BFP, Vol. 36, pp. 369-376, Dezember 2012 • Copyright © by Walter de Gruyter • Berlin • Boston. DOI10.1515/bfp-2012-0050

PROJEKTE

Discovery jenseits von "all you can eat" und "one size fits all"

Roland Bertelmann
Leiter der Bibliothek des Wissenschaftsparks Albert Einstein
Deutsches GeoForschungsZentrum GFZ
Telegrafenberg
D-14473 Potsdam
E-Mail: rab@gfz-potsdam.de

Sascha Szott
Kooperativer Bibliotheksverbund Berlin-Brandenburg
Verbundzentrale
c/o Konrad-Zuse-Zuse Zentrum für Informationstechnik Berlin
Takustraße 7
D-14195 Berlin
E-Mail: szott@zib.de

Tobias Höhnow
Bibliothek des Wissenschaftsparks Albert Einstein
Deutsches GeoForschungsZentrum GFZ
Telegrafenberg
D-14473 Potsdam
E-Mail: hoehnow@gfz-potsdam.de

Auffindbarkeit, Zugänglichkeit, aber auch Filterung von wissenschaftlichen Informationen sind heute mehr denn je eine Grundvoraussetzung für erfolgreiches wissenschaftliches Arbeiten. Je besser eine Bibliothek solche Prämissen mit ihren Werkzeugen und Diensten ermöglicht und je näher sie dabei an den Bedürfnissen ihrer wissenschaftlichen Nutzer agiert, umso besser erfüllt sie ihre Aufgabe. Das innovative Suchportal ALBERT ist ein solches Werkzeug.
Schlüsselwörter: Discoverysystem; Suchportal; Informationsinfrastruktur

Discovery Beyond „all you can eat" and „one size fits all"

Discovery, accessibility, but also features to refine scientific information are nowadays elementary premises of scientific work. Tools and services of a library should meet these requirements, working close to scientists' needs. ALBERT – a cutting-edge search portal – is such a tool.
Keywords: Discovery; search portal; information infrastructure

1 Einleitung

In einem gemeinsamen Projekt der Bibliothek des Wissenschaftsparks Albert Einstein, Potsdam,[1] und der Verbundzentrale des Kooperativen Bibliotheksverbunds Berlin-Brandenburg (KOBV) wurde das Suchportal ALBERT[2] entwickelt, das für die Wissenschaftler auf dem größten außeruniversitären Forschungscampus im Land Brandenburg den zentralen Sucheinstieg in die für die geowissenschaftlichen Forschungsinstitutionen relevanten Ressourcen bietet. Die Anwendung wird inzwischen auf der Basis eines Hosting-Angebots des KOBV von ersten Bibliotheken nachgenutzt.[3]

Der folgende Werkstattbericht stellt die entwickelte Lösung ALBERT detailliert vor und verdeutlicht dabei an ausgewählten Beispielen, wie Benutzererwartungen schließlich innerhalb der technischen Lösung umgesetzt wurden und welche Herausforderungen dabei zu bewältigen waren. Neben der benutzerorientierten Sicht – orientiert an den heutigen Erwartungen an eine wissenschaftliche Spezialbibliothek und ihre Dienste – geben wir auch einen Einblick in die Bereiche des zugrundeliegenden Metadatenmanagements, einer Kernkompetenz von Bibliotheken, die bei ALBERT besonders zum Tragen kommt. So ermöglicht die Erhaltung der „Datenhoheit" flexibel anpassbare und pragmatische Lösungen entgegen dem Grundsatz von *one size fits all*, dem letztlich die großen am Markt befindlichen Discovery-Systeme folgen. Des Weiteren schafft ALBERT mit der konsequenten Fokussierung auf die spezifischen Bedürfnisse von wissenschaftlichen Spezial- und Forschungsbibliotheken ein Alleinstellungsmerkmal gegenüber Discovery-Lösungen, die vornehmlich als Katalogersatz (Stichwort *Next-Generation-Catalog*) fungieren. Die Fokussierung der indexierten Inhalte entgegen einem *all you can eat*-Ansatz erlaubt eine bedarfsgerechte und kostenoptimierte Versorgung, die Irrelevantes bewusst ignoriert und damit letztendlich auch innerhalb wirtschaftlicher Sparzwänge überhaupt eine Entscheidung für den Einsatz eines modernen Discovery-Systems ermöglicht.

1 Bibliothek des Wissenschaftsparks Albert Einstein, eine gemeinsame Bibliothek des Deutschen GeoForschungsZentrums GFZ, des Potsdam Instituts für Klimafolgenforschung, der Forschungsstelle Potsdam des Alfred Wegener Instituts für Polar- und Meeresforschung und des IASS Institute for Advanced Sustainability Studies. http://bib.gfz-potsdam.de.
2 http://waesearch.kobv.de.
3 http://albert.kobv.de.

2 Ein Zugangssystem orientiert am Bedarf vor Ort

Mit der Orientierung und Optimierung entlang lokal relevanter Quellen und unter maximaler Nutzung von offenen Metadaten und Volltexten einerseits und der Entwicklung von Funktionalitäten mit Blick auf den Mehrwert für wissenschaftliche Nutzer andererseits positioniert sich ALBERT in einem eigenen Feld. Der Ansatz konkurriert nicht mit den großen kommerziellen Discovery-Systemen oder den VuFind-Katalogen als „bessere“ Bibliothekskataloge. ALBERT versteht sich als Wissensportal, zugeschnitten auf die Bedürfnisse von Wissenschaftlern „vor Ort“.

In die Suche können verschiedenste Quellen eingebunden werden. ALBERT soll dabei als Zugangssystem, als erster Ort der Suche dienen. In vielen Fällen wird der Suchende bereits von ALBERT direkt mit genügend Informationen versorgt, nicht zuletzt in den Fällen, in denen frei zugängliche Volltexte indexiert sind. Konzept ist aber, nicht Funktionen anderer Systeme nachzubilden, sondern den Weg in diese weiterführenden Systeme zu ebnen. Das Beispiel Forschungsdaten zeigt dies besonders deutlich. In ALBERT ist nur ein elementarer Satz an Metadaten indexiert. Dies ermöglicht eine gemeinsame Suche nach Literatur und veröffentlichten Forschungsdaten, leitet den Suchenden dann aber in das jeweilige spezifische System weiter, aus dem die Metadaten stammen. Dort findet sich dann auch eine Vielfalt weiterer Informationen und regt u.U. zur Weitersuche im Zielsystem an.

Abb. 1: Erschlossene Inhalte am Beispiel der Anwendung in der Bibliothek des Wissenschaftsparks Albert Einstein

Dies beinhaltet u.a. Zeitschriften (lizenziert, gedruckt, Open Access, National- und Allianzlizenz), Bücher aus dem Bestand vor Ort, aber auch elektronische Bücher genauso wie Metadaten zu einschlägigen Aufsätzen und natürlich fachlich einschlägige Open-Access-Texte. Ergänzt wird die Sammlung durch themenbezogene Bibliografien, die in der Wissenschaft entstanden sind (etwa Literaturdokumentationen zu Projekten), sowie, zeitgemäß, disziplinspezifische publizierte Forschungsdaten.

Abb. 2: Mit der Auswahl der ‚Collection‘ wird die Suche verfeinert.

3 Integrierte Datenquellen

Welche Inhalte mithilfe des Werkzeugs ALBERT angeboten werden, ist natürlich eng verbunden mit den definierten Aufgaben der Bibliothek und ihrem spezifischen Umfeld, d.h. die Bibliothek entwickelt selbst das Portfolio der Inhaltsangebote. Da ALBERT im Kontext naturwissenschaftlicher Forschungsinstitute entstanden ist, stehen wissenschaftliche Zeitschriften und ihre Inhalte im Fokus. Die folgende Übersicht verdeutlicht das Spektrum am Beispiel der Bibliothek des Wissenschaftsparks Albert Einstein.

Inwieweit nachnutzende Bibliotheken andere Schwerpunkte setzen, wird sich zeigen. Voraussetzung ist lediglich die Möglichkeit, die definierten Inhaltsschnittstellen (XML-Formate, OAI-PMH) bedienen zu können.

Beim genaueren Blick auf ALBERT muss also differenziert werden zwischen der Suchoberfläche und der je nach Bibliothek differierenden Zusammenstellung der zugrundeliegenden Datenquellen.

3.1 Mashups

Verschiedene dieser Quellen werden bereits automatisiert im Vorfeld aufbereitet, damit sie für den Einsatz in ALBERT „reif“ sind und dort optimal genutzt werden können. Die dabei entwickelten Werkzeuge werden weiter unten

näher beschrieben und stehen Nachnutzern von ALBERT zur Verfügung.

Besonderen Wert wurde dabei auf den Umgang mit Zeitschriften, dem Kerninformationsbestand in den Naturwissenschaften, gelegt. So wird dem Wissenschaftler beispielsweise bei der Suche nach einer Zeitschrift ein einziger Datensatz angeboten, für den im Hintergrund alle Bestandsinformationen zusammengeführt wurden, egal ob die Zugänglichkeit für den Gesamt- oder einen Teilbestand über print lokal, via abgelaufenen oder laufenden Lizenzvertrag, Allianz- oder Nationallizenz, regionalem Konsortium oder als Open-Access-Zeitschrift geregelt ist. Da EZB und/oder ZDB bei den meisten Bibliotheken in der Zeitschriftenverwaltung eine zentrale Bedeutung zukommt, wird dies für ALBERT eingesetzt, und *Mashups* aus ZDB und EZB spielen u.a. zur Datenanreicherung eine wichtige Rolle.

Für die meisten laufend gehaltenen Zeitschriften werden darüber hinaus die aktuellen Metadaten zu den Aufsätzen angeboten. Jene Aufsätze also, die nach der *Peer Review* auf den Verlagsservern schon elektronisch verfügbar sind, aber erst nach Monaten im Druck erscheinen und in klassischen Fachdatenbanken auch erst dann ausgewertet werden. Auf diese Weise sind die aktuellen Aufsätze über ALBERT sofort für die Forschung verfügbar, d.h. die dramatische Lücke für die Recherche zwischen Publikation auf Verlagsservern und Nachweis in Fachdatenbanken (wie z.B. Web of Science) wird geschlossen.

Diese Aufsätze sind, genauso wie die Aufsätze aus den indexierten Metadaten der Nationallizenzen, wiederum über ein *Mashup* mit Informationen aus der EZB, inhaltlich erweitert und fachspezifisch klassifiziert und damit für Filtervorgänge vorbereitet. Unseres Wissens ist es das erste Mal, dass auf diese Weise die verfügbaren Metadaten der Nationallizenzen mit solchen Mehrwerten versehen wurden.

Zur Zeit sind mehr als 6 Millionen Dokumente indexiert, momentan wird mit Verlagen verhandelt, um weitere fachspezifisch interessante Artikel einzubinden. In nächster Zeit sollen weitere Nationallizenz-Pakete eingebunden werden (z.B. des Elsevier-Verlags).

Mit dieser Ausrichtung ist ALBERT nicht nur eine neue Technikanwendung, sondern versteht sich als innovativer Benutzerservice. Das positive Nutzerfeedback äußerst sich nicht nur in Gesprächen mit Wissenschaftlern, sondern schlägt sich auch deutlich in den Nutzungsstatistiken nieder.

4 Wissenschaftsnähe

Diese Rückmeldung basiert sicherlich darauf, dass Wissenschaftlerinnen und Wissenschaftler eine Vielzahl weiterer Mehrwerte zur Verfügung stehen. So können in ALBERT jegliche Suchanfragen per Feed abonniert werden, so dass der Nutzer automatisch über neu aufgenommene Dokumente informiert wird, die mit der Suchanfrage korrelieren. Forschende können sich aber auch Sammlungen „ihrer" Zeitschriften zusammenstellen und bekommen automatisch entsprechende Benachrichtigungen beim Erscheinen neuer Aufsätze. Diese Art der Filterung entlang persönlicher Schwerpunkte kommt der Arbeitsweise von Wissenschaftlern sehr entgegen. Über die Einbindung in die Oberfläche als *Journal Watch List* ist dieser Dienst einfach und unkompliziert einzusetzen. Diese Angebote ersetzen, bzw. erweitern traditionelle Zeitschrifteninhaltsdienste. So löste dies in der Bibliothek des Wissenschaftsparks Albert Einstein einen seit zehn Jahren mit hohem Aufwand betriebenen separaten Zeitschrifteninhaltsdienst ab.

Obwohl bei der Auswahl der angebotenen Inhalte der Nutzen für die Institution im Vordergrund steht, dürfen in Zeiten interdisziplinärer Forschung die Grenzen nicht zu eng gezogen werden. Nationallizenzen und offen zugängliche Quellen bieten vielfältiges Material, das im Zweifel doch für den Einzelnen interessant sein kann. Um diesen Zwiespalt zwischen Vorfilterung und Angebotsbreite zu umgehen, erlaubt ein weiteres Werkzeug dem einzelnen Wissenschaftler, den Suchraum zu erweitern oder ihn auf seine spezifischen Bedürfnisse einzuschränken. Er kann sich so den Suchraum auf der Basis einer Fachsystematik selbst zusammenstellen und über einen permanenten Link diese Vorauswahl in Zukunft zur Basis seiner Suchen machen. Voraussetzung für eine solche Funktionalität ist die bereits erwähnte Erweiterung der indexierten Quellen, etwa um die Fachsystematik der EZB.

Solcher Art von Personalisierung kommt in ALBERT eine wichtige Rolle zu. So können auch weitere Sucheinstellungen (z.B. Anzahl der Treffer pro Seite, Sortierkriterium) per Bookmark gespeichert werden. Erfolgt nun später der Aufruf über den abgelegten Link, so werden die persönlichen Sucheinstellungen übernommen, ohne dass erst eine Authentifizierung erforderlich ist. Gerade durch die Vorauswahl von EZB-Fachgebieten kann der Suchraum von vornherein erheblich begrenzt und dadurch die Übersichtlichkeit erhöht werden.

Das Web-Frontend selbst bietet sämtliche Funktionen, die an eine moderne Suchoberfläche gestellt werden. So bieten sich verschiedene Einstiege (einfache bzw. erweiterte Suche und Browsing). Autovervollständigung schon während der Eingabe von Suchanfragen oder eine kontextsensitive[4] Facettierung über die Suchtreffer bieten dem Benutzer die Möglichkeit, schnell die gewünschten Informationen aufzufinden. Verknüpfungen zwischen den Datenquellen (die bereits bei der Datenvorbereitung hergestellt wurden) werden genutzt, so z.B. zwischen den Zeitschriften und den dazugehörigen Artikeln. Die Anbindung von externen Systemen ist möglich. So werden Buchcover aus GoogleBooks[5] bezogen, kontextsensitive Weiterleitungen auf *Link Resolver* (wie z.B. SFX oder LinkSource) angeboten, der aktuelle Verfügbarkeitsstatus aus dem Bibliothekssystem abgefragt oder ein *Deeplink* in das Bibliothekssystem für die Bestellung von ausgeliehenen Medien angeboten. Bei der Entwicklung von ALBERT wurde konsequent darauf geachtet, Anknüpfungspunkte für die Integration solcher Dienste zu schaffen und bereits bestehende Services nicht redundant nachzubauen.

4 Bestimmte Facetten erscheinen erst im Kontext der Auswahl anderer Facetten, z.B. wird die Unterscheidung nach Zweigbibliothek oder printed/electronic books erst dem Benutzer zur Filterung angeboten, wenn dieser die Collection (Facette) „Books" ausgewählt hat.

5 https://developers.google.com/books/.

Abb. 3: Einbindung lokaler Katalogdaten

Die Applikation lässt sich selbst in andere Systeme (z.B. Content-Management-Systeme) integrieren, so dass ein in der Institution vorhandenes Corporate Design auch in der Suchoberfläche wiedergespiegelt werden kann und es aus Benutzersicht zu keinem Bruch zwischen den unterschiedlichen Bibliothekswebdiensten (z.B. dem Webauftritt und der Suchmaschine) kommt. Sämtliche URLs sind „stateful" im dem Sinne, dass der gesamte Kontext in ihnen codiert ist und sie damit später (auch nach abgelaufener Benutzersession) weiterhin aufgerufen werden können. Diese Eigenschaft ermöglicht auch Suchmaschinen-Crawlern die Indexierung, so dass auch über Websuchmaschinen ein Einstieg in die Bibliothekssuchmaschine erfolgen kann. Funktionen wie „Neuerwerbungsliste" und „Bestellliste" ermöglichen der Bibliothek die gezielte Hervorhebung von Medien.

Aufgrund der konsequenten Verwendung von „stateful" URLs können spezifische Sucheinstiege gebildet werden. Ein Sucheinstieg ist dabei durch eine Suchanfrage definiert und kann beliebige Kollektionen von Indexdokumenten umfassen. Jedes Indexdokument kann über einen Permalink referenziert werden. Damit kann beispielsweise eine Literaturliste aufgebaut werden, ohne die Metadaten erst umständlich übernehmen zu müssen. ALBERT stellt aber selbstverständlich auch verschiedene Mechanismen für den Export von Metadaten bereit (E-Mail, Datei-Download, Auszeichnung als Mikroformat COinS).

5 Hinter den Kulissen: Lucene/Solr

Die von der Bibliothek vorbereiteten Exportdateien werden z.B. auf einem FTP-Server für die weitere Verarbeitung zur Verfügung gestellt. Dabei können je nach Datenquelle unterschiedliche Updatezyklen vereinbart werden (z.B. soll der Katalogexport täglich; eine „langsame" Datenquelle mit nur wenigen monatlichen Änderungen aber nur einmal pro Monat im Suchindex aktualisiert werden). Der Indexer lädt die betreffenden Dateien zum definierten Zeitpunkt vom FTP-Server und beginnt mit der Verarbeitung. Aktuell werden für alle Datenquellen XML-basierte Austauschformate (z.B. MABXML[6] der DNB oder ein ALBERT-spezifisches XML) verwendet, so dass in einem ersten Schritt die Wohlgeformtheit und Schemakonformität überprüft werden kann. Im Falle von Fehlern werden automatisch E-Mails an die Verantwortlichen verschickt, so dass direkt im Nachgang eine Korrektur erfolgen kann. In solchen Fällen wird die betroffene Datenquelle bis zum nächsten Indexupdate nicht aktualisiert.

Nach der syntaktischen Prüfung werden die XML-Dateien ausgelesen und die Inhalte der einzelnen Elemente in die vorgesehenen Indexfelder geschrieben. Dabei findet teilweise auch noch eine Normalisierung oder Datenkorrektur statt. So werden beispielsweise ISSNs und ISBNs mit und ohne Bindestriche in den Index geschrieben, so dass der Benutzer in den Suchanfragen beide Varianten synonym verwenden kann. Datenkorrekturen umfassen z.B. den Umgang mit Umlauten (wenn diese in den Ausgangsdaten nicht korrekt codiert sind) bzw. mit Zusätzen, die im Suchindex bzw. der späteren Anzeige nicht erscheinen sollen (z.B. die nach MAB-Standard erlaubte Preisangabe im ISBN-Feld 540).

Neben den Exportdateien bietet ALBERT in der aktuellen Ausbaustufe auch die Möglichkeit, entfernte Quellen (z.B. Publikations- und Dokumentenserver sowie Repositories) über die OAI-PMH-Schnittstelle[7] zu indexieren. Hierbei wird in ALBERT lediglich die Basis-URL der Schnittstelle sowie ggf. ein OAI-Set[8] registriert. Da das verwendete Protokoll explizit die Möglichkeit inkrementeller Updates (mittels der Parameter *from* und *until*) vorsieht, werden hier nur die Änderungen seit dem letzten Update abgefragt. Eine Reindexierung der kompletten Datenquelle entfällt. Sind in den Datensätzen Direktverweise auf Volltexte angegeben, werden diese abgerufen, extrahiert und ebenfalls in den Suchindex aufgenommen. Werden lediglich Verweise auf die Übersichtsseiten der einzelnen Dokumente (*front doors*) angegeben, so erlaubt es ALBERT, auf diesen Seiten nach Volltext-URLs zu suchen (*screen scraping*) und diese abzurufen. Das Verhältnis von Indexdokumenten, zu denen Metadaten und Volltexte existieren, im Vergleich zu allen im Index abgespeicherten Dokumenten beträgt aktuell weniger als ein Prozent.

6 http://www.dnb.de/DE/Standardisierung/Formate/MABxml/mabxml_node.html.

7 http://www.openarchives.org/OAI/openarchivesprotocol.html.

8 http://www.openarchives.org/OAI/openarchivesprotocol.html#Set.

ALBERT ermöglicht durch das Definieren von Filtern die in einer Exportdatei enthaltenen Datensätze in verschiedene Indexdatenquellen („virtuelle Quellen") aufzuteilen. Dies ist z.B. sinnvoll, wenn innerhalb des Bibliothekskatalogs nicht nur gedruckte Bücher, sondern z.B. auch E-Books, AV-Medien oder Vortragsaufzeichnungen nachgewiesen werden und diese in der späteren Suchoberfläche für den Benutzer getrennt angezeigt und facettiert werden sollen. Die Filter können dabei programmatisch bis auf Feldebene definiert werden, so dass beliebig komplexe Kriterien abgebildet werden können.

Als Retrievaltechnologie nutzt ALBERT die etablierten Open-Source-Lösungen Lucene/Solr[9] der Apache Software Foundation. Der Solr-Index dient dabei als „Schnittstelle" zwischen Front- und Backend. In ihm werden sämtliche Metadaten der Dokumente für Retrieval und Anzeige abgelegt. Eine weitere Datenhaltung (z.B. in Form einer relationalen Datenbank) wird nicht benötigt. Damit Suche und Indexierung sich nicht gegenseitig beeinflussen, wird ein Solr-Server mit zwei Kernen verwendet (*multi core setup*), so dass Suche und Indexierung getrennt auf zwei unterschiedlichen Kernen erfolgen können. Erst nachdem die Indexierung erfolgreich beendet wurde, findet eine dynamische Umschaltung zwischen den beiden Kernen statt, so dass anschließend der gerade aktualisierte Kern in der Suchtur. Solche Analysen können auch in regelmäßigen Abständen wiederholt werden, so dass eine kontinuierliche Überwachung der Datenqualität möglich wird.

ALBERT ist aber bewusst sehr tolerant bezüglich der Datenqualität der zu indexierenden Datenquellen implementiert. Die aufgedeckten Probleme sind daher nur als Verbesserungshinweise zu verstehen, die aber nicht zwingend behoben werden müssen (hier ist im Einzelfall auch immer der Aufwand und Nutzen zu hinterfragen bzw. die personellen Kapazitäten der Bibliothek zu beachten). Hierbei zeigt sich nochmals, dass auch innerhalb des Hosting-Modells (s.u.) die „Datenhoheit" bei der Bibliothek verbleibt.

Die Anwendung stellt keine übermäßigen Anforderungen an die technische Infrastruktur und kann nicht zuletzt auch aus diesem Grund weiteren Bibliotheken (auch außerhalb des Verbunds) kostengünstig angeboten werden. Die KOBV-Verbundzentrale greift hierbei auf die IT-Infrastrukturdienste des Konrad-Zuse-Zentrums für Informationstechnik (ZIB) zurück. Ein internes und externes Monitoringsystem überwacht die Applikation und ermöglicht im Fehlerfall ein schnelles Eingreifen. Seit der Liveschaltung der ersten Version im Jahre 2007 hat sich die Applikation als sehr robust erwiesen, so dass wir im Durchschnitt eine Verfügbarkeit von 99,5 Prozent (d.h. weniger als zwei Ausfalltage pro Jahr) erlangen konnten.[10]

Abb. 4: Schematische Darstellung der Verarbeitungsstufen im Index-Backend

oberfläche zur Verfügung steht, ohne dass der Benutzer dies in Form eines Ausfalls bemerkt.

Im Rahmen des Aufbaus neuer Instanzen (siehe Hosting-Angebot) werden die bereitgestellten Exportdaten auch einer inhaltlichen Prüfung unterzogen (*data profiling*). So wird beispielsweise der Katalogexport detailliert analysiert, und es werden Vorschläge für eine Verbesserung der Datenqualität unterbreitet, die sich dann direkt in einer besseren Auffindbarkeit bzw. Präsentation niederschlagen. So können z.B. syntaktisch invalide ISBNs, ISSNs, URLs, URNs oder DOIs erkannt werden. Verstöße gegen Katalogisierungsregeln, die im Rahmen der Validierung, z.B. gegen das MABXML-Schema, so nicht aufgedeckt werden können, sind ebenfalls detektierbar. Solche Verstöße können z.B. fehlende Pflichtfelder (wie das MAB-Feld 331 für den Hauptsachtitel) oder MAB-Untersätze sein, die auf nicht im Export enthaltene Hauptsätze verweisen. Die Bibliothek entscheidet dann letztendlich über die Korrek-

6 Hosting-Angebot – Trennung der Verantwortung für Technik und Inhalt

Wie angesprochen gibt es die ersten Institutionen, die ALBERT einsetzen und ein entsprechendes Hosting-Angebot des KOBV nutzen. So werden die Zentralbibliothek des Deutschen Krebsforschungszentrums Heidelberg und die Bibliothek der Technischen Hochschule Wildau (Bibliothek des Jahres 2012) demnächst live gehen. Aus unserer Sicht wichtig und für den erfolgreichen Betrieb entscheidend ist die klare Trennung der Verantwortung. Die KOBV-Zentrale hostet den technischen Rahmen des Werkzeugs inklusive der definierten Quellenschnittstellen, die Bibliothek ist verantwortlich für die Auswahl, Aufbereitung und Bereitstellung

9 http://lucene.apache.org/solr/.

10 Durch den jährlich stattfindenden zentralen Wartungstag am Konrad-Zuse-Zentrum für Informationstechnik ergibt sich mindestens ein eintägiger, wenn auch geplanter, Ausfall der Anwendung im Jahr.

(sowie ggf. Korrektur) der Inhalte. Inhaltlicher Zuschnitt und Schwerpunktsetzung liegen also in der Hand der Bibliothek.

7 Hinter den Kulissen: Alles XML *oder* Wie kommen meine Zeitschriften überhaupt in ALBERT?

Jenseits des Einsatzes von *Electronic Resource Management-Systemen* (ERM) gibt es die Möglichkeit, den Inhalt externer Anbieter mit Standarddatenformaten wie XML zu verarbeiten[11].

Viele Bibliotheken in Deutschland verwalten ihre elektronischen Zeitschriften mithilfe der Elektronischen Zeitschriftenbibliothek (EZB) in Regensburg[12]. Ausgehend davon können die dort verwalteten bibliografischen Angaben sowie die Lizenzangaben einzelner Institutionen getrennt (manuell) heruntergeladen werden. Die Verarbeitung eines solchen Downloads kann wiederum auf vielfältige Weise vorgenommen werden, etwa als SQL-Dump, um diesen direkt in eine Datenbank einzulesen, welche die Datenbasis eines ERM-Systems bildet.

Die progressivere Methode ist das automatisierte Abholen und Weiterverarbeiten der Daten aus der hier als Wissensbasis fungierenden EZB. Als Mehrwertdienst sowie als Vorstufe für die Integration der Zeitschriftenbestände in ALBERT wird dies mittels einer XSLT-Anwendung (*Extensible Stylesheet Language Transformations*) vorgenommen. Die Applikation besteht aus mehreren XSLT-Stylesheets mit aufeinander aufbauenden mehrstufigen Prozessen und ist im Rahmen einer Bachelorarbeit[13] an der Bibliothek des Wissenschaftsparks Albert Einstein entstanden. Basierend auf der XML-Schnittstelle der EZB[14], die sonst zur Integration der Zeitschriftenbibliothek in das eigene Umfeld bzw. CMS genutzt wird, lädt die XSLT-Anwendung die entsprechenden Daten herunter. Zur Spezifizierung des Downloads lässt sich die XSLT-Anwendung parametrisieren. Als Attribute können die Bibliothek (EZB-ID), die „Ampelfarbe“ oder das Fachgebiet angegeben werden. Je nach Bedarf ist so das differenzierte Abholen der Daten, z.B. aller „gelben“ (d.h. lizenzierten) Titel eines Fachgebiets, realisierbar, bis hin zum vollständigen Download der EZB. Dass hier, im Gegensatz zum manuellen Herunterladen, zusätzliche Felder wie die frei vergebenen Schlagwörter berücksichtigt werden, wird als positiver Nebeneffekt gerne mitgenommen.

Die vielfältigen Lizensierungsmöglichkeiten wie Nationallizenzen, Allianzlizenzen und Konsortien bringen für viele Bibliotheken einander überschneidende, ergänzende oder komplettierende Bestandszeiträume für einzelne Zeitschriften hervor. Nicht selten ergibt sich aus der Kombination des Archivs der Nationallizenzen und dem laufenden Bezug via Konsortium oder Allianzlizenz Zugriff auf den gesamten Bestand einer Zeitschrift. Die Herkunft des Bestands wird in der EZB separat abgebildet und ist für den Bibliothekar wissenswert, für den Nutzer jedoch irrelevant. Er möchte ohne tiefergehende Kenntnis diverser Lizenzen lediglich wissen, was verfügbar ist und was nicht.

Die XSLT-Anwendung, die nicht als bibliothekarisches Backend agiert, um separate Zugänge zu verwalten, sondern direkt nach ALBERT transformiert, maximalisiert solche Bestandszeiträume und stellt den Bezugszeitraum in einer konzentrierten Einheit dar.[15] Nicht nur bei Extrembeispielen kann so eine gewisse Übersichtlichkeit hergestellt werden.

Verwaltet eine Bibliothek ihre Bestandsangaben gedruckter Zeitschriften in der Zeitschriftendatenbank (ZDB), können über die XSLT-Anwendung zusätzlich zu den elektronischen auch die gedruckten Bestände integriert werden. Die Realisierung erfolgt über die gemeinsame Verfügbarkeitsrecherche von EZB und ZDB[16], einem XML-Webservice. Diese Option ist als Parameter in der XSLT-Anwendung hinterlegt, wobei als Attribut das Sigel der jeweiligen Bibliothek benutzt wird.

Ist die ZDB-Option gesetzt, generiert die XSLT-Anwendung nach Ausführung eine ALBERT-konforme XML-Datei, die die elektronischen und gedruckten Bestände in einer Titelanzeige zusammenführt. Als Updatefrequenz hat sich für die EZB ein wöchentlicher Rhythmus als praktikabel erwiesen. Die Option des Integrierens der Print-Zeitschriften aus der ZDB dürfte, aufgrund der selteneren Bewegung der Bestände dort, mit größeren Abständen ausreichend sein.

Neben den vielfältigen Zugängen über Lizenzen ist in den letzten Jahren vor allem der Bereich der freien („grünen“) Zeitschriften stark gewachsen. Oftmals unterliegen diese Zeitschriften einem Embargo (z.B. „älter als 12 Monate“), treten sonst aber als ergänzender Inhalt zum ggf. lizensierten Inhalt auf. Nicht selten existieren dadurch für ein und denselben Titel zwei Einträge mit verschiedenen Zugangszeiträumen, die aber auf dieselbe Plattform führen bzw. denselben Inhaltsanbieter. Um nicht gänzlich auf die „grünen“ Zeitschriften zu verzichten, entstehen bei dieser Konstellation somit ungewollte Dubletten. Diesem Umstand kann begegnet werden, indem die in der XSLT-Anwendung hinterlegte Option des Zusammenführens „grüner“ und „gelber“ Zeitschriften ausgewählt wird. Das Aggregieren erfolgt einerseits unter der Bedingung des Vorliegens derselben ZDB-ID, so dass die zusammenzuführenden Zeitschriften eindeutig detektierbar sind und das Vermischen unterschiedlicher Zeitschriften ausgeschlossen ist. Andererseits muss

11 Höhnow, Tobias: Suchmaschine, ERM & Co.: Ressourcenmanagement im Backend des Bibliothekars. In: eLibrary – den Wandel gestalten: 5. Konferenz der Zentralbibliothek, Forschungszentrum Jülich, 8.–10. November 2010. Jülich 2010, S. 163-176.

12 Elektronische Zeitschriftenbibliothek Jahresbericht 2010. Regensburg: Universitätsbibliothek Regensburg, 2011. http://ezb.uni-regensburg.de/anwender/Jahresbericht_EZB_2010.pdf.

13 Nolte, Linda: XSL Transformationen als Vorbereitung für die Bibliothekssuchmaschine ALBERT. Bachelorarbeit im Studiengang Bibliotheksmanagement an der Fachhochschule Potsdam. Potsdam 2012.

14 XML-Ausgabeformat der EZB. Version 0.2. Regensburg : Universitätsbibliothek Regensburg, 2005. URL: http://ezb.uni-regensburg.de/anwender/info_XML.htm.

15 Beispiel: aus „1.1939–159.2002“ und „149.1997 – “ wird somit „1.1939 –“

16 Journals Online & Print. Gemeinsame Verfügbarkeitsrecherche von ZDB und EZB. Berlin : Staatsbibliothek zu Berlin, 2011. http://www.zeitschriftendatenbank.de/services/journals-online-print/.

dieselbe Zugangs-URL[17] angegeben sein, letzteres um auszuschließen, dass Archive bzw. URLs, die zu anderen Plattformen führen (z.B. JSTOR)[18], nicht abhanden kommen.

Häufig besteht der Bedarf, weitere qualitätsverbessernde Angaben zu den Standardangaben der EZB hinzuzufügen. Zu nennen wären da etwa zusätzliche Titelangaben, Abkürzungen, Akronyme, vorherige oder nachfolgende Titelangaben, Körperschaften, Schlagwörter, Informationen über ein vorliegendes Nationallizenz-Archiv oder die Bild-URL des Titelblatts einer Zeitschrift. Diese Angaben können gemäß eines Schemas in einer separaten XML-Datei hinterlegt werden und fließen beim Ausführen der XSLT-Anwendung in den EZB-Export ein. Besonders hervorzuheben ist die Möglichkeit des Zusammenstellens spezifischer auf den eigenen Bedarf abgestimmter Kollektionen von Zeitschriften über eindeutige Suchterme bzw. *Keywords*. Beispielsweise für Forschungsgruppen kann so eine Sammlung thematisch enger „Lieblingszeitschriften" problemlos und ohne viel Aufwand zusammengestellt werden.

Die generierte ALBERT-konforme XML-Datei, die nun sämtliche Zeitschriftenangaben einer Bibliothek enthält, kann gegen ein eigens dafür entwickeltes XML-Schema (XSD) validiert werden, um die Verarbeitung der Datei durch den ALBERT-Parser zu standardisieren und Fehlerquellen zu vermeiden. Die Erfahrungen zeigen aber eine problemlose Verarbeitung der Daten und die Produktion von durchweg validem und wohlgeformtem XML.

Nach Einrichten eines entsprechenden Cronjobs agiert die XSLT-Anwendung vollkommen autark, angefangen beim Herunterladen des Inhalts über die Transformation, dem Hinzufügen bibliografischer Informationen, bis hin zum Bereitstellen des fertigen Zeitschriften-XMLs für den ALBERT-Indexer.

8 Aktueller geht es nicht: Integration neuester Artikel direkt von der Quelle

Die Fülle und ständige Verfügbarkeit elektronischer Zeitschriften vor allem in den Naturwissenschaften bringt es mit sich, dass diese als primäre Informationsquellen für Wissenschaftler gelten. Neben diesem Aspekt steht der Zugriff auf Artikelebene über Aggregatoren wie Google Scholar oder PubMed immer mehr im Vordergrund[19]. Um diese Entwicklung aufzunehmen und die Lücke zwischen der Online-Publikation (*ahead of print*, *early first*, AOP etc.) und der Aufnahme in etablierten bibliografischen Datenbanken zu schließen, wird der aktuelle abonnierte sowie freie Bestand an Zeitschriftenartikeln per RSS-Feed (von den einzelnen Verlagsservern) in die Suchmaschine ALBERT integriert. Bedingung dafür ist die laufende Verfügbarkeit der Zeitschrift und ein von der Plattform angebotener RSS-Feed.

Die XSLT-Anwendung, die zu diesem Zweck um ein weiteres Modul (XSLT-Stylesheet) erweitert wurde, bezieht ihre Informationen über jene Zeitschriften, die integriert werden sollen, aus einer OPML-Datei (*Outline Processor Markup Language*), welche die Feed-Adressen beinhaltet. Diese Auswahl kann mithilfe des RSS-Aggregators JournalTOCs[20] getroffen werden, einem Service der Heriot-Watt University Edinburgh[21], der die RSS-Feeds von nahezu 20.000 Zeitschriften erfasst. Soweit keine weiteren Feeds hinzukommen sollen, ist dies ein einmaliger Arbeitsschritt. Allerdings kann auch jederzeit eine überarbeitete Liste von Zeitschriften aus JournalTOCs generiert werden.

Auf Basis der erzeugten OPML-Datei fragt die XSLT-Anwendung sämtliche definierte Zeitschriften auf neue Artikel bei JournalTOCs ab[22] und exportiert diese in eine XML-Datei (RSS-XML). Zusätzlich wird das RSS-XML mit Informationen aus der Zeitschriften-XML angereichert, beispielsweise mit ISSNs sowie den Themengebieten der EZB.

Die XSLT-Anwendung kann explizit auf das Abholen der RSS-Feeds eingerichtet werden, so dass eine differenzierte

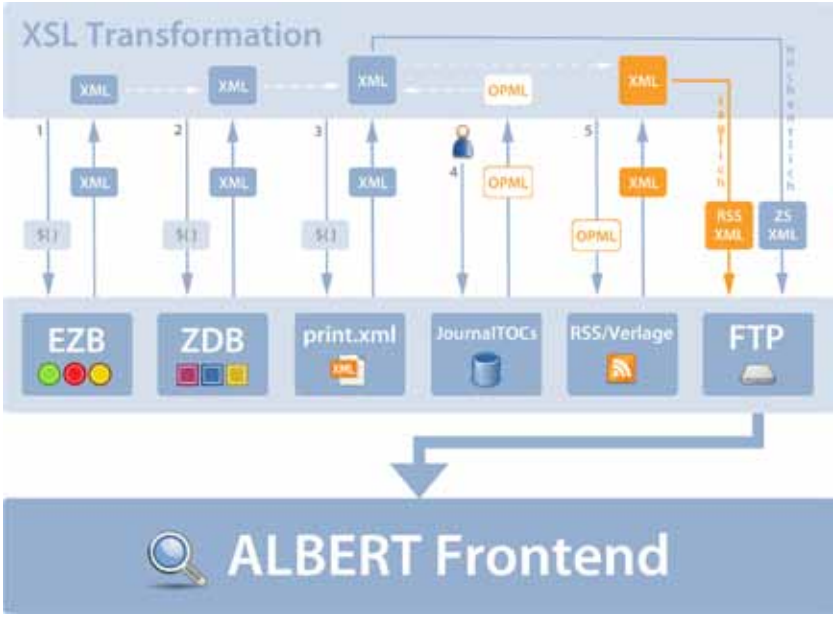

Abb. 5: Metadatenaggregation und Transformation im Vorfeld der Indexierung in ALBERT

17 Beispiel „Journal of Climate", vgl. EZB-IDs 5917 und 61984.
18 Beispiel „Ecology", vgl. EZB-IDs 6124 und 7568.
19 Nicholas, D., Williams, P., Rowlands, I., Jamali, H.R.: Researchers' e-journal use and information seeking behaviour. In: Journal of Information Science 36 (2010) S. 494-516.
20 http://www.journaltocs.hw.ac.uk/.
21 http://www.hw.ac.uk/.
22 Die Artikel werden aufgrund der einheitlichen Darstellung und somit einfacheren Verarbeitung bei JournalTOCs abgefragt, da die einzelnen Verlage zum Teil qualitativ sehr unterschiedliche Daten liefern. Unter dieser Praxis leidet aber geringfügig die Aktualität des Inhalts, weil JournalTOCs erfahrungsgemäß nicht täglich alle 20.000 Feeds bei den Verlagen abfragt. Obwohl die Inhalte von der XSLT-Anwendung auch direkt bei den Verlagen abgefangen werden können, haben wir uns für die Qualität entschieden.

Updatefrequenz im Gegensatz zum Abholen der Zeitschriften aus der EZB eingestellt werden kann. Dies empfiehlt sich im Bereich der RSS-Feeds deshalb, weil so eine tagesaktuelle Bereitstellung der neuesten Artikel gewährleistet werden kann.

Falls fehlerhafte XML-Dateien in den gelieferten Daten enthalten sind (was durchaus vorkommt, z.B. bei Mängeln in der Codierung), ist in der XSLT-Anwendung eine Schleife zur Reparatur des XMLs und Wiederholung des Prozesses eingebaut, so dass keine Artikel verloren gehen.

9 Fazit und Ausblick

ALBERT hat sich als Werkzeug zur Profilierung der Bibliothek als wissenschaftszugewandter Dienstleister bewährt. Die Möglichkeit, Inhalte in der Verantwortung der Bibliothek zu gestalten, eröffnet den Freiraum, wissenschaftsnah zu agieren. Letztlich positioniert sich ALBERT damit als Modul in einer modernen Informationsinfrastruktur.

Mit der Übernahme durch weitere Bibliotheken im Rahmen des Hosting-Angebots entsteht eine ALBERT-Community, die das Werkzeug gemeinsam weiterentwickeln wird. Eine Vielzahl von Ausbaumöglichkeiten liegt bereits auf der Hand. Neben der Integration weiterer Metadatenpakete aus den Nationallizenzen bietet sich der Ausbau der Personalisierungsmöglichkeiten an. Darüber hinaus bleibt zu überlegen, ob die Portalfunktion nicht durch die Integration von nichtbibliografischen Quellen gestärkt werden kann. Lokale wissenschaftliche Inhaltsbeschreibungen wie Projektübersichten und etwa Lehrangebote bieten sich dabei an.

Die hier vorgestellte Version 2 von ALBERT hat noch längst nicht alle Möglichkeiten des Ansatzes dieses Suchportals ausgeschöpft.