Terms. Denn selbst im Rahmen von Dublin Core wäre es schon möglich – um im Anwendungsfeld der inhaltlichen Erschliessung zu bleiben – mittels der Qualifiers, die im Vocabulary Encoding Scheme der DCMI Metadata Terms zur Verfügung stehen, zumindest die weitverbreiteten kontrollierten Vokabularien DDC, UDK, LCC, LCSH oder MeSH, maschinenlesbar zu identifizieren.[21] Wie schnell eine solche Anhebung des Standard Metadatenformats umgesetzt und wie gut sie in der OA-IR-Community akzeptiert würde, bliebe allerdings abzuwarten. Auch würde sowohl die Niederschwelligkeit des Protokolls, als auch das bestehende OAI-PMH-Ökosystem von Data Providern und Service Providern möglicherweise einem Test unterzogen.

Als zweiter möglicher Ausweg aus dem gegenwärtig verlustbehafteten Datentransfer über OAI-PMH wäre denkbar, die Identifikation der verwendeten Vokabularien nicht über die DC-Elemente, sondern über die darin erfassten Werte selber zu erreichen. Indem nicht nur die menschenlesbaren Zeichenketten, sondern auch maschinenlesbare HTTP-URI für die Beschlagwortung in die subject-Felder eines DC-Datensatzes eingefügt würden, wäre das Vokabular über den Namespace des URI eindeutig identifizierbar und für die maschinelle Verarbeitung gerüstet.[22] Ob die technische Umsetzbarkeit mit den heute im OAI-PMH-Kontext gängigen Softwarelösungen gegeben ist, müsste im Einzelnen geklärt werden. Kann beispielsweise die IR-Software Eprints HTTP-URI so vorhalten, dass sie zwar in der Präsentationschicht als menschenlesbare Strings erscheinen, für das Harvesting aber als

[20] Vgl. Mathias Lösch, „Automatische Klassifikation von OAI-Metadaten mit linguistischen Methoden. Vortrag im Kolloquium Wissensinfrastruktur an der UB Bielefeld" (Bielefeld, 30. Oktober 2009).

[21] Vgl. „DCMI Metadata Terms", zugegriffen 22. Februar 2015, http://dublincore.org/documents/2012/06/14/dcmi-terms/?v=dcam#H4.

[22] Zwar ist dieser auf Verlinkung basierende Ansatz nahe an der Idee von Linked Open Data, indem dabei auf bestehende Ressourcen, auf anderswo definierte Entitäten verwiesen wird, dennoch wären damit die Linked Data Prinzipien nach Tim Berners-Lee nur teilweise verwirklicht. Vgl. dazu Kapitel 5.2. weiter unten.

dereferenzierbare HTTP-URI ausgegeben werden?[23] Können die auf Seiten der Service Provider heute gängigen Softwarelösungen die gelieferten URI auch tatsächlich zielführend verarbeiten? Falls sich die Speicherung von HTTP-URI in IR-Metadaten aufgrund von Granularitätsproblemen im Zusammenhang mit DC, oder aufgrund technischer Einschränkungen innerhalb der Repository-Software tatsächlich nicht bewerkstelligen lässt, muss der Weg über ein genügend granulares und HTTP-URI-freundliches Internformat (das dann für die OAI-konforme Ausgabe auf ein einfaches DC gemappt werden müsste) und eine geeignete Softwarealternative gegangen werden.

3.3.2. Indexierung durch Websuchmaschinen

Aufgrund der Popularität von Websuchmaschinen wie Google, Bing, Yahoo etc. sind deren Indices sicherlich die Instrumente zur webbasierten Verbreitung von IR-Inhalten mit der grössten Reichweite und sollten unbedingt in die Strategie zur Verbreitung von IR-Inhalten miteinbezogen werden.[24] Die Indexierung von Repositorienmetadaten und –volltexten durch Websuchmaschinen geschieht mit Hilfe von Webcrawlern im Zuge der regelmässigen automatische Durchforstung und Analyse von Webseiten im gesamten WWW, ohne dass dazu ein Webserver selbst aktiv werden muss. Die am häufigsten in OA-Repositorien angewendeten Softwarelösungen (DSpace, Eprints) sind standardmässig so konfiguriert, dass diese durch Websuchmaschinen problemlos gecrawlt werden können. Dabei folgen Webcrawler oder auch Robots genannte Programme wie Googles Googlebot den in den HTML-Seiten einer Website codierten Hyperlinks und indexieren dabei sukzessive die auf den gefundenen HTML-Seiten vorgehaltenen Textdaten. Dazu zählen im Falle von Repositorien neben den bibliographischen Metadaten auch die Abstracts und die verlinkten Volltexte, sofern deren PDF-Dateien offen und suchbar sind. Grundsätzlich gilt daher für das integrale Crawling eines Repositoriums, dass jeder Eintrag über einen Hyperlink erreichbar sein muss. Stellt ein IR also eine Browsingfunktion zur Verfügung, die nur mit Klicks zu jedem Eintrag navigieren lässt, ist diese Bedingung erfüllt. So gelangt ein Webcrawler – analog zu menschlichen Webnutzenden – beispielsweise über den Link „Browsen nach Jahr" auf Listen von Titeln, die wiederum mit Links hinterlegt sind, die zu den einzelnen zu indexierenden Einträgen führen. Sobald aber Inhalte – wie in einem OPAC – nur über ein Suchinterface aufrufbar sind, haben Webcrawler keinen Zugriff darauf. Da Webcrawler je

[23] Becker bezweifelt für Eprints diese Funktion. Vgl. Becker, „Repositorien und das Semantic Web – Repositorieninhalte als Linked Data bereitstellen", S. 43.

[24] Die Dominanz von Websuchmaschinen auch im akademischen Umfeld ist seit längerem verschiedentlich belegt. Für eine zusammenfassende Übersicht vgl. beispielsweise Kenning Arlitsch und Patrick S. O'Brien, „Invisible institutional repositories", *Library Hi Tech* 30, Nr. 1 (2. März 2012): S. 62–63, doi:10.1108/07378831211213210.

nach Algorithmus nur eine beschränkte Zeit auf einem bestimmten Webserver verweilen, oder nur bis zu einer bestimmten Tiefe in der Seitenstruktur vordringen, sollte die Hierarchie der IR-Website für eine gründliche Indexierung zudem möglichst flach gehalten werden.[25] Neben zahlreichen weiteren Optimierungsmöglichkeiten, die OA-IR-Inhalte durch Suchmaschinen gut indexieren helfen, können sich Repositorien (wie jegliche Website-Betreiber) des Sitemap-Protokolls bedienen, um beispielsweise Google beim Crawling der eigenen Seiten zu unterstützen. Somit können Suchmaschinen aktiv darüber informiert werden, welche Seiten auf einem Webserver existieren, wann sie geändert wurden etc., und so dem Webcrawler den Weg für ein effizientes Indexieren weisen.[26]

Sind also die genannten Bedingungen für das Crawling erfüllt, kann sich ein Repositorium die enorme Reichweite und Sichtbarkeit zunutze machen, ohne dass dazu zusätzliche Leistungen erbracht werden müssen. Da die Websuchdienstanbieter ein eigenes Interesse haben, möglichst das gesamte WWW zu erschliessen, werden für die Indexierung auch keine Lizenzgebühren fällig. Dem steht allerdings auch die weitgehende Abhängigkeit von der „Gutmütigkeit" der hinter dem Webcrawlern und Suchmaschinen stehenden Algorithmen gegenüber.

Für die Recherche in mittels Crawling automatisch erstellten Indices kann ein Nachteil allerdings sein, dass die Suchabfragen nicht durch Metadatenfeldabfrage unterstützt wird, sondern auf Volltextindexierung basiert. Somit bieten Suchmaschinen wie Google Scholar keine spezifische Suche mit kontrollierten Vokabularien, aber auch die in der erweiterten Suche angebotenen Autorensuche kann problembehaftet sein. Da die Indexierung von Autorennamen offenbar auf texttechnologischer Nachbearbeitung der Volltextindexate basiert, Named Entity Recognition aber schwierig fehlerfrei zu gestalten ist, ergeben sich teilweise absurde Autorenzuschreibungen, wie Jacsó verschiedentlich dokumentiert hat und hier erläutert:

> Instead the software of Google Scholar has produced literally millions of ghost authors from section headings, descriptor terms, side-bar headers, or other prominently displayed terms in the source documents which they assumed to be the authors. The parsers also created publication years from page numbers, street addresses, and postal code numbers as long as they were four digits long. There have been millions of lost authors who were deprived of their authorship and millions of missing citations.[27]

[25] Die hier dargestellten Abläufe des Crawlings, sowie Best Practices für Repositorien dokumentiert das Repositories Support Project: „Optimisation ~ Grow ~ Repositories Support Project", zugegriffen 23. Februar 2015, http://www.rsp.ac.uk/grow/optimisation/. Weitere informative Dokumentation bezüglich Crawling von Repositorien bieten Peter Suber, „How to facilitate Google crawling of OA repositories", zugegriffen 23. Februar 2015, http://legacy.earlham.edu/~peters/fos/googlecrawling.htm; und Jody L. DeRidder, „Googlizing a Digital Library", The Code4Lib Journal, Nr. 2 (24. März 2008), http://journal.code4lib.org/articles/43.

[26] Vgl. „sitemaps.org", zugegriffen 23. Februar 2015, http://www.sitemaps.org/de/.

[27] Péter Jacsó, „Google Scholar Author Citation Tracker: is it too little, too late?", Online Information Review 36, Nr. 1 (17. Februar 2012): S. 129, doi:10.1108/14684521211209581.

Seither sind diese Fehler in Google Scholar kontinuierlich minimiert worden, ob mit besseren texttechnologischen Algorithmen oder bloss Listen von Stopwörtern ist unklar. Ohne die Möglichkeit auf eindeutige Identifier für Autorschaften zurückgreifen zu können, bleibt das Problem aber wahrscheinlich behelfsmässig gelöst. Immerhin besteht mittlerweile die Möglichkeit, mittels HTML-„meta"-Tags die bibliographischen Angaben auf einer HTML-Repository-Seite als solche zu kennzeichnen, wodurch die Indexierung viel besser gelenkt werden kann. Google Scholar unterstützt beispielsweise prinzipiell den Gebrauch solcher Tags und gibt Hinweise, wie diese konfiguriert sein müssen und welche Vokabularien dabei verwendet werden können, damit sie der Crawler bestmöglich interpretieren kann.[28] Repositorien-Software wie Eprints oder DSpace lassen sich so konfigurieren, dass sie diese Tags im HTML-Header der einzelnen Repositorieneinträge automatisch integrieren und so Titel, Autorennamen und Erscheinungsdaten als solche in einer Form gekennzeichnet werden, die Googles Webcrawler gut indexieren können.

Wieweit und in welcher Weise diese Massnahmen aber tatsächlich die Indexierung und in der Folge die Suchresultate in gewünschtem Masse unterstützen, bleibt hingegen einerseits Betriebsgeheimnis der Suchmaschinenanbieter, hängt aber andererseits von weiteren Faktoren ab. Beispielsweise ist schwer abzuschätzen, warum BORIS-Einträge bei einer Suche mit dem „site"-Operator („site:boris.unibe.ch") nur zu einem Bruchteil in Google Scholar auftauchen, obwohl bei BORIS eine von Google Scholar für ihre Indexierung als bestens geeignet empfohlene Software zum Einsatz kommt. Arlitsch und O'Brien beschreiben in einer 2011 durchgeführten Studie eine generell niedrige Integration von IR-Inhalten in den Index von Google Scholar, unabhängig von der verwendeten Software.[29] Als Hauptgrund nennen sie eine mangelnde Darstellbarkeit von bibliographischen Angaben zu Zeitschriftenliteratur in Dublin Core und schlagen eine Metadatenkonversion zu einem von Google Scholar empfohlenen Schema vor. Im Testfall wurde das Schema Highwire Press mit grossem Erfolg eingesetzt. Gleichzeitig stellen sie aber auch klar, dass eine Anzahl weiterer Faktoren die Indexierung in Google Scholar beeinflusst. Darunter dürfte der wichtigste sein, dass Google Scholar unter verschiedenen Versionen eines Dokuments die Verlagsversion als Primärversion festgelegt, während Sekundärversionen daran angehängt werden. Aufgrund dieser Form der Dedublierung verschwinden IR-Einträge in der Resultateliste in Google Scholar hinter einem der Primärversion beigefügten Link und werden auch mit einer „site"-Operatoren-Suche nicht gefunden, obwohl sie von Google Scholar indexiert worden sind. Somit wird also der Ertrag der Bemühungen um Kompatibilität mit Google Scholars Metadatentags-Empfehlungen seitens IR durch die Dedublierungspolitik vermindert. Dennoch lohnen sich diese Bemühungen, denn sie verbessern nicht nur die Indexierung durch Google

[28] Vgl. „Inclusion Guidelines for Webmasters: Indexing Guidelines", zugegriffen 24. Februar 2015, http://scholar.google.com/intl/en/scholar/inclusion.html#indexing.

[29] Vgl. Arlitsch und O'Brien, „Invisible institutional repositories", S. 72.

Scholar, sondern gleichermassen auch die Indexierung durch weitere Websuchmaschinen, Googles Hauptindex eingeschlossen.[30]

Einerseits bieten das Crawling und Indexieren von IR-Inhalten durch Suchmaschinen also ein enormes Potential an Sichtbarkeit und Reichweite, andererseits haben diese abgesehen von u.U. sehr aufwändigen Suchmaschinenoptimierung des eigenen Webservers nur sehr beschränkt Kontrolle über die Art und Weise der Verbreitung ihrer Metadaten, wie das Beispiel Google Scholar zeigt. Im Bereich der Nachnutzung durch die Indices kommerzieller Websuchmaschinen besteht die Herausforderung für IR-Betreiber somit darin, Wege zu finden, wie dieser Kontrollverlust wettgemacht werden kann. Search Engine Optimization (SEO; Suchmaschinen-optimierung) über die Einbindung von interoperablen Schemata in den HTML-Header ist einer dieser Wege – die optimale Vorbereitung für die Crawlbots als Vorbedingung miteingeschlossen.

Als weiterer vielversprechender Weg dürfte der Einsatz von semantischer SEO interessant sein. Zwar hat sich die Begeisterung für das Semantic Web seit seiner anfänglichen Propagierung durch Tim Berners-Lee 2001 zwischenzeitlich wieder gelegt,[31] aber das Beispiel Google lässt in der Websuchlandschaft einen erneuerten Trend in dieser Richtung ausmachen. Im August 2013 hat Google einen komplett neuen Suchalgorithmus in Betrieb genommen, der deutlich darauf ausgerichtet ist, nicht mehr nur Zeichenfolgen in Dokumenten wiederzufinden, sondern Konzepte oder Entitäten, wie er auch Beziehungen zwischen diesen darstellen kann. Das Schlagwort „Things, not strings", das im 2012 beim Launch von Google Knowledge Graph in den Umlauf gebracht wurde, bezeichnet diesen sich abzeichnenden Paradigmenwechsel prägnant.[32] Angesichts dieser Entwicklung erscheint eine Ausrichtung des SEO auf die semantische Auszeichnung von IR-Inhalten als zukunftsgerichtete Strategie der Metadatenstrukturierung im Hinblick auf deren Nachnutzung im WWW.

[30] Für Google erreichten Arlitsch und O'Brien mit eine Testkollektion eine Abdeckung von beinahe 100 % (10'306/10'536) gegenüber 18 % vor der Optimierung. Vgl. Ebd., S. 75.

[31] Vgl. Adrian Pohl und Patrick Danowski, „Linked Open Data in der Bibliothekswelt: Grundlagen und Überblick", in (Open) Linked Data in Bibliotheken (Berlin, Boston: De Gruyter, 2013), S. 5, http://www.degruyter.com/view/books/9783110278736/9783110278736.1/9783110278736.1.xml.

[32] Vgl. den offiziellen Blogeintrag „Introducing the Knowledge Graph: things, not strings", Official Google Blog, zugegriffen 25. Februar 2015, http://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html.

3.3.3. Semantische Auszeichnung von IR-Einträgen mit schema.org

Die vier grossen Suchdienstanbieter Bing, Google, Yahoo und Yandex lancierten 2011 gemeinsam den Standard schema.org,[33] der es ihren Suchmaschinen erlaubt, HTML-Seiten durch semantische Auszeichnung noch besser indexierbar und suchbar zu machen. Schema.org stellt Webmastern eine Ontologie zur Verfügung, mit deren Terminologie sie die Inhalte von Webseiten so beschreiben können, dass deren Bedeutung maschinenlesbar und automatisch verarbeitbar wird. In schema.org gemachte Aussagen können – mit microdata, microformats oder RDFa codiert[34] – direkt im HTML-Code eingebettet werden. Schema.org wurde als ein für jegliche Art von Inhalten anwendbares Vokabular angelegt und ist nach Themen in verschiedene Schemas gegliedert, die wiederum Types und Properties als Klassen und Eigenschaften aufweisen. Zwar ist schema.org nicht auf die Bibliotheksdomäne ausgelegt, ist aber gerade dadurch für die weitere Webcommunity anschlussfähig und wird von einem viel grösseren Kreis verstanden als ein spezifischeres Vokabular wie DC.

Für die bibliographische Beschreibung kommt v.a. das Schema „CreativeWork" mit Types wie „Book", „Article", „WebPage" etc. in Betracht, auch wenn sich damit nicht das ganze bibliographische Universum abbilden lässt. Da schema.org aber erweiterungsfähig konzipiert ist, lassen sich Sets von noch fehlenden Types und Properties gut ergänzen und beispielsweise auch die nötige Granularität für die Beschreibung von Artikelpublikationen erreichen. Eine solche Ergänzung wird denn auch von der W3C-Interessengruppe Schema Bib Extend vorangetrieben, und sukzessive durch schema.org offiziell unterstützt.[35] Eine speziell auf die Beschreibung von IR-Inhalten ausgerichtete Erweiterung wird an der Montana University Library entwickelt und ist für Dissertationen und studentischen Arbeiten bereits erfolgreich getestet worden.[36]

Während bibliographische Feldbezeichnungen auf einer herkömmlichen HTML-Seite durch Menschen gut interpretiert werden können, ist eine eindeutige Interpretation derselben für Maschinen hingegen kaum möglich. Hier kommen die Vorteile der Verwendung einer Ontologie wie schema.org zum Tragen. Denn eine zentrale Funktion von semantischer Auszeichnung liegt darin, dass sie Beziehungen (und deren Bedeutung) zwischen Entitäten eindeutig und maschinenlesbar darstellen kann. So können beispielsweise die Rollen von mit einem Dokument

[33] http://schema.org; Für weitere Erläuterungen vgl. Carsten Klee, „Vokabulare für bibliographische Daten: Zwischen Dublin Core und bibliothekarischem Anspruch", in (Open) Linked Data in Bibliotheken (Berlin, Boston: De Gruyter, 2013), S. 53, http://www.degruyter.com/view/books/9783110278736/9783110278736.45/9783110278736.45.xml; sowie „Schema.org", Wikipedia, the Free Encyclopedia, 17. Februar 2015, http://en.wikipedia.org/wiki/Schema.org.

[34] Google unterstützte anfänglich alle drei Formate, hat aber nun in seiner Empfehlung microdata monopolisiert. Vgl. „schema.org FAQ - Webmaster Tools Help", zugegriffen 26. Februar 2015, https://support.google.com/webmasters/answer/1211158.

[35] „Schema Bib Extend Community Group", zugegriffen 25. Februar 2015, https://www.w3.org/community/schemabibex/wiki/Main_Page.

[36] Vgl. Jeff Keith Mixter, Patrick S. O'Brien, und Kenning Arlitsch, „Describing Theses and Dissertations Using Schema.org", International Conference on Dublin Core and Metadata Applications, 8. Oktober 2014, 138–46.

assoziierten Personen durch Rückgriff auf eine Ontologie eindeutig und maschinenlesbar beschrieben werden. Referent und Autor einer Dissertation lassen sich durch die Vergabe von Properties so kennzeichnen, dass eine Suchmaschine diesen Rollenunterschied auswerten kann. Auch Affiliationen von Personen mit Organisationen, also z.B. einer Forscherin mit einer Universität, lassen sich mit einer Property auf diese Weise ausweisen. Generell lassen sich also durch den Einsatz einer Ontologie Beziehungen zwischen einzelnen Feldinhalten eindeutig und maschinenlesbar darstellen, während diese Beziehungen ohne den Einsatz einer Ontologie implizit bleiben und nur dadurch bestehen, dass sich die Felder gemeinsam auf derselben HTML-Seite befinden, wodurch sie durch eine Maschine aber nicht eindeutig interpretiert werden können.

Die Auszeichnung von HTTP-URIs mit schema.org führt dazu, dass Suchmaschinen nicht nur verstehen, dass sie einen URI vor sich haben, sondern auch dessen Funktion in Bezug auf das beschriebene Dokument erfassen. Wenn also beispielsweise eine URI für einen bestimmten GND-Datensatz mit einem Property „about" eindeutig in eine inhaltsbeschreibende Beziehung zum Dokument gesetzt wird, wird für die Maschine klar, dass dieser GND-URI eine Aussage über den Inhalt des Dokuments macht, und könnte somit in einer Resultateliste dieses Schlagwort gezielt als Inhaltsangabe zum Dokument darstellen. Dadurch würden IR-Inhalte in Suchresultaten mit reichhaltigeren Kontextinformationen darstellbar, was die Recherche entscheidend unterstützen könnte. Eine ähnliche Anwendung kommt bei Google mit Rich Snippets schon zum Einsatz. Dabei werden aufgrund von schema.org-Auszeichnungen Suchresultate für Personen, Veranstaltungen, Produkte, Musik, Besprechungen oder Kochrezepte mit zusätzlichen Informationen angereichert, die helfen sollen, die Relevanz einer gefundenen Website einzuschätzen. Zurzeit unterstützt diese Funktion nur eine geringe Anzahl von Klassen (Types) aus schema.org, weitere könnten aber laut Google folgen.[37]

Prinzipiell können in schema.org nicht nur HTTP-URI ausgezeichnet werden. Es ist aber auch für die Darstellung von IR-Inhalten in schema.org hilfreich und wichtig, möglichst maschinenverarbeitbare Identifikatoren und nicht blosse Zeichenfolgen zu verwenden, um ihre Attraktivität und Kompatibilität für die Nachnutzung in solchen Applikationen zu steigern. Der Slogan „Things not strings" mag hier die Bedeutung von Identifikatoren für die neue Generation der (semantischen) Websuche nochmals in Erinnerung rufen.

Konkrete Anwendungen von schema.org in der HTML-Schicht von institutionellen Repositorien sind vereinzelt schon anzutreffen. Ein Beispiel ist das IR der Columbia State University, das ganz

[37] „ Not every type of information in schema.org will be surfaced in search results but over time you can expect that more data will be used in more ways. [...] Google currently supports rich snippets for people, events, reviews, products, recipes, and breadcrumb navigation, and you can use the new schema.org markup for these types, [...]. Because we're always working to expand our functionality and improve the relevance and presentation of our search results, schema.org contains many new types that Google may use in future applications. In addition, since the markup is publicly accessible from your web pages, other organizations may find interesting new ways to make use of it as well." Vgl. „schema.org FAQ - Webmaster Tools Help".

auf meta-Tags und schema.org für die Einbindung in Indices von Websuchmaschinen setzt, und mit der Aufgabe ihrer OAI-PMH-Schnittstelle auf die Verbreitung über diesen Kanal verzichtet.[38] Interoperabilität und Sichtbarkeit für Metadaten sollen in diesem IR immer mehr auf Technologien des Semantic Web abgestützt werden, unter anderem mit der Begründung, „[...] OAI-PMH is not the most efficient way to increase discoverability of repository contents through the major search engines."[39] An der Montana State University wurde in einem Pilotprojekt mit einem IR-Teilbestand von Dissertationen und studentischen Arbeiten ein spezifisches, auf schema.org beruhendes Vokabular entwickelt und erfolgreich getestet. Aufgrund dieser Konzeptstudie sollen weitere Bestände mit schema.org ausgezeichnet, und das entwickelte Vokabular in weiterer existierende IR integriert werden, wie auch die Auswirkungen auf die Nachnutzung in Webservices genauer untersucht werden.[40] Dahinter steht die Überzeugung, dass mit mehr Bedeutung und Kontext angereicherte IR-Inhalte bessere Suchresultate für Websuchmaschinennutzer ermöglichen.

Zur Anreicherung oder Konvertierung von existierenden IR-Metadaten mit schema.org scheint es keine Standardapplikation zu geben. Mit Islandora, Blacklight (bei AC Columbia im Einsatz), ScholarSphere oder der am CERN entwickelten Lösung Invenio gibt es aber einige Beispiele von IR-Management-Software, die schema.org standardmässig in ihre Präsentationsschicht integrieren können. Auch DSpace lässt sich so abändern, dass schema.org mit JSON-LD in die HTML-Darstellung integriert werden kann, wie dies das IR Scholar Works in Montana für ihr Testset vollzogen hat.[41]

Die Effizienz von semantischer Auszeichnung im Hinblick auf Suchmaschinenoptimierung ist zurzeit noch eine nicht empirisch getestete Hypothese, aber erste Hinweise auf erhöhte Zugriffszahlen und bessere Rankings für IR-Inhalte gibt es.[42] Wie weit das Versprechen einer besseren Auffindbarkeit, einer besseren Integration in Applikationen wie Rich Snippets und letztendlich besser auf die Endnutzenden abgestimmte Suchresultate durch semantische Suchmaschinenoptimierung mit schema.org eingelöst werden kann, bliebe also noch abzuwarten. Ihr Potential ist durch die Unterstützung der weltgrössten Websuchdienstanbieter aber vielversprechend und die verschiedentlich geäusserten grossen Hoffnungen auf ihren

[38] „Academic Commons", zugegriffen 26. Februar 2015, http://academiccommons.columbia.edu/.

[39] Vgl. Robert J. Hilliker, Melanie Wacker, und Amy L. Nurnberger, „Improving Discovery of and Access to Digital Repository Contents Using Semantic Web Standards: Columbia University's Academic Commons", Journal of Library Metadata 13, Nr. 2–3 (Juli 2013): S. 2, doi:10.1080/19386389.2013.826036.

[40] Vgl. Mixter, O'Brien, und Arlitsch, „Describing Theses and Dissertations Using Schema.org".

[41] „Islandora Website", zugegriffen 27. Februar 2015, http://islandora.ca/; „Home – Blacklight", zugegriffen 27. Februar 2015, http://projectblacklight.org/; Eine Eigenentwicklung der Penn State University: „ScholarSphere", zugegriffen 27. Februar 2015, https://scholarsphere.psu.edu/; „Invenio", zugegriffen 27. Februar 2015, http://invenio-software.org/.

[42] Vgl. z.B. Hilliker, Wacker, und Nurnberger, „Improving Discovery of and Access to Digital Repository Contents Using Semantic Web Standards", S. 4.

Nutzen für die Verbreitung von IR-Inhalten scheinen berechtigt. Denn wenn IR-Inhalte so im WWW präsentiert werden, dass Suchmaschinen ihre Metadaten gut verstehen können, ist die Wahrscheinlichkeit für eine gründliche Indexierung und ein optimales Ranking in Suchresultaten am grössten. Wenn dabei zudem existierende Identifier-Sets (für Personen, Körperschaften, Schlagwortdateien) nachgenutzt werden, wird die Interoperabilität mutmasslich noch zusätzlich erhöht. Denn es ist davon auszugehen – so die Grundthese dieser Arbeit – dass granular strukturierte Metadaten grundsätzlich umso attraktiver für eine Nachnutzung in anderen Discovery Systemen werden, je reichhaltiger sie insbesondere an normierten Erschliessungsmetadaten sind. Da die Einbindung in neue semantische Webapplikationen in der Art eines Google Knowledge Graph auf maschinelle Verarbeitung von Identifikatoren abstützt, wird insbesondere auch die Vorhaltung dieser Identifikatoren als eindeutig dereferenzierbare HTTP-URI gefordert sein.[43]

4. Anreicherung

Für viele institutionelle Repositorien ist die Ausgangssituation für die Nachnutzung der IR-Inhalte durch oft wenig reichhaltige Metadaten geprägt, besonders was die inhaltliche Erschliessung mit kontrollierten Vokabularien betrifft. So weist in der Schweizer IR-Landschaft beispielsweise edoc.unibas.ch gar keine Inhaltserschliessung auf, BORIS oder ZORA greifen auf eine grobe DDC-Klassifikation aufgrund von Institutszugehörigkeiten oder unkontrollierte Schlagworte und Volltextsuche ohne ausgefeilte Filtermöglichkeiten zurück. Auch die Erschliessung der beteiligten Personen und Körperschaften wird nicht über Normdaten zur Identifikation von AutorInnen abgewickelt, wodurch mehrfache Ansetzungen keine Seltenheit sind. Bei BORIS reicht die Kapazität für eine spezielle Pflege der Berner Autorinnen und Autoren, aber diese eigenen Ansetzungen kommen nur BORIS-intern zur Anwendung und sind nach aussen nicht verknüpft mit eindeutigen Identifikatoren. Zum einen ist diese Situation begründet durch begrenzt zur Verfügung stehende Ressourcen, zum anderen durch eine Minimalanforderung an Metadaten der im OAI-PMH-Umfeld de facto zum Standard geworden ist und im Sinne einer möglichst niederschwelligen Interoperabilität keine bestimmten Vokabularien zur Erschliessung vorschreibt. Für alle beschriebenen Nachnutzungsszenarien sind reichhaltige Metadaten aber ein grosser Vorteil, wodurch sich Anreicherung von IR-Metadaten als ein wichtiges Ziel im Hinblick auf eine möglichst weite Verbreitung und Nachnutzung der IR-Inhalte begründen lässt.

[43] Dies ist denn auch eine der zentralen durch die Linked-Data-Prinzipien gestellten Forderungen, vgl. Kapitel 5.2. weiter unten.

Unter Anreicherung soll hier die Erweiterung von IR-Einträgen in einem weiten Sinn verstanden werden, fokussiert auf die Erweiterung der Metadaten, wobei darunter auch bei der Erfassung zum Einsatz kommende Hilfsmittel oder Datenimport fallen können. Hingegen soll hier Anreicherung mit zusätzlichen digitalen Objekten wie Coverbildern, Inhaltsverzeichnissen oder Volltexten, wie sie im Bibliothekskatalog oft vorgenommen wird kein Thema sein, da digitale Objekte als per definitionem schon vorhandene integrale Bestandteile eines IR gelten können.

4.1. Anreicherungsarten

Um sich eine genauere Vorstellung darüber zu verschaffen, auf welche Arten Metadaten in einem IR mit Fremddaten angereichert werden können, kann die konzeptuelle Unterscheidung in drei grundsätzliche Formen wie sie Christoph für die Kataloganreicherung beschreibt hilfreich sein.[44] Seine Unterscheidung basiert einerseits auf den involvierten technischen Prozessen, andererseits – in enger Verbindung mit diesen – aber auch auf lizenzrechtlichen Bedingungen. So lassen sich die blosse **Verlinkung**, die **dynamische Anreicherung** (d.h. die Übernahme von ad-hoc und automatisch über Links abgerufene Fremddaten), sowie die **komplette Datenübernahme** in die eigene Datenbank voneinander unterscheiden. Bei der technisch wenig aufwändigen reinen Verlinkung auf fremde Ressourcen stehen die wenigsten lizenzbedingten Einschränkungen im Wege, allerdings ergeben sich für die Endnutzer beim Folgen eines Links immer Portalanwendungsbrüche. Zudem lassen sich die nur verlinkten Ressourcen nicht im eigenen Index speichern, und somit sind integrierte Suche und Facettierung über diese Daten nicht möglich. Auch bei einer dynamischen Anreicherung sind diese Funktionen nur schwer zu integrieren, da die Ressourcen nur in der Präsentationsschicht aufgerufen werden, was ausserdem u.U. auch zu Performanzproblemen führen kann. Wie bei blosser Verlinkung sind für stets aktuelle Daten aber keine Updates nötig und es können gegebenenfalls auch restriktiv lizenzierte Ressourcen integriert werden. Dadurch, dass bei Suchabfragen in dynamisch angereicherten Diensten die Daten automatisch in die Resultateseite eingebunden werden, lassen sich zudem Brüche in der Portalanwendung vermeiden. Während solche Einschränkungen der Usability durch die komplette Datenübernahme genauso vermieden werden, ist deren grösster Vorteil die Möglichkeit, durch optimale Integration in die Suchfunktionalitäten der eigenen Plattform einheitlich über alle Daten suchen zu können. Gleichzeitig ist durch die Vorhaltung der Daten im eigenen Backend die optimale Performanz der Suchoberfläche gewährleistet, und der Dienst macht sich nicht abhängig von der Infrastruktur der

[44] Vgl. Pascal Christoph, „Datenanreicherung auf LOD-Basis", in (Open) Linked Data in Bibliotheken, Bd. 50, Bibliotheks- und Informationspraxis (Berlin: De Gruyter, 2013), S. 139–67.

Datenlieferanten. Ein Nachteil der Verspeicherung im eigenen Index kann allerdings die zwingende Einschränkung auf offen lizenzierte Datenquellen sein. Auch ist ein erhöhter Integrationsaufwand gegenüber der anderen beiden Anreicherungsarten zu erwarten – abhängig davon wie stark die übernommen Metadaten an die IR-Umgebung angepasst werden müssen, je nachdem, welche Mappings bezüglich Formaten oder Vokabularien vorgenommen werden.
Primär kommt von den drei Arten der Anreicherung für ein IR wie BORIS im Hinblick auf die in Kapitel 3 besprochenen Nachnutzungs-Szenarien (RDS-Integration, Web-Nachnutzung) die komplette Datenübername in Frage. Denn nur durch komplette Integration der Fremddaten in die IR-Metadaten können sie anschliessend integral als Open Data zur Nachnutzung über die verschiedenen Kanäle wieder zur Verfügung gestellt werden. Daraus folgt, dass die Anreicherung durch Datenübernahme hauptsächlich in der Datenbank von BORIS geschehen sollte, auch wenn sich noch andere Ansatzpunkte für die Anreicherung denken lassen.[45] Unabhängig davon, auf welchen Wegen die IR-Metadaten nachgenutzt werden, ihre Reichhaltigkeit ist gegeben, ihre Attraktivität also optimal, was besonders für die zukunftsträchtige Web-Nachnutzung zentral sein wird. Somit beantworten diese Überlegungen auch die Frage nach dem primären Ort der Anreicherung: damit für alle Nachnutzungsszenarien die gleichen reichhaltigen Metadaten zur Verfügung stehen, soll die Anreicherung direkt an der Quelle, im IR vorgenommen werden. Schlussendlich soll ein IR ein gefragter Datenlieferant sein, nicht ein Aggregator, der nur Daten für Ad-hoc-Abfragen bündelt.
Wenn auch die drei Anreicherungsarten sich in einem spezifischen IR mischen können, macht die Anreicherung primär durch Datenübernahme am meisten Sinn und soll für die weiteren Überlegungen zum Ausgangspunkt genommen werden.

4.2. Anreicherungsverfahren

Die infrage kommenden Anreicherungsverfahren lassen sich im IR-Kontext grob in zwei Bereiche einteilen. Die manuelle und intellektuelle Anreicherung umfasst die bei der Dokumenten-Erfassung (durch die Forschenden oder auch das Bibliothekspersonal) vorgenommenen Eintragungen zur Inhaltsklassifizierung, Dokumentenidentifikation, Autorenkontrolle oder auch den manuellen Import von zusätzlichen Daten aus einer anderen Datenbank. Dabei können auch unterstützende Lookup-Tools zum Einsatz kommen. Die automatische Anreicherung hingegen

[45] Beispielsweise wäre als dynamische Anreicherung ein Mashup denkbar, das über verspeicherten Personenidentifikatoren (z.B. ORCID) Forschendenprofile mit zusätzlichen Informationen wie Affiliation, Forschungsgebiete etc. zu Autoren abruft, allerdings wäre damit höchstens die Attraktivität der eigenen Suchoberfläche des IR gesteigert, die Metadaten selber werden dadurch nicht reichhaltiger oder attraktiver für eine Nachnutzung.

bedeutet naturgemäss Bearbeitungen von grösseren Mengen von Datensätzen und umfasst Prozesse wie Matching, Merging, Mapping, und Transformation der Metadaten.
Grundsätzlich hängt das Anreicherungsverfahren immer direkt von den zur Anreicherung nutzbaren Quellen und den Vorgaben der anzureichernden Zielindices ab. In welcher Form stellen diese Quellen ihre Daten zur Verfügung? Über welche Wege muss ein spezifisches Datenset importiert werden, wie muss es bearbeitet werden? Daher ist es unmöglich und nicht sinnvoll, hier Anreicherungsverfahren über eine allgemeine Charakterisierung hinausgehend zu beschreiben.

4.3. Anreicherungsinhalte

Welche Inhalte für die Anreicherung in Frage kommen, hängt von verschiedenen Faktoren ab. Einen zentralen Faktor bilden die Erschliessungsstandards, die von den IR-Metadaten erfüllt werden sollen. Diese ergeben sich einerseits aus den Anforderungen im IR selber, andererseits im Hinblick auf eine Nachnutzung in weiteren Indices auch aus deren Voraussetzungen. In einem konkreten Setting, wie im beschriebenen Fall von BORIS, bestimmen diese Anforderungen die bekannten Zielindices der IR-Metadatennachnutzung (also die im hauseigenen RDS eingesetzten Indices) das Minimum, das an Reichhaltigkeit gewährleistet sein sollte. Alles, was über diesen Standard hinausgeht, ist aber grundsätzlich ebenso erwünscht, insbesondere im Hinblick auf die Nachnutzung der IR-Metadaten durch Dienste, die nicht im Hoheitsbereich der Heiminstitution liegen. Der Grundsatz der Reichhaltigkeit kommt auch in dieser Frage Bedeutung zu. Je mehr Erschliessungsdaten vorhanden sind, desto flexibler sind die Metadaten im Bezug auf die Bedürfnisse verschiedener Nachnutzer, die aus diesem „Überangebot" an Erschliessungsdaten jene Aspekte auswählen können, die für sie nützlich sind. Dieser Grundsatz entbindet allerdings nicht von einer Priorisierung der Anreicherungsinhalte. Ein wichtiges Priorisierungskriterium besteht in einer möglichst guten Abstimmung der einzelnen fachlich gegliederten Teilbestände eines IR auf die erwartete Erschliessungpraxis in der jeweiligen Disziplin. Übergreifend haben aber auch durch den Wissenschaftsbetrieb generell geforderte Ziele wie die Autoritätskontrolle eine hohe Priorität. Zudem ist die offene Lizenzierung eine zwingend zu erfüllende Vorbedingung für die urheberrechtskonforme Datenintegration.

4.3.1. Inhaltliche Erschliessungsdaten

Aufgrund ihrer internationale Verbreitung gehören kontrollierte Erschliessungsvokabularien wie die Dewey Decimal Classification (DDC), Library of Congress Classification (LCC), Library of Congress Subject Headings (LCSH) und die Medical Subject Headings (MeSH) zu den Favoriten für eine Integration, sowohl was ihr Potential zur Anreicherung aber auch zur anschliessenden Nachnutzung von IR-Einträgen betrifft. Zudem sind sie zur Anwendung auf ein breitestes Spektrum von Wissensgebieten geeignet, auch MeSH beschränkt sich nicht auf Medizin, sondern ist in weiteren Bereichen, insbesondere der Biowissenschaften, anwendbar. MeSH hat ausserdem in der Erschliessung von Zeitschriftenartikeln einen hohen Stellenwert durch die flächendeckende Anwendung im Index der medizinischen Publikationsdatenbank MEDLINE, deren Metadaten über PubMed frei zugänglich sind.[46]

Im Hinblick auf die Metadatenintegration in die hauseigenen Zielindices stehen vor allem die Gemeinsame Normdatei (GND) und MeSH im Vordergrund, da diese beiden Vokabulare zur thematischen Facettierung in baselbern.swissbib.ch und swissbib.ch genutzt werden. Zudem sind via Fremddatenübernahme auch LCSH und RAMEAU (Répertoire d'autorité-matière encyclopédique et alphabétique unifié) in den Metadaten der Zielindices eingebunden. Auch DDC und die Regensburger Verbundklassifikation (RVK) sind im Index des Aleph-Katalogs des IDS Basel Bern vertreten, werden aber zurzeit nicht zur Facettierung in baselbern.swissbib.ch eingesetzt. Die DDC wird von einigen IR als Browsinginstrument eingesetzt, jedoch beziehen sich diese Klassifikationsvergaben für einzelne Artikel oft nur auf die Zeitschriftenebene oder sie werden aufgrund der institutionellen Zugehörigkeit der Forschenden vergeben und sind dementsprechend grob. Um hier zu einer feineren Erschliessung zu gelangen, böten sich evtl. ein Eingabetool zur Klassifikationsvergabe durch die Forschenden selbst oder aber ein Rückimport der in BASE aufgrund von Abstracts texttechnologisch generierten DDC an.

Welche Vokabulare in kommerziellen RDS zur Anwendung kommen, ist im Einzelfall zu bestimmen. Der in baselbern.swissbib.ch eingesetzte Summon Index basiert primär auf den „ProQuest Subject Headings", unterstützt aber auch DDC und die integrierten BORIS-Bestände lassen sich anhand der übernommen DDC-Vergaben facettieren. Wie weit aber sich eine Anreicherung mit proprietären Vokabularien kommerzieller Anbieter lohnen würde, erscheint nicht zuletzt auch angesichts den Widersprüchen zum Open-Access-Gedanken fraglich.[47]

Der hohe Anteil an medizinischer Fachliteratur in BORIS lässt eine Priorisierung aufgrund der fachlichen Zuordnung der Repositorieninhalte zu. Allein die Publikationen der medizinischen

[46] Über 21 Millionen Artikel sind in Medline indexiert. Vgl. „Fact SheetMEDLINE, PubMed, and PMC (PubMed Central): How Are They Different?", Fact Sheets, zugegriffen 2. März 2015, http://www.nlm.nih.gov/pubs/factsheets/dif_med_pub.html.

[47] Nichtsdestotrotz hat die Columbia Universtity für ihr IR Academic Commons zur Facettierung des IR-Recherchetools die ProQuest Beschlagwortung übernommen.Vgl. Hilliker, Wacker, und Nurnberger, „Improving Discovery of and Access to Digital Repository Contents Using Semantic Web Standards", S. 2.

Fakultät machen über ein Drittel der gesamten Einträge in BORIS aus.[48] Eine Anreicherung dieses mengenmässig dominierenden Ausschnitts mit MeSH-Beschlagwortungen würde nicht nur einem fachlichen Standard entsprechen, sondern erscheint auch unter dem Aspekt der nahtlosen Einbindung in die Facettierung der Zielindices (swissbib.ch, baselbern.swissbib.ch) als lohnenswert.

Für Zeitschriftenliteratur besteht das grundsätzliche Problem, dass sie seltener inhaltlich erschlossen wird, respektive die betreffende Inhaltserschliessung nicht in offenen Quellen sondern eher in lizenzpflichtigen bibliographischen Datenbanken zu finden ist. Gerade für den Bereich Medizin bietet MEDLINE aber eine willkommene Ausnahme. Durch die grosse Abdeckung von BORIS-Einträgen durch MEDLINE und deren freien Zugänglichkeit über PubMed ergibt sich ein grosses Potential zur Nachnutzung dieser MeSH beinhaltenden Metadaten, das bei BORIS für die Erfassung von medizinischen Einträgen manuell mittels PubMed-ID[49] über ein Plugin zur Metadatenübernahme genutzt wird. Einer (automatisierten) Übernahme auch der MeSH-Daten nach BORIS stehen allerdings zwei Probleme im Wege. Zum einen hält das aktuell verwendete interne Metadatenformat kein Feld für MeSH bereit. So werden momentan MeSH-Terms nur manuell und selektiv in „keywords" Felder eingetragen, die sich aber nicht als MeSH kennzeichnen lassen und dementsprechend für eine spätere Nachnutzung ungenügend vorbereitet sind. Dieses Problem liesse sich mit der Anhebung auf ein genügend granulares Metadatenformat lösen. Mit Qualified DC beispielsweise wäre für MeSH eine eindeutige Feldcodierung möglich, da ein entsprechendes Vocabulary Encoding Scheme zur Verfügung steht.[50]

Zum anderen ist ein automatisches Harvesting nur für Berner Autoren via OAI-PMH wegen fehlendem entsprechend vordefinierten Set nicht möglich.[51] Hier bieten sich aber mögliche Lösungsansätze entweder über ein dem OAI-PMH-Harvesting nachgeschaltetes Programm, das ein Filtern nach Berner Autoren erlauben würde, oder über ein alternatives Protokoll für einen selektiven Datenaustausch. Houssos et al. haben ein Toolset entwickelt, das selektives Harvesting über OAI-PMH ermöglichen soll, offenbar sogar von Quellen, die nicht primär über OAI-PMH erreichbar sind. Dies wäre insofern sehr interessant, als die MEDLINE/PubMed-Metadaten nur in einem kleineren Ausschnitt – für die in PubMed Central in Open Access vorgehaltenen Inhalte – über OAI-PMH erhältlich sind.[52] Aufbauend auf dem SWORD-Protokoll (Simple Web-service

[48] Zurzeit enthält BORIS 53'761 Einträge, wovon 19'464 auf die medizinische Fakultät kommen. Potentiell erweitert sich der für MeSH-Beschlagwortung infrage kommende Anteil von BORIS noch um weitere Fächer der Life Sciences. in MEDLINE mit MeSH indexierten Einträge

[49] PubMed ID ist die von PubMed eingesetzte eindeutige Dokumentenidentifikationsnummer.

[50] „DCMI Metadata Terms: Vocabulary Encoding Schemes", zugegriffen 19. Januar 2015, http://dublincore.org/documents/dcmi-terms/#H4.

[51] Vgl. Kapitel 3.3.1. zu den Limitierungen von OAI-PMH.

[52] Vgl. Nikos Houssos u. a., „Enhanced OAI-PMH Services for Metadata Sharing in Heterogeneous Environments", *Library Review* 63, Nr. 6/7 (26. August 2014): 465–89, doi:10.1108/LR-05-2014-0051.

Offering Repository Deposit) und einer DSpace-Infrastruktur hat das MIT Repositorium erfolgreich einen Prozess erarbeitet, mit dem Metadaten aus BioMed Central, SpringerOpen und zukünftig möglicherweise auch Hindawi automatisch ins Repositorium übertragen werden können. Für eine erfolgsversprechende Lösung auf diesem Weg ist allerdings eine enge Zusammenarbeit mit dem Datenlieferanten unabdingbar, nicht zuletzt deshalb, weil über das SWORD-Protokoll der Austausch nicht vom Empfänger initiiert wird.[53]

4.3.2. Personen- und Körperschaftsidentifikatoren

Um die eindeutige Zuschreibung von Autorschaften mittels Namenseinträgen zu gewährleisten, sind Repositorien durch Dedublierung von Namensvarianten, Disambiguierung von gleichlautenden Namen, sowie durch Namensänderungen besonders gefordert. In vielen IR – wie auch bei BORIS – wird diese Arbeit primär über universitätsinterne Identifikatoren geleistet. Diese haben ihren Wert hauptsächlich für die Verwaltung in der Trägerinstitution und werden im IR einfach übernommen (Hauptzweck ist die Generierung von möglichst vollständigen Publikationslisten zur Evaluation des Forschungsoutputs) und eignen sich daher kaum zur universellen Nachnutzung. Um eine möglichst breite Interoperabilität zu erreichen, sind Anreicherung und Verknüpfung mit existierenden eindeutigen Identifikatoren mit einer weiten Verbreitung essentiell.
Die 2012 eingeführte Open Researcher and Contributor ID (ORCID) hat in der akademischen Welt schon einige Verbreitung erfahren und der u.a. von Thomson Reuters und Elsevier mitbegründeten Non-Profit-Initiative wird dementsprechend viel Potential für einen massgeblichen Standard im IR-Kontext zugesprochen.[54] Zu den Stärken von ORCID zählen die interdisziplinäre Verbreitung, die schnelle Verfügbarkeit, eine gute Interoperabilität mit wissenschaftlichen Datenbanken wie Europe PubMed Central, Scopus und Web of Science, sowie die gemeinfreie Lizenzierung des Datensets unter CC0.[55] ORCID können von Forschenden selber in kurzer Zeit erstellt werden, aber auch Organisationen können dies für ihre Mitglieder tun. Der einfache, dezentrale Registrierungsprozess ist aber zugleich ein Schwäche, denn doppelte

[53] Vgl. Ellen Finnie Duranceau und Sue Kriegsman, „Implementing Open Access Policies Using Institutional Repositories", in *The Institutional Repository: Benefits and Challenges* (Association for Library Collections & Technical Services, American Library Association, 2013), S. 84, http://dspace.mit.edu/handle/1721.1/76721; und Ellen Finnie Duranceau und Richard Rodgers, „Automated IR deposit via the SWORD protocol: an MIT/BioMed Central experiment", *Serials: The Journal for the Serials Community* 23, Nr. 3 (1. Januar 2010): 212–14, doi:10.1629/23212.
[54] Vgl. Lizzy A. Walker und Michelle Armstrong, „‚I cannot tell what the dickens his name is' - Name Disambiguation in Institutional Repositories", *Journal of Librarianship and Scholarly Communication* 2, Nr. 2 (2014): S. 2–3, doi:10.7710/2162-3309.1095. Zurzeit sind mehr als eine Million Profile registriert, diese wiederum sind mit mehr als sechs Millionen Publikationen verknüpft. Vgl. „ORCID Statistics", zugegriffen 4. März 2015, https://orcid.org/statistics.
[55] Vgl. Laure Haak, „ORCID Public Data File Use Policy", Text, (2. Mai 2013), https://orcid.org/content/orcid-public-data-file-use-policy.

Vergaben können nicht immer verhindert werden und grundsätzlich sind die Autoren selbst für die Zusammenführung von mehrfach vergebenen ORCID zuständig.[56]

Demgegenüber ist die Personennormdatei der GND sehr gut gepflegt und kontrolliert. Personen, Familien, Körperschaften und Konferenzen werden seit Juli 2014 nach dem internationalen Standard Resource Description and Access (RDA) erfasst,[57] wodurch eine optimale Interoperabilität vor allem mit bibliothekarischen Services, insbesondere auch durch die Einbindung in das Virtual International Authority File (VIAF), gegeben ist. Im Hinblick auf eine Nachnutzung in bibliothekarischen RDS ist die Anreicherung mit GND also eine naheliegende Option. Die Voraussetzung der offenen Verfügbarkeit wird mit der Lizenz CC0 erfüllt, zudem steht die GND als Linked Open Data zur Verfügung. Der hohe Qualitätsstandard der Personendaten in der GND ist allerdings mit grossem Aufwand seitens der bibliothekarischen Lokalredaktionen erkauft, die Erfassung einer Person braucht eine gewisse Zeit und kann nicht von Forschenden selbst vorgenommen werden.

Die Personennormdaten des von OCLC gehosteten Dienstes VIAF bieten aufgrund des weltweiten Netzes von beitragenden Institutionen einen riesigen Datenpool an verlinkten Normdaten aus verschiedener Provenienz, die zu Clustern geordnet eine eindeutige (auch als HTTP URI codierten) VIAF ID zugeschrieben erhalten. Die gesamte Datei steht unter einer Open Data Commons Attribution Lizenz (ODC-By)[58] zur Nachnutzung bereit. Da VIAF auf den Autoritätsdaten der teilnehmenden Nationalbibliotheken, Museen und Archiven basiert, kann ein neuer Normeintrag nicht direkt eingebracht werden und eine schnelle Erfassung eines neuen Normeintrags ist nicht möglich. Das verwendete automatische Verknüpfungsverfahren offenbart schon in wenigen Stichproben Schwächen punkto Dedublierung.[59] Wenn verschiedene Varianten durch das automatische Verfahren nicht verknüpft werden, erhalten sie je eigene VIAF ID und URI.[60]

Zwar sind Autorenidentifikatoren kommerzieller bibliographischer Datenbanken wie Scopus Author ID (Scopus, Elsevier) oder ResearcherID (Web of Science, Thomson Reuters) vielfältig mit Informationen verknüpft, sodass sich reichhaltige Forschenden-Profile ergeben. Auch erlauben diese Dienste oft das Laden der Publikationsmetadaten in verknüpfte ORCID-Profile, untereinander besteht aufgrund von Rivalitäten am Markt aber keine Interoperabilität. Für eine ResearcherID können sich Forschende selber registrieren, Scopus generiert ihre Author ID hingegen mittels Algorithmus auf Basis ihrer bibliographischen Datenbank, Korrekturen müssen von betroffenen Autoren selber beantragt werden. Für eine Integration, die auf Weiterverbreitung

[56] Vgl. Laure Haak, „Managing Duplicate ORCID iDs", Text, (9. Januar 2014), https://orcid.org/blog/2014/01/09/managing-duplicate-iDs.
[57] Vgl. „Informationsseite zur GND", zugegriffen 4. März 2015, https://wiki.dnb.de/display/ILTIS/Informationsseite+zur+GND.
[58] Vgl. „Open Data Commons Attribution License", zugegriffen 4. März 2015, http://opendefinition.org/licenses/odc-by/.
[59] Z.B. Rolf Zinkernagel, Regula Schatzmann, Thomas Klöti. Vgl. „VIAF", zugegriffen 4. März 2015, http://viaf.org/.
[60] Vgl. „Duplicate Detection and Resolution", zugegriffen 4. März 2015, http://www.oclc.org/services/metadata/quality/ddr.en.html.

von IR-Metadaten abzielt, sind diese proprietären Identifikatoren allerdings nicht geeignet, da sie nur im Rahmen einer entsprechenden Datenbank-Lizenzierung voll nutzbar sind und nicht als offene Datensets zur Verfügung stehen.

4.3.3. Dokumentenidentifikatoren, Dokumenttypen

Dokumentenidentifikatoren wie PubMed ID, DOI, URN sind nicht nur für eine eindeutige und persistente Zuordnung eines bibliographischen Eintrags zu einem Volltext wichtig.[61] Sie sind auch eine Vorbedingung für ein automatisches Rights Management, denn nur eine eindeutige Dokuementenidentifikation ermöglicht die maschinenlesbare Zuordnung von Lizenzen zu einem Dokument. Zudem erlaubt die Integration dieser Identifier als HTTP-URI eine Verlinkung auf mögliche zusätzliche Information. Über die PubMed ID gelangt man biespielsweise auf die korrespondierenden Einträge, die ausführlich mit MeSH beschlagwortet sind. Diese Funktion ist aber nicht zwingend gegeben, insbesondere nicht wenn DOIs auf Closed Access Plattformen verlinken. Grundsätzlich ist die Verspeicherung von Dokumentenidentifikatoren ein Beispiel von blosser Verlinkung als Anreicherungsform. Somit bestehen auch keine rechtlichen Einschränkungen, was die Nachnutzung der verspeicherten Information anbelangt.
Für die universitätsinterne Nachnutzung der IR-Metadaten können genaue Angaben zum Publikationstyp eine entscheidende Bedeutung haben. Eine Typenzuschreibung „review" ist beispielsweise problematisch, da nicht in allen Disziplinen dasselbe unter Review verstanden wird. Während in den Geisteswissenschaften eine Rezension als weniger gewichtige Leistung gewertet wird, ist eine Review in der Medizin ein anderes Format und hat den Status einer vollwertigen akademische Arbeit. Solchen Umständen muss auch in den beschreibenden Metadaten Rechnung getragen werden können, damit sie als solide Basis für die Evaluation dient. Ein mögliches Vokabular zu diesem Zweck bietet das Unterset „Publication Types" der MeSH.

4.3.4. Angaben zu Lizenz, Funding, Projektaffiliation

Das Management von Lizenzen ist für OA IR eine wichtige Aufgabe, da sie meistens auch Dokumente enthalten, die nur unter bestimmten Bedingungen, z.B. erst eine gewissen Zeit nach Publikationsdatum, Open Access zugänglich gemacht werden können. Grundsätzlich muss für jedes Dokument klar ersichtlich sein, unter welchen Bedingungen es konsumiert und

[61] Zwar werden Dokumente in OA IR auch mit dem Ziel einer bestmöglichen Langzeiterhaltung archiviert, je mehr Kopien eines Dokuments aber an verschiedenen Orten vorgehalten werden, desto besser ist dieses Ziel unterstützt (vgl. LOCKSS, Lots of Copies Keep Stuff Safe).

gegebenenfalls weiterverarbeitet werden kann.[62] Digitales Management von Lizenzen erfordert möglichst eindeutige Identifikatoren, einerseits für das Dokument und andererseits für die dazugehörige Lizenz. Dazu gibt es verschiedene Empfehlungen und Vokabularien, die für die Anwendung in OA IR entwickelt werden.

Die amerikanische National Information Standards Organization (NISO) empfiehlt zur Verwendung in Kombination mit dem Metadaten-Element license_reference stabile URIs als Identifikatoren, die auf eine frei zugängliche Ressource im WWW verweisen.[63] Diese Ressource kann ein PDF des Lizenztextes selber, aber auch eine Zusammenfassung auf einer HTML-Seite sein. Dabei sollen die URIs in der vom Lizenz-Provider empfohlenen Form verwendet werden. Im Fall von Creative Commons Lizenzen beispielsweise: http://creativecommons.org/licenses/by/2.0/. Das ursprünglich von DRIVER und OpenAIRE entwickelte Vokabular info:eu-repo [64] zur Beschreibung der Zugangsrechte ist besonders in Europa verbreitet und sieht „Identifier" zur Einbettung im HTML-Code in folgender Form vor: info:eu-repo/semantics/openAccess.[65] Zur Verdeutlichung wird die Kombination mit einem HTTP-URI, der auf eine Lizenz (z.B. CC) verweist, empfohlen. Für die Codierung von Embargos gibt es auch verschiedene Lösungsvorschläge, dies es gleichermassen ermöglichen, das Ablaufdatum einer restriktiven Lizenzierung und den Übergang zu einer offenen Lizenz in den Metadaten festzuhalten.[66]

Die Integration von Identifikatoren für geldgebende Organisationen, Projekte oder Stipendien ermöglichen eine eindeutige Zuordnung zu resultierendem Publikationsoutput. FundRef hält in einer offenen Datenbank[67] für über 9'000 Fonds als DOIs codierte persistente Identifikatoren vor (Bsp. SNF: http://dx.doi.org/10.13039/501100001711). Der Schweizerische Nationalfonds wiederum unterhält eine eigene Forschungsdatenbank (http://p3.snf.ch/), die URIs für die Fördernummern von Projekten und Stipendien verzeichnet.

Werden diese hier beschriebenen Identifikatoren neben menschenlesbarem Klartext auch in *maschinenlesbarer* Form, als HTTP-URIs in IR-Metadaten verspeichert, werden sie mit einem geeigneten Vokabular semantisch ausgezeichnet auch *maschinenverstehbar*. So lassen sich auf Basis von solchermassen codierten IR-Metadaten komplexe Abfragen durchführen, indem die Maschine verstehen kann, welche Personen für wie viele Publikationen aus welchen Fonds Unterstützung erhalten haben. Auch Identifikatoren für Lizenzen werden in einem geeigneten

[62] Die Lizenz für die Metadaten (CC0) in den Metadatensätze selbst nachzuweisen könnte heikel sein, die Zuordnung der Lizenzen für das Dokument, bzw. die beschreibenden Metadaten müsste zwingend eindeutig gemacht werden können.

[63] Vgl. National Information Standards Organization, *NISO RP-22-2015, Access License and Indicators: A Recommended Practice of the National Information Standards Organization*, 2015, S. 6, http://www.niso.org/publications/rp/rp-22-2015.

[64] „COAR » info:eu-repo", zugegriffen 4. März 2015, https://www.coar-repositories.org/activities/repository-interoperability/ig-controlled-vocabularies-for-repository-assets/wiki/info-eu-repo/.

[65] Die Terms in info:eu-repo haben zwar eine definierte Bedeutung, sind aber zurzeit nicht als HTTP-URI direkt abrufbar.

[66] Vgl. „RIOXX - UK Metadata Guidelines for Open Access Repositories, Version 3.0 January 2015", Januar 2015, S. 9, http://rioxx.net/guidelines/RIOXX_Metadata_Guidelines_v_3.0.pdf. Diese Empfehlung nimmt Bezug auf NISO RP-22-2015. Zudem erlaubt auch info:eu-repo die Codierung von Embargos, z.B: info:eu-repo/date/embargoEnd/2016-02-13.

[67] „FundRef Search beta", zugegriffen 5. März 2015, http://search.crossref.org/fundref; „FundRef Registry", zugegriffen 5. März 2015, http://www.crossref.org/fundref/fundref_registry.html. Im Registry ist die gesamte Liste der über 9'000 eingetragenen Fonds unter CC0 als RDF-Dokument veröffentlicht.

Vokabular[68] eingebettet eindeutig auf Dokumente beziehbar und die Metadaten so maschinell verarbeitbar und reichhaltig, dass sie für die Nachnutzung auch semantisch operierender Web-Services attraktiv sind.

5. LOD-Einsatz zur Anreicherung und Erfassung von Metadaten

Die Beschäftigung mit Austausch, Nachnutzung und Anreicherung von Metadaten in Open Access Repositorien bringt unweigerlich den Vernetzungsgedanken in den Vordergrund. Dabei spielen in diesem Kontext nicht nur die Vernetzung von digitalen Sammlungen und den dazugehörigen Metadaten durch die Integration in weitere Services eine Rolle, sondern ganz zentral ist das Ideal der grösstmöglichen Offenheit. Zum einen ist diese Offenheit definiert durch rechtliche Voraussetzungen, zum anderen basiert sie ganz wesentlich auf technologischen Infrastrukturen. Im Bemühen um verbesserte Sichtbarkeit für die Inhalte von OA IR weist ein starker Trend in Richtung Einsatz von neueren Web-Technologien zur noch besseren – semantischen – Vernetzung von Inhalten und Metadaten (vgl. Kapitel 3.3.). Aus diesem Grund sollen die Möglichkeiten von Linked Open Data zur Anreicherung von Metadaten in institutionellen Open-Access-Respositorien aufgezeigt werden.

5.1. Gründe für LOD-Einsatz zur Anreicherungsarbeit

Linked Open Data ist eine aktuell viel diskutierte Technologie, nicht nur in der in der weiteren WWW-Community, sondern auch in der Bibliothekswelt. Das Bestreben, Bibliotheken und ihre Dienstleistungen mittels neuer Technologien an der vernetzten Welt teilhaben zu lassen, äussert sich seit längerem. In diesem Kontext entstanden beispielsweise ein neues Format (BIBFRAME) und ein neues Regelwerke (RDA), denen ein Datenmodell (FRBR) zugrunde liegt, bei dem die vielfältige bedeutungstragende Vernetzung von Entitäten der zentrale theoretische Gedanke ist und das sich auf ein in der Informatik geprägtes Entity-Relationship-Modell bezieht. Mittlerweile haben mit u.a. OCLC, der Library of Congress und der Deutschen Nationalbibliothek einige grosse Player eigene LOD-Projekte initiiert, womit sich LOD über einen anfänglichen Hype hinaus gut etabliert zu haben scheint. Vor diesem Hintergrund erstaunt es nicht, das LOD insbesondere auch in der informatikaffinen OA-IR-Szene als zukunftsträchtiges Ziel für die Weiterentwicklung von

[68] Für die Lizenzenbeschreibung eignen sich z.B. schema.org (property: „license") oder die Creative Commons Rights Expression Language ccREL (RDF/RDFa-basiert). Vgl. „ccREL: The Creative Commons Rights Expression Language", zugegriffen 5. März 2015, http://www.w3.org/Submission/ccREL/.

Repositorien gesehen wird. Dies kommt beispielsweise in der Grundthese zum Ausdruck, „[…] dass LOD prädestiniert ist, einen wichtigen Eckpfeiler einer nachhaltigen Metadateninfrastruktur für die Wissenschaft zu bilden."[69]

Die meisten der unter 4.3. beschriebenen für Anreicherung und Nachnutzung interessanten Datenbestände liegen schon als LOD-Datensets vor. Das heisst, als Linked *Open* Data sind sie frei nutzbar, die offene Lizenzierung macht also die komplette Datenübernahme möglich. LOD ist auf Nachnutzung hin angelegt und daher geradezu prädestiniert für den Verwendungszweck Datenanreicherung. Dass die zur Nachnutzung geeignete Normdateien und Identifikatoren in vielen Fällen bereits als LOD vorliegen (GND, DDC, MeSH, VIAF, u.a.),[70] heisst aber auch, dass die Entitäten dieser Datenpools jeweils als eindeutig referenzierte URIs vorhanden sind, wodurch die maschinelle Verarbeitung gerade im Hinblick auf die RDS-Integration und damit verbundene Mappings von Metadatensätzen besonders gut automatisiert erstellt werden können.[71] Da in LOD auch die Bedeutung von Beziehungen zwischen Entitäten maschinengerecht codiert sind, werden Anwendungen möglich, die auf der semantischen Auswertbarkeit von Fremddaten beruhen. Beispielsweise lassen sich umfangreiche Crosskonkordanzen zwischen verschiedenen Erschliessungsvokabularien mittels automatisiertem Vocabulary Alignment erstellen.[72]

Die Nachnutzung von Linked Open Data ist aber immer potentiell eine Vorbereitung für die Einbindung ins Semantic Web (vgl. Kapitel 3.3.3.), denn wenn nicht nur strings, sondern things, also Entitäten und nicht blosse Zeichenketten, in die IR-Datenbestände integriert werden, haben die Suchmaschinen der neuen Generation die besten Voraussetzungen IR-Inhalte ins semantische Netzwerk einzubinden. Selbst wenn die dabei entstandenen Anreicherungen nicht selbst wieder als LOD zur Verfügung stehen, kann die Verwendung von LOD-Technologie also ein Schritt in Richtung Semantic Web darstellen.

5.2. Was ist Linked Open Data?

In Kapitel 3.3.3. war davon die Rede, wie Beziehungen von Entitäten mit Hilfe von Ontologien semantisch ausgezeichnet werden können, sodass die Bedeutung dieser Beziehungen von Maschinen „verstanden", d.h. ähnlich interpretiert werden können wie Menschen dies tun

[69] Vgl. Pascal Christoph und Adrian Pohl, „Dezentral, offen, vernetzt – Überlegungen zum Aufbau eines LOD-basierten FID-Fachinformationssystems", *Bibliothek Forschung und Praxis* 38, Nr. 1 (2014): S. 114, doi:10.1515/bfp-2014-0005.

[70] Vgl. „Datahub", zugegriffen 30. Dezember 2014, http://datahub.io/. Dies ist eine von der Open Knowledge Foundation betriebene Datenbank, die grössere offene Datasets (v.a. open government data) verzeichnet. Ein Untersammlung des Dathub verzeichnet LOD-Datasets.

[71] Metafacture gilt als ein LOD-basiertes Werkzeug, mit dem sich ohne vertiefte Programmierkenntnisse Mappings definieren und ausführen lassen. Vgl. Christoph und Pohl, „Dezentral, offen, vernetzt – Überlegungen zum Aufbau eines LOD-basierten FID-Fachinformationssystems", S. 121.

[72] Vgl. Kapitel 5.3.2. weiter unten.

würden. Diese Berechenbarkeit von Bedeutung ist das grundlegende Prinzip des Semantic Web. Linked Data (wovon LOD eine besondere Variante darstellt) beruht primär auf denselben Grundsätzen und benützt ähnliche Techniken, insbesondere Vokabularien (Ontologien) zur Modellierung und Definition der Beziehungen zwischen Entitäten. Eine klare Unterscheidung der Begriffe „Semantic Web" und „Linked Data" ist nicht einfach, und die Begriffe werden oft gleichbedeutend verwendet.[73] Vielleicht lässt sich verkürzend sagen, dass Linked Data eine Spielart des Semantic Web darstellt, bei der der Fokus ganz auf der Verlinkung von Entitäten liegt. Linked Data bedeutet also nicht die Verknüpfung von Dokumenten, sondern von Daten. Diese Verlinkung von Daten wird in den vier als Linked Data Principles bekannt gewordenen Konventionen beschrieben:[74]

1. Use URIs as names for things.
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information, using the standards (RDF*, SPARQL).
4. Include links to other URIs so that they can discover more things.

Die vier Prinzipien besagen also, dass Linked Data auf den allgemeinen Web-Standards Uniform Resource Identifier (URI) und Hypertext Transfer Protocol (HTTP) aufbauen, dazu aber noch das Resource Description Framework (RDF) als Datenmodell verwendet, um die Verlinkungen zu definieren, also mit Bedeutung versehen zu können.[75] Werden nun nach diesen Prinzipien verknüpfte Daten noch offen lizenziert, spricht man von Linked Open Data.

5.3. Anwendungsfelder für LOD-Technologien in der Datenanreicherung

Im Folgenden sollen exemplarisch drei Anwendungsfelder für LOD-basierte Metadatenanreicherung im IR-Kontext skizziert werden. Die tatsächliche Umsetzung der einzelnen Anwendungen sind aber u.U. von der vorhandenen Infrastruktur und Formatvorgaben

[73] Tim Berners-Lee wird die Prägung beider Begriffe zugeschrieben (2001 und 2006) und er wird zitiert: „Linked data is semantic web done right." Vgl. „Frequently Asked Questions (FAQs) | Linked Data - Connect Distributed Data across the Web", zugegriffen 6. März 2015, http://linkeddata.org/faq.

[74] Vgl. Tim Berners-Lee, „Linked Data - Design Issues", 2009 2006, http://www.w3.org/DesignIssues/LinkedData.html.

[75] Das Grundprinzip von RDF ist die Darstellung von Aussagen als Tripel: Subjekt – Prädikat – Objekt, wobei Subjekt und Prädikat HTTP-URIs sein müssen, das Objekt aber auch Freitext, ein sogenanntes Literal sein kann. Das Prädikat stellt die definierte Beziehung zwischen den Entitäten dar, indem es via HTTP-URI auf einen bestimmten Term einer Ontologie Bezug nimmt. Ein RDF-Tripel kann auch als die kleinste Form eines Graphen bezeichnet werden. Je mehr Verknüpfungen gemacht werden, desto grösser wird ein Graph. SPARQL (SPARQL Protocol and RDF Query Language) ist die Standardabfragesprache, mit der Abfragen über einen RDF-Graphen möglich sind. Vgl. beispielsweise Pohl und Danowski, „Linked Open Data in der Bibliothekswelt", S. 23–26.

abhängig. Die Voraussetzung ist aber nicht in allen Fällen eine RDF-basierte Infrastruktur (bspw. sind Autosuggest-Funktionalitäten ohne IR-Datenhaltung in einem Triple-Store möglich). Auch bedingt die Nachnutzung der über LOD-basierte Anreicherung entstandenen Produkte nicht zwingend eine LOD-Infrastruktur, denn die Ausgabe in herkömmlichen Metadatenformaten ist immer möglich.[76] In einem LOD-basierten Zukunftsszenario wären aber sowohl IR als auch RDS als RDF-basierte Triple Stores denkbar, die untereinander optimal interoperieren könnten und deren Einbindung auch in die LOD-Cloud gegeben wäre, weil ihre als Graphen vorgehaltenen bibliographischen Datenpools mit externen Graphen verknüpft wären.

5.3.1. Autosuggest-Tools

Gute LOD-basierte Lookup-Tools liefern über eine Suchvorschlagsfunktion einerseits einen Wortvorschlag, wie man es von vielen Suchschlitzen als Funktionalität gewohnt ist. Hinter dem Klartextvorschlag verbirgt sich aber ein Link zur Beschreibung des vorgeschlagenen Begriffs, über den das Tool zusätzliche Informationen abrufen und einblenden kann. Diese über Links abgerufenen Informationen ermöglichen den Benutzenden die Entscheidung zwischen zwei ähnlichen oder gleichlautenden Begriffen, Personennamen oder Körperschaften, ermöglichen also eine schnelle Disambiguation und Auswahl des richtigen Bezeichners.[77] Bei der Auswahl eines Bezeichners zur Erschliessung eines Dokuments wird in der Folge die dazugehörige Entität (als HTTP-URI) gespeichert, womit ein eindeutiger Identifikator und nicht bloss Freitext in die Metadaten übernommen wird. In ähnlicher Weise lassen sich solche Autosuggest-Tools auch für die Recherche einsetzen, mit dem Unterschied, dass die gewählte Entität nicht gespeichert wird, sondern die mit ihr verknüpften Dokumente oder Informationen abgerufen werden.
Das Hochschulbibliothekszentrum des Landes Nordrhein-Westfalen (hbz) bietet innerhalb seines LOD-Services (http://lobid.org/) eine solche LOD-basierte Vorschlagsuche zur freien Weiterverwendung an.[78] Dieses Autosuggest-Tool greift tagesaktuell auf das LOD-Datenset der GND zu und lässt sich in andere Webanwendungen einbauen und ist grundsätzlich um weitere Datensets erweiterbar. Im IR-Kontext wäre eine solche Vorschlagsuche beispielsweise als Tool zur unterstützten Inhaltserschliessung mit kontrolliertem Vokabular durch Forschende gleich bei der Deposition von Publikationen im IR denkbar. In ähnlicher Weise könnte damit eine tiefere DDC-Erschliessung von Artikeln LOD-basiert unterstützt werden. Weitere Datensets sind für die

[76] Vgl. Christoph, „Datenanreicherung auf LOD-Basis", S. 164.
[77] Eine solche Funktion baut also auf den Linked-Data-Prinzipien auf, wonach HTTP-URIs verwendet werden sollen, und dazu durch die Anwendung von RDF nützliche Information zur Verfügung gestellt werden soll, damit Menschen Begriffe nachschlagen und dazu reichhaltige Informationen erhalten können.
[78] Vgl. „lobid - api", zugegriffen 7. März 2015, http://api.lobid.org/api. Zwar sind die Funktionalitäten dieser Vorschlagsuche nicht so ausgereift wie oben beschrieben, aber da die Software offen ist, liesse sie sich gegebenenfalls dahin weiterentwickeln.

Integration in eine Vorschlagsuche denkbar (ORCID, MeSH, FundRef, ...), wobei als Bedingung für die Umsetzung immer die offene Lizenz vorhanden sein muss.

5.3.2. Vocabulary-Alignment: Beispiel MeSH – GND

Für die Überführung von vorhandenen Beschlagwortungen in die Terme eines anderen Erschliessungsvokabular (z.B. MeSH nach GND), bietet sich als LOD-basierte Technologie Vocabulary-Alignment an.[79] Auf der Grundlage eines solchen Vocabulary-Alignment lassen sich äquivalente Terme unterstützt durch automatisierte Prozesse verknüpfen und somit vom einen System ins andere übersetzen. Die Grundlagenarbeit besteht darin, dass die beiden Vokabulare mittels einer geeigneten Ontologie zur Darstellung von Wissensorganisationssystemen in RDF modelliert werden. Die W3C-Empfehlung SKOS (Simple Knowledge Organization System)[80] ermöglicht die semantische Modellierung von Thesauri, Klassifikationen oder sonstigen kontrollierten Vokabularen und kann somit auch der Verknüpfung von Termen verschiedener Begriffssysteme in einer Crosskonkordanz dienen. Als Beispiel einer solcherart erstellten Crosskonkordanz kann AGROVOC dienen, ein agrarwissenschaftlicher, multilingualer Thesaurus der zurzeit mit Teilbeständen aus 16 weiteren Vokabularen verknüpft und auch als LOD veröffentlicht ist.[81]

Für die GND wurde eine auf OWL basierende Ontologie entwickelt, die schlussendlich die Verlinkung zu verschiedenen weiteren Erschliessungssystemen ermöglichen soll, darunter die DDC (deutsch), LCSH, VIAF, RAMEAU.[82] Sind diese Arbeiten auch zu einem grossen Teil noch Zukunft, so lässt sich doch konzeptuell denken, wie beispielsweise eine LOD-basierte Verknüpfung und Anreicherung von MeSH-Termen (aus einem PubMed-Import ins IR) mit GND-Termen ermöglichen könnte. Auf diese Weise angereicherte Metadaten würden für die Nachnutzung in Indices mit GND-basierter Facettierung einen Mehrwert darstellen. Dies wäre im Fall der Integration von BORIS-Inhalten in swissbib.ch allerdings abhängig von einer Anhebung

[79] Im Linked Data Glossary des W3C wird Vocabulary Alignment definiert als „the process of analyzing multiple vocabularies to determine terms that are common across them and to record those relationships." Unter Vocabulary können gemäss Glossary einfache Sammlungen von Termen, aber auch sehr umfangreiche Erschliessungsvokabulare verstanden werden. Vgl. „Linked Data Glossary", zugegriffen 8. März 2015, http://www.w3.org/TR/ld-glossary/#vocabulary-alignment.

[80] „Mit SKOS soll die einfache Veröffentlichung und Kombination kontrollierter, strukturierter und maschinenlesbarer Vokabulare für das Semantische Web ermöglicht werden." Vgl. „Simple Knowledge Organization System - Wikipedia, the free encyclopedia", zugegriffen 7. März 2015, http://en.wikipedia.org/wiki/Simple_Knowledge_Organization_System.

[81] Vgl. „AGROVOC Linked Open Data | Agricultural Information Management Standards (AIMS)", zugegriffen 8. März 2015, http://aims.fao.org/standards/agrovoc/linked-open-data; und Caterina Caracciolo u. a., „The AGROVOC Linked Dataset", *Semantic Web*, Mai 2013, http://eprints.rclis.org/20648/.

[82] Vgl. Alexander Haffner, *Internationalisierung der GND durch das Semantic Web* (Frankfurt am Main: KIM, 16. Juli 2012), S. 12, http://www.kim-forum.org/Subsites/kim/SharedDocs/Downloads/DE/Berichte/internationalisierungDerGndDurchDasSemanticWeb.html.

des Metadatenformats von Dublin Core auf ein höher granulares Format, das die Auszeichnung von GND unterstützt.

5.3.3. Mapping, automatisierte Datenanreicherung

Bereits existieren Open Source Softwarelösungen zur automatisierten Datentransformation (z.B. eignet sich Metafacture für das Erstellen von Mappings) und Datenanreicherung (z.B. Silk, eine Lösung die bei lobid.org zur Anreicherung mit DBpedia-Daten angewendet wurde) auf LOD-Basis.[83] Dabei sind allerdings beide zu verknüpfenden Datensets als LOD (als RDF-Dumps) nötig. Wo also keine LOD-Infrastruktur vorhanden ist, muss vorgängig eine Konversion der Daten in LOD-konformes RDF stattfinden.[84] Mit Silk lassen sich aufgrund von eindeutigen Identifiern Beziehungen zwischen Metadatensätzen herstellen, die dann als RDF-Tripel gespeichert werden, sodass verschiedene Ressourcen zu einem gemeinsamen Identifier verlinken und unter diesem gebündelt werden können. Durch geeignete – auch intellektuelle – Postprozessierung zur Disambiguierung werden die Datensätze für das Merging vorbereitet. Für das Zusammenführen können mit Silk Algorithmen konfiguriert werden, die bestimmen, in welchen Fällen welche Felder übernommen werden sollen.[85]

Im IR-Kontext könnte ein solches Verfahren bei der Anreicherung von IR-Inhalten mit MeSH-Daten aus PubMed zur Anwendung kommen. Dabei wäre der eindeutige Identifikator die PubMed ID oder die DOI-Nummer. Hier bleibt aber zu fragen, ob die Anwendung von LOD-Technologie zur automatisierten Metadatenanreicherung immer Vorteile gegenüber nicht LOD-basierten Technologien bringt. LOD bietet eine Technologie, die vieles verspricht, aber in bestimmten Anwendungsfällen die Aufgabe nicht unbedingt immer besser oder schneller lösen kann als andere Technologien (beispielsweise sind mit SPARQL sehr mächtige Abfragen realisierbar, für die RDS-Recherche ist ein SPARQL-Endpoint aber definitiv kein Gewinn an Usability für den gemeinen Nutzer). Der grosse Vorteil einer LOD-basierten Verarbeitung kann darin gesehen werden, dass das Endprodukt bestmöglich auf das Semantic Web vorbereitet ist, indem es semantisch definierte Beziehungen zwischen als HTTP-URIs codierten Entitäten herstellt und so für eine maschinelle Nachnutzung optimiert ist.

---

[83] Vgl. Christoph und Pohl, „Dezentral, offen, vernetzt – Überlegungen zum Aufbau eines LOD-basierten FID-Fachinformationssystems", S. 121.

[84] Konvertierungstools wie OAI2LOD Server können LOD konforme RDF-Repräsentationen von relationalen Datenbanken erstellen. Vgl. Bernhard Haslhofer und Bernhard Schandl, „The OAI2LOD Server: Exposing OAI-PMH Metadata as Linked Data" (International Workshop on Linked Data on the Web (LDOW2008), co-located with WWW 2008, Beijing, 2008), http://events.linkeddata.org/ldow2008/papers/03-haslhofer-schandl-oai2lod-server.pdf.

[85] Diese Darstellung lehnt sich an den Prozessbeschreibungen für ein Anreicherungsprojekt in lobid.org an. Vgl. Christoph, „Datenanreicherung auf LOD-Basis", S. 153–157.

6. Zusammenfassung & Ausblick

Institutionelle Open Access Repositorien sind zwar in der Regel eng mit Bibliotheken verbunden, haben aber diesen gegenüber aufgrund ihrer etwas anders gearteten Funktionen zunächst andere Prioritäten in der Herstellung von Findbarkeit und Zugänglichkeit von Wissen. Zum einen sind OA IR Evaluationsinstrument wie auch Präsentationsplattform für die Trägerinstitution, zum anderen sind sie dem Open Access Gedanken verpflichtet. Somit sind Offenheit und Sichtbarkeit von Anfang an zentrale Punkte in der Bereitstellung von Inhalten. OA IR stehen meist ausserhalb der traditionellen bibliothekarischen Kataloginfrastruktur und haben weniger rigide Anforderungen an die bibliographische Kontrolle, mit der sie ihre Dokumente beschreiben. Der Gedanke der Offenheit und die Anbindung an das WWW hat Priorität.

Dennoch kommen auch OA IR nicht umhin, sich neue Distributionswege zu erschliessen und bestehende weiterzuentwickeln. Dabei werden die verbreiteten Standards OAI-PMH und Dublin Core herausgefordert, da durch neue Nachnutzungsszenarios reichere und besser auf die Nachnutzungsziele abgestimmte Metadaten verlangt werden. Zum einen sind diese Ziele bibliothekarische Resources Discovery Services, zum andern eine sich verändernde WWW-Landschaft, in der vermehrt ganz anders strukturierte Metadaten und Inhalte gefragt sind, wenn die Sichtbarkeit über Websuchmaschinen optimal genutzt werden soll. Hier zeigen sich Techniken der Suchmaschinenoptimierung als vielversprechende Möglichkeiten, sich um eine bessere Verständlichkeit der IR-Metadaten zu bemühen. Dabei ergibt sich mit der semantischen Auszeichnung von Metadaten auch die Chance, ein IR an zukünftige Standards semantischer Suchmaschinentechnologie anzupassen.

Neben der Darbietungsform bestimmen aber auch die Inhalte der Metadaten ihre Attraktivität für eine Nachnutzung durch die verschiedenen Systeme. Die dazu nötige Reichhaltigkeit von Metadaten kann neben primärer Erschliessungsarbeit durch Anreicherung erreicht werden. Von Bibliotheken und Repositorienbetreibern schon geleistete Arbeiten im Bereich Linked Open Data geben Hinweise darauf, wie diese Technologien die Anreicherungs- und Erschliessungsarbeit durch Autosuggest-Tools, Vocabulary Alignment und Anreicherungsverfahren erleichtern und zumindest teilweise automatisieren können. LOD könnte sich somit als ein Feld erweisen, auf dem sich sowohl die Metadaten selbst, wie auch ihre Verbreitungswege optimieren lassen. Indem sich IR-Repositorien auf das Semantic Web zubewegen und so ihre Interoperabilität und Sichtbarkeit weiter stärken, machen sie sich zukunftsfähig. Unter diesen Aspekten lohnt es sich, das LOD-Feld weiter zu beobachten.

Bibliographie

Arlitsch, Kenning, und Patrick S. O'Brien. „Invisible institutional repositories". *Library Hi Tech* 30, Nr. 1 (2. März 2012): 60–81. doi:10.1108/07378831211213210.

Becker, Pascal-Nicolas. „Repositorien und das Semantic Web – Repositorieninhalte als Linked Data bereitstellen", 2014. http://www.pnjb.de/uni/diplomarbeit/repositorien_und_das_semantic_web.pdf.

Borst, Timo. „Repositorien auf ihrem Weg in das Semantic Web: Semantisch hergeleitete Interoperabilität als Zielstellung für künftige Repository-Entwicklungen". *Bibliothek Forschung und Praxis* 38, Nr. 2 (2014): 257–65. doi:10.1515/bfp-2014-0034.

Calhoun, Karen. „Supporting Digital Scholarship: Bibliographic Control, Library Cooperatives and Open Access Repositories". In *Catalogue 2.0 - the future of the library catalogue*, 143–77. London: Facet Publishing, 2013.

Caracciolo, Caterina, Armando Stellato, Ahsan Morshed, Gudrun Johannsen, Sachit Rajbahndari, Yves Jaques, und Johannes Keizer. „The AGROVOC Linked Dataset". *Semantic Web*, Mai 2013. http://eprints.rclis.org/20648/.

Christoph, Pascal. „Datenanreicherung auf LOD-Basis". In *(Open) Linked Data in Bibliotheken*, 50:139–67. Bibliotheks- und Informationspraxis. Berlin: De Gruyter, 2013.

Christoph, Pascal, und Adrian Pohl. „Dezentral, offen, vernetzt – Überlegungen zum Aufbau eines LOD-basierten FID-Fachinformationssystems". *Bibliothek Forschung und Praxis* 38, Nr. 1 (2014): 114–23. doi:10.1515/bfp-2014-0005.

DeRidder, Jody L. „Googlizing a Digital Library". *The Code4Lib Journal*, Nr. 2 (24. März 2008). http://journal.code4lib.org/articles/43.

Duranceau, Ellen Finnie, und Sue Kriegsman. „Implementing Open Access Policies Using Institutional Repositories". In *The Institutional Repository: Benefits and Challenges*, 75–97. Association for Library Collections & Technical Services, American Library Association, 2013. http://dspace.mit.edu/handle/1721.1/76721.

Duranceau, Ellen Finnie, und Richard Rodgers. „Automated IR deposit via the SWORD protocol: an MIT/BioMed Central experiment". *Serials: The Journal for the Serials Community* 23, Nr. 3 (1. Januar 2010): 212–14. doi:10.1629/23212.

Haffner, Alexander. *Internationalisierung der GND durch das Semantic Web*. Frankfurt am Main: KIM, 16. Juli 2012. http://www.kimforum.org/Subsites/kim/SharedDocs/Downloads/DE/Berichte/internationalisierungDerGndDurchDasSemanticWeb.html.

Haslhofer, Bernhard, und Bernhard Schandl. „The OAI2LOD Server: Exposing OAI-PMH Metadata as Linked Data". Beijing, 2008. http://events.linkeddata.org/ldow2008/papers/03-haslhofer-schandl-oai2lod-server.pdf.

Hilliker, Robert J., Melanie Wacker, und Amy L. Nurnberger. „Improving Discovery of and Access to Digital Repository Contents Using Semantic Web Standards: Columbia University's Academic Commons". *Journal of Library Metadata* 13, Nr. 2–3 (Juli 2013): 80–94. doi:10.1080/19386389.2013.826036.

Houssos, Nikos, Kostas Stamatis, Panagiotis Koutsourakis, Sarantos Kapidakis, Emmanouel Garoufallou, und Alexandros Koulouris. „Enhanced OAI-PMH Services for Metadata Sharing in Heterogeneous Environments". *Library Review* 63, Nr. 6/7 (26. August 2014): 465–89. doi:10.1108/LR-05-2014-0051.

Jacsó, Péter. „Google Scholar Author Citation Tracker: is it too little, too late?". *Online Information Review* 36, Nr. 1 (17. Februar 2012): 126–41. doi:10.1108/14684521211209581.

Jones, Richard E., Theo Andrew, und John MacColl. *The Institutional Repository*. Oxford: Chandos, 2006.

Klee, Carsten. „Vokabulare für bibliographische Daten: Zwischen Dublin Core und bibliothekarischem Anspruch". In *(Open) Linked Data in Bibliotheken*. Berlin, Boston: De Gruyter, 2013. http://www.degruyter.com/view/books/9783110278736/9783110278736.45/9783110278736.45.xml.

Lösch, Mathias. „Automatische Klassifikation von OAI-Metadaten mit linguistischen Methoden. Vortrag im Kolloquium Wissensinfrastruktur an der UB Bielefeld". Bielefeld, 30. Oktober 2009.

Mixter, Jeff Keith, Patrick S. O'Brien, und Kenning Arlitsch. „Describing Theses and Dissertations Using Schema.org". *International Conference on Dublin Core and Metadata Applications*, 8. Oktober 2014, 138–46.

National Information Standards Organization. *NISO RP-22-2015, Access License and Indicators: A Recommended Practice of the National Information Standards Organization*, 2015. http://www.niso.org/publications/rp/rp-22-2015.

Pohl, Adrian, und Patrick Danowski. „Linked Open Data in der Bibliothekswelt: Grundlagen und Überblick". In *(Open) Linked Data in Bibliotheken*. Berlin, Boston: De Gruyter, 2013. http://www.degruyter.com/view/books/9783110278736/9783110278736.1/9783110278736.1.xml.

„RIOXX - UK Metadata Guidelines for Open Access Repositories, Version 3.0 January 2015", Januar 2015. http://rioxx.net/guidelines/RIOXX_Metadata_Guidelines_v_3.0.pdf.

Suber, Peter. „How to facilitate Google crawling of OA repositories". Zugegriffen 23. Februar 2015. http://legacy.earlham.edu/~peters/fos/googlecrawling.htm.

Walker, Lizzy A., und Michelle Armstrong. „‚I cannot tell what the dickens his name is' - Name Disambiguation in Institutional Repositories". *Journal of Librarianship and Scholarly Communication* 2, Nr. 2 (2014). doi:10.7710/2162-3309.1095.

Webressourcen

„Academic Commons". Zugegriffen 26. Februar 2015. http://academiccommons.columbia.edu/.

„AGROVOC Linked Open Data | Agricultural Information Management Standards (AIMS)". Zugegriffen 8. März 2015. http://aims.fao.org/standards/agrovoc/linked-open-data.

„BASE Weblog: 10 Jahre BASE". Zugegriffen 8. Februar 2015. http://ekvv.uni-bielefeld.de/blog/base/entry/10_jahre_suchmaschine_base.

„Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities". Zugegriffen 1. Februar 2015. http://openaccess.mpg.de/Berlin-Declaration.

Berners-Lee, Tim. „Linked Data - Design Issues", 2009 2006. http://www.w3.org/DesignIssues/LinkedData.html.

„ccREL: The Creative Commons Rights Expression Language". Zugegriffen 5. März 2015. http://www.w3.org/Submission/ccREL/.

„COAR » info:eu-repo". Zugegriffen 4. März 2015. https://www.coar-repositories.org/activities/repository-interoperability/ig-controlled-vocabularies-for-repository-assets/wiki/info-eu-repo/.

„Datahub". Zugegriffen 30. Dezember 2014. http://datahub.io/.

„DCMI Metadata Terms". Zugegriffen 22. Februar 2015. http://dublincore.org/documents/2012/06/14/dcmi-terms/?v=elements#subject.

„DCMI Metadata Terms". Zugegriffen 22. Februar 2015. http://dublincore.org/documents/2012/06/14/dcmi-terms/?v=dcam#H4.

„Duplicate Detection and Resolution". Zugegriffen 4. März 2015. http://www.oclc.org/services/metadata/quality/ddr.en.html.

„Fact SheetMEDLINE, PubMed, and PMC (PubMed Central): How Are They Different?". Fact Sheets. Zugegriffen 2. März 2015. http://www.nlm.nih.gov/pubs/factsheets/dif_med_pub.html.

„Frequently Asked Questions (FAQs) | Linked Data - Connect Distributed Data across the Web". Zugegriffen 6. März 2015. http://linkeddata.org/faq.

„FundRef Registry". Zugegriffen 5. März 2015. http://www.crossref.org/fundref/fundref_registry.html.

„FundRef Search beta". Zugegriffen 5. März 2015. http://search.crossref.org/fundref.

Haak, Laure. „Managing Duplicate ORCID iDs". Text, 9. Januar 2014. https://orcid.org/blog/2014/01/09/managing-duplicate-iDs.

———. „ORCID Public Data File Use Policy". Text, 2. Mai 2013. https://orcid.org/content/orcid-public-data-file-use-policy.

„Home – Blacklight". Zugegriffen 27. Februar 2015. http://projectblacklight.org/.

„Inclusion Guidelines for Webmasters: Indexing Guidelines". Zugegriffen 24. Februar 2015. http://scholar.google.com/intl/en/scholar/inclusion.html#indexing.

„Informationsseite zur GND". Zugegriffen 4. März 2015. https://wiki.dnb.de/display/ILTIS/Informationsseite+zur+GND.

„Introducing the Knowledge Graph: things, not strings". *Official Google Blog*. Zugegriffen 25. Februar 2015. http://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html.

„Invenio". Zugegriffen 27. Februar 2015. http://invenio-software.org/.

„Islandora Website". Zugegriffen 27. Februar 2015. http://islandora.ca/.

„Linked Data Glossary". Zugegriffen 8. März 2015. http://www.w3.org/TR/ld-glossary/#vocabulary-alignment.

„lobid - api". Zugegriffen 7. März 2015. http://api.lobid.org/api.

„OAI 2.0 Request Results". Zugegriffen 23. Februar 2015. http://boris.unibe.ch/cgi/oai2?verb=Identify.

„Open Access-Policy der Universität Bern", 29. April 2013. http://www.ub.unibe.ch/openaccess/content/open_access_policy/index_ger.html.

„Open Archives Initiative Protocol for Metadata Harvesting". Zugegriffen 8. Februar 2015. http://www.openarchives.org/pmh/.

„Open Archives Initiative - Protocol for Metadata Harvesting - v.2.0". Zugegriffen 22. Februar 2015. http://www.openarchives.org/OAI/openarchivesprotocol.html#Record.

„Open Archives Initiative Service Providers". Zugegriffen 21. Februar 2015. http://www.openarchives.org/service/listproviders.html.

„Open Data Commons Attribution License". Zugegriffen 4. März 2015. http://opendefinition.org/licenses/odc-by/.

„Optimisation ~ Grow ~ Repositories Support Project". Zugegriffen 23. Februar 2015. http://www.rsp.ac.uk/grow/optimisation/.

„ORCID Statistics". Zugegriffen 4. März 2015. https://orcid.org/statistics.

„Schema Bib Extend Community Group". Zugegriffen 25. Februar 2015. https://www.w3.org/community/schemabibex/wiki/Main_Page.

„Schema.org". *Wikipedia, the Free Encyclopedia*, 17. Februar 2015. http://en.wikipedia.org/wiki/Schema.org.

„schema.org FAQ - Webmaster Tools Help". Zugegriffen 26. Februar 2015. https://support.google.com/webmasters/answer/1211158.

„ScholarSphere". Zugegriffen 27. Februar 2015. https://scholarsphere.psu.edu/.

„Simple Knowledge Organization System - Wikipedia, the free encyclopedia". Zugegriffen 7. März 2015. http://en.wikipedia.org/wiki/Simple_Knowledge_Organization_System.

„sitemaps.org". Zugegriffen 23. Februar 2015. http://www.sitemaps.org/de/.

„UB Wiki - Oeffentlich/BASE". Zugegriffen 21. Februar 2015. http://www.ub.uni-bielefeld.de/wiki/BASE%20.

„VIAF". Zugegriffen 4. März 2015. http://viaf.org/.