Vom Katalog zur Bibliothek: Zwischenschritt und Zwischenstand „Kataloganreicherung“

Manfred Hauer, Reiner Diedrichs

Bibliotheken als Information-Retrieval-Systeme

Bibliotheken sind nicht chaotische Sammlungen von Medien, sondern sie werden stets als Information-Retrieval-Systeme implementiert, denn das Wiederfinden bei exakten und vagen Anfragen ist von Anfang an das Sammlungsziel. Niemand kennt zum Zeitpunkt der Sammlung die Fragestellung, welche einen zukünftigen Benutzer zu diesem Medium führen soll. Die Art der Sammlungsorganisation bestimmt in hohem Maße die zukünftig mögliche Antwortmenge und deren informatorische Qualität.

Viele Sammler kennen die einzelnen Medien recht gut durch eigene Lektüre oder Nutzung, durch Verarbeitung der Inhalte in eigenen Schriften, durch Kenntnis der Autoren oder durch Einschätzungen, Empfehlungen, Meinungen von anderen. Diese Sammler sind „Antwortmaschinen“ – können oft mit hoher Präzision und angepasst auf den Kenntnisstand des Fragenden Erklärungen geben, welche aus einer Summe von Medieninhalten gelernt wurden. Genau diesen Typ wünschen sich die meisten Benutzer, führt er doch schnell, fachlich ausgewogen und verständlich ans Ziel.

Der nicht ganz so begabte Sammler oder Sammler deutlich größerer Medienmengen kann nur mehr oder weniger gut auf einzelne Medien oder Mediengruppen hinführen, gibt aber keine fachliche Auskunft mehr – sondern liefert nur Hinweise auf mögliche „Antwort-Container“. Zumindest seit der berühmten Bibliothek von Alexandria hilft dieser Sammler seinem eigenen Gedächtnis mit kurzen Notizen nach: Listen von Titeln, von Autoren, von Themen und Referenzen auf den Standort. Praktischer als geschriebene Listen sind wegen der leichteren Sortierbarkeit Karteikarten. Der digitale *Record* in den Datenbanksystemen der Bibliotheken ist logisch nichts anderes, nur deutlich schneller sortierbar.

Als sich in den 1970er Jahren die heutigen Bibliothekssysteme entwickelten, entschieden sich fast alle Anbieter für zumeist relationale Datenbank-Management-Systeme als Basis-Technologie – gut bewährt in Lagerverwaltung und Buchhaltung, in Wirtschaft und Verwaltung, also in Welten mit sehr wenig Textinformation. Vage Suche war im Ansatz nicht vorgesehen, ging es doch zunächst nur um digitale „Karteikarten“.

Digitalisierung, Speichersysteme, Virtualisierung und Weiterentwicklungen der Datenbank-Management-Systeme erlauben heute aber, dass Katalog und Medium technisch zusammenfallen.

Damals in den 1970er Jahren fristeten die Experten für Information Retrieval, ein weiteres Forschungsfeld der Informatik, noch ein Schattendasein. Als im World-Wide-Web endlich viel Text verfügbar war, traten sie ab Mitte der 1990er Jahre aus dem Schatten, zeigten ihre Stärken. Google ist wohl heute weltweit vielleicht die bekannteste Marke nach Coca Cola, auch wenn Google hier noch zur Kategorie des „weniger begabten Sammlers" zählt.

Kataloganreicherung, wie seit 2002 von der Gruppe um dandelon.com betrieben, versucht Information-Retrieval-Technologie mit den relationalen Datenbanksystemen der Bibliotheken zusammenzubringen. Dazu werden mehr Daten benötigt als die bisherigen bibliothekarischen Titelbeschreibungen hergeben. Sprachverarbeitungskonzepte sind notwendig, um die Vielfalt der Sprache wieder einzufangen und neue Konzepte für die Anzeige dieser Texte. Die alten Konzepte der inhaltlichen Gruppierung – hierarchische Bäume (Systematiken) bzw. polyhierarchisch und assoziativ vernetzten Schlagworte – können für die Suche hinterlegt oder für die Anzeigesortierung genutzt werden.

Kataloganreicherung ist die derzeit technisch sinnvolle Voraussetzung zum Einsatz von moderner Information-Retrieval-Technologie in Bibliotheken. Deren Einsatz ist bei Bibliotheksverbünden auf dem Vormarsch; die Einzelbibliothek ist darin zunehmend nur noch eine virtuelle Sicht auf weit größere Datenkollektionen.

Die maschinelle Indexierung – mit linguistischen und/oder statistischen Methoden – ist ein mögliches Verfahren in Information-Retrieval-Systemen, dessen Resultate auch direkt in OPAC-Systemen nachgenutzt werden kann und damit auch dort eine Recherche auf breiterer terminologischer Basis gestattet. Diese Erweiterung des OPACs um die maschinellen Indexierungsergebnisse stand bei der Vorarlberger Landesbibliothek, dem Pionier unter den „Kataloganreichern" und wohl noch immer größten Einzel-Produzenten, schon 2002 vor der Anzeige der Inhaltsverzeichnisse im Vordergrund. Alle dandelon.com-Bibliotheken übernehmen die maschinelle Indexierung in ihre Kataloge. Schon bald zeigte sich aber im jeweiligen Bibliothekssystem das fehlende Ranking. 2004 startete deshalb „dandelon.com". Die maschinelle Indexierung wird in homöopathischer Dosis in den HEBIS-Katalog übernommen. Für die DNB hat Frau Dr. Niggemann auf dem Bibliothekartag 2009 in Erfurt die maschinelle Indexierung, Klassifizierung, Extraktion weiterer Metadaten und die Ergänzung um zusätzliche Daten – auch abweichend von bisherigen Normdateien – als Projekt angekündigt und deren Übernahme in den Katalog ist ein Meilenstein. Die DNB kündigt damit den Perspektivenwechsel der bibliothekarischen Sicht hin zur Sicht des Endbenutzers an. Dieser will und benötigt im internationalen Wettbewerb reiche, nützliche Kataloge bzw. Bibliotheken!

Seit 2003 entwickelten das BSZ, das HBZ, der Bayerische Verbund, der Österreichische Bibliothekenverbund und die ETH Zürich sowie drei weitere Anbieter von

Scannern jeweils eigene Lösungen für die Digitalisierung und Texterkennung (OCR) von Inhaltsverzeichnissen. BSZ, ÖBV und ETH befassten sich mit maschinellen Erschließungsverfahren. 2007 einigten sich die deutschsprachigen Verbünde (ohne die Deutsch-Schweiz) auf grobe Standards für PDFs (Image im Vordergrund, Text dahinter) und seit 2008/2009 tauschen sie Daten untereinander aus. In den USA begann die ISBN-Agentur Bowker („Syndetics Solutions" ist seit 2004 Teil von Bowker) von allen Verlagen für neue Titel ein Inhaltsverzeichnis einzufordern, sichtbar bei der Library of Congress und bei einigen US-Bibliotheken, sowie in Deutschland u.a. beim GBV. Zusätzlich zieht Bowker/Syndetics große Mengen von Metadaten aus Library Thing (siehe Tabelle) und arbeitet mit weiteren Content Providern zusammen, z.B. Libri in Deutschland.

Was alles ist „Kataloganreicherung"?

Einfach gesagt, alles was über die bisherigen Katalogregeln und „Karten" hinausgeht und der Suche, Navigation und Information des Benutzers dient. Es geht nicht nur um Inhaltsverzeichnisse. Doch sie sind wichtig für alle drei Funktionen.

Kataloganreicherungen stammen nach unserem Verständnis aus drei Quellen:

1. die Integration von *Originaldaten* aus dem jeweiligen Medium: Cover Page, Klappentext, Inhaltsverzeichnis, andere Verzeichnisse, Volltexte, Link auf Volltexte oder auf Non-Text-Dokumenttypen, unselbstständige Werken, also Aufsätze, Artikel aus Konferenzbänden, Festschriften und natürlich Zeitschriften, aber auch aus Zeitungen und digitalen Quellen jeder Art.
2. *Erschließungsdaten:* von Menschen vergebene Deskriptoren, insbesondere von Verlagen, Bibliothekaren oder durch Leser (Social Tagging), von Programmen aus Originaldaten errechneten maschinellen Extraktionen möglichst mit Relevanzgewichtung bis hin zur Zusammenfassung (Mensch oder Maschine). Diese Daten hängen jeweils am einzelnen Datensatz.
3. *Ergänzende Daten:* Query-Expansion auf Basis von semantischen Konstrukten wie Thesaurus oder statistisch assoziierten Wortlisten oder maschinelle Recommender-Systeme auf Basis statistischer Analyse von massenhaftem Benutzerverhalten, aus den Suchergebnissen herausgerechnete Navigationen wie WordClouds, semantische Wolken sowie Sortierungen, teils mit Visualisierung von Suchergebnisgruppen und schließlich die große Gruppe des Crosslinking zu sachlich verwandten Daten wie Rezensionen, Autorenprofilen, Wortdefinitionen, Dokumenten auf Volltext-Servern, eBooks auf Verlagsservern, zu Rohdaten-Speichern und vieles mehr.
4. *Schnittstellen und Programme:* Hierzu zählen Link-Techniken und Bookmark-Services, dann Download-Schnittstellen für lokale oder andere Literaturverwaltung und schließlich Programme zur Verarbeitung, sprich Analyse, Manipulation und Kombination von Texten, Tabellen, Graphen und Rohdaten.

Quantitativer Zwischenstand

Auf dem Deutschen Bibliothekartag im Juni 2009 in Erfurt gab Christof Mainberger (BSZ) einen Überblick über die Menge der bisher produzierten und verfügbaren Inhaltsverzeichnisse. Die Produktion findet stets in und durch die Bibliotheken oder dort engagierten Dienstleistern statt. Die Verbünde sind meist zentrale Sammelstellen. Wir ergänzen seine Zahlen um eigene Erhebungen und Daten. Da sich die produzierenden Bibliotheken außerhalb des dandelon.com-Netzwerks innerhalb der Verbünde und zwischen den Verbünden bislang insbesondere bei der rückwärtigen Erschließung nicht pro Medium, sondern grob über Sachgebiete abstimmen, muss man von einer nicht unerheblichen Mengen an Dubletten ausgehen. So schätzen wir z.B. in Bezug auf die Reihe A der Deutschen Bibliothek – seit 2008 wird dort gescannt – eine Redundanz von 50% mit dem Netzwerk von dandelon.com.

Um Dubletten zwischen den Bibliotheken innerhalb von Verbundgrenzen und zwischen Verbünden zu reduzieren, wurden über die Verbünde Sammelgebiete abgesprochen. Dennoch rechnet Reiner Diedrichs, Direktor der Verbundzentrale des GBV (VZG) mit 6% Dubletten, da viele Titel nicht eindeutig genug abgrenzbar sind. Nur das Netzwerk dandelon.com stimmt sich automatisch pro Titel ab und erzeugt bei identischer ISBN und Jahr keine Dublette. Somit ist die Gesamtzahl aller verschiedenen von Bibliotheken produzierten Inhaltsverzeichnisse schwer zu schätzen, ca. 1 Million vermutet Reiner Diedrichs. Da allein mit intelligentCAPTURE bis Anfang Juli 2009 bereits 750.000 Inhaltsverzeichnisse produziert wurden und hier der monatliche Zuwachs seit 2 Jahren bei 20.000 Titeln liegt, ist die tatsächliche Zahl aller Produzenten eher bei 1,2 oder 1,3 Millionen, vermutet Manfred Hauer.

In vier Bibliotheken von Max-Planck-Instituten werden Inhaltsverzeichnisse produziert, ebenso in der Universitätsbibliothek Bologna. Diese Daten werden nicht ausgetauscht und wir schätzen die Gesamtmenge auf grob 100.000 Titel.

Im Durchschnitt werden in wissenschaftlichen Bibliotheken Bücher 4,5 mal ausgeliehen, vorwiegend in den ersten 2 bis 5 Jahren nach Erscheinen. Mit der Möglichkeit der zusätzlichen maschinellen Indexierung und/oder der Volltexte der PDF-Dateien der Inhaltsverzeichnisse steigt die Nutzung bei älteren Werken deutlich an, wie die DNB jetzt feststellt, seit sie teils bis zu 200 Jahre alte Werke (erworben ab 1913) mit den Inhaltsverzeichnissen online suchbar macht.

Bei aktuellen Titeln ist die Abdeckung teils recht gut, im GBV sind z.B. 35% der Monografien und Mikroformen der letzten 10 Erscheinungsjahre mit einem Inhaltsverzeichnis erschlossen. Auf seinen internen Datenpool mit PDFs konnte der GBV 2008 366.000 Zugriffe pro Monat registrieren. 2009 stieg die Zugriffzahl bereits auf 622.000 pro Monat (Zuwachs von 70%). Für 2009 werden ca. 7,5 Millionen Zugriffe erwartet. Die Nutzung wächst weiter durch die Einspielung von Links

und/oder Daten aus dandelon.com, anderen Verbünden, Library of Congress und Casalini Libri. Immer mehr Benutzer klicken immer öfter auf das zusätzlich angebotene Inhaltsverzeichnis. Es ist das am stärksten genutzte Zusatzangebot laut Reiner Diedrichs.

Das freie, internationale kollaborative Netzwerk dandelon.com zählte 750.000 Einzelfragen in 2008, 80% mehr als in 2007. Dabei wurden durchschnittlich 2,5 Inhaltsverzeichnisse geöffnet. Zusätzlich greift der GBV auf das externe dandelon.com im Hintergrund direkt auf die PDFs zu, statistisch nicht ausgewertet, mehrere hundertausend mal pro Monat.

Kataloganreicherung senkt die Total Cost of Ownership

Während ein Stellplatz im Regal ca. 3 Euro pro Buch und ein Buch über seinen gesamten Lebenszyklus ca. 138 Euro kostet, sind die Kosten für die Kataloganreicherung mit ca. 1,50 Euro doch sehr gering. Bei Austausch über viele Bibliotheken und Verbünde fallen die Kosten und der Aufwand praktisch nicht mehr ins Gewicht. Die bessere inhaltliche Erschließung führt darüber hinaus zu Kosteneinsparungen durch nicht getätigte Ausleihen (nur Holen, Durchblättern, Zurückgeben) und Fernleihen (sehr niedrige Leserate durch die Benutzer), wobei eine Fernleihe mit Kosten von über 30 Euro pro Buch veranschlagt werden muss. Die, bisher nicht gemessen, aber mit Abstand größte Zeit- und Kosteneinsparung liegt bei den Benutzern der Bibliotheken – und damit volkswirtschaftlich der Nutzen bei uns allen.

Während die Bibliotheken im dandelon.com-Netzwerk schon immer ohne Kostenverrechnung und vollautomatisch ihre Inhaltsverzeichnisse getauscht und auf Dubletten kontrolliert haben, gab es unter den Verbünden nach dem ersten Großprojekt beim HBZ die Idee, im Verhältnis 1:1 zu tauschen. Der Aufwand zur Kontrolle wäre höher als der Nutzen gewesen, es wird inzwischen ohne Einschränkung auf Gegenseitigkeit getauscht. Nur die DNB berechnet pro Verbund 1075 Euro pro Jahr als Bearbeitungskosten für die Lieferungen. Ob diese kostenpflichtigen und durch Logo und URL stets nur auf die DNB zurückweisenden Daten so angenommen werden, muss sich noch zeigen.

Make or Buy?

Ganz ohne Eigenleistung wird keine Bibliothek eine komplette Anreicherung ihrer Bestände erreichen. Alte oder ausländische Schriften ohne ISBN, Noten, Schallplatten-, CDs- oder DVD-Covers, graue Literatur, regionale Zeitschriften und Jahresbände, Abschlussarbeiten gibt es bislang nicht auf dem Tauschmarkt; mangels eindeutiger Identifikationsschlüssel wie ISBN oder DOI wird dies auch mittelfristig nicht zu erreichen sein. Bei der gedruckten Literatur wird es bei Büchern noch lange große Lücken geben und noch länger bei Aufsätzen in Zeitschriften und

	Orignaldaten				Metadaten		
Anbieter	**TOCs**	**Titelblatt intern**	**Cover Pages**	**Lese-probe**	**Maschinelle Indexierung**	**Abstract u.ä.**	**Rezensionen**
Dandelon.com	580.000	Ja	39.000		Ja	50.000	
HBZ	400.000 (teilweise TOCs)	Nein			Nein		
GBV	352.000 *	Ja	Cover von Nilson Book-data	35.000	Teilweise beabsichtigt	3.100	
BSZ	128.645	Nein	53.000	10.000	Nein	2.200	4.700
BVB	150.000 (teilweise TOCs)	Nein			Nein		
HEBIS	75.000 *	Ja			Ja		
KOBV	18.000	Ja			Ja		
OBV	55.000	Nein			Nein		
IDS	30.000 * 103.000 11.000	Ja			Ja		
DNB ab 2008	90.000	Nein			Beabsichtigt		
DNB ab 1913	35.000	Nein			Beabsichtigt		
Casalini Libri, Italien	70.000	Nein			Nein		
Bowker/Syndetics	900.000	Nein	4,9 Mio, vor-wiegend US	225.000		2.3 Mio	900.000

*Vollständig oder weitgehend in dandelon.com enthalten.

Weitere Daten				
	Produktionssystem	**Bemerkungen**	**Fachgebiete**	**Suchsystem**
19.500 eBooks (2.700 Videos geplant)	intelligentCAPTURE	Offener internationaler Verbund, bisher D, A, CH, LI, N	alle, zahlreiche Sondersammelgebiete – aus GBV, HEBIS u.a.	Ranking, Query-Expansion Highlighting aller Suchterme auch im PDF
	Adam	davon 180.000 Projektfinanziert Nutzt zusätzlich DNB, GBV,+?	Medizin, Ernährung, Agrar, Wirtschaft, Mathematik, Romanistik, Kunst, Sozialwissenschaft	Boolean
FAZ Rezensionen	intelligentCAPTURE überwiegend und C3	Davon ca. 225.000 verbundfinanziert Nutzt zusätzlich Dandelon.com, Bowker(LOC) und Casalini Libri Sowie DNB, HBZ,HEBIS, BSZ; OBV, BVB in VorbereitungS	Pharmazie, Technik, Recht, Anglistik, Politikwissenschaft, Lateinamerika Skandinavistik, Astronomie, Geografie, Forst	Volltext Relevanz, unscharf Highlighting des ersten Treffers
Schriften auf Hochschul-Server, EKZ-Rezensionen	SWBplus	Nutzt zusätzlich HBZ, GBV, HEBIS, OBV, BSZ	Kunst, Archaologie, Ägyptologie, Südasien Psychologie, Theologie, Kriminalistik	Volltext Relevanz, unscharf Highlighting des ersten Treffers
	Adam	Nutzt zusätzlich DNB, BSZ, +?	Geschichte, Sprachen, OstEurop. Sprachen	Volltext Relevanz, unscharf Highlighting des ersten Treffers
	intelligentCAPTURE	Nutzt zusätzlich DNB, GBV,+?	alle, Technik, Sozialwissenschaft, Romanistik	Boolean
	intelligentCAPTURE	HTW und TU	Wirtschaft, Technik u.a.	Boolean
	Eigenentwicklung, dann Adam			Boolean
	intelligentCAPTURE, Index/Abstract C3	Nicht zentral, jeweils nur in den Bibliotheken	Wirtschaft u.a.	Boolean
	C3		Reihe A	Boolean mit Volltext
	intelligentCAPTURE		Alles – u.a. mit Wissenschaft	Boolean mit Volltext
	intelligentCAPTURE		Neu italienische Publikationen,	Lizenzpflichtig
	Von US-Verlagen, Library Thing u.a. Lieferanten	Sammelt und vertreibt weltweit	Alle Fachgebiete	Lizenzpflichtig

Konferenzbänden. Nicht alles muss immer nach außen getauscht werden – die Zentralbibliothek Wirtschaft in Kiel und Hamburg arbeitet neben Inhaltsverzeichnissen von Monografien mit intelligentCAPTURE die Publikationen des Instituts für Weltwirtschaft und von anderen Leibnitz-Instituten als komplette Volltexte auf – manche Titel über 400 Seiten lang, zumeist wissenschaftliche Titel, die über Libreka, der Suchmaschine des Börsenvereins des deutschen Buchhandels, nie verfügbar sein werden. Menge, Dringlichkeit, Personalressourcen und Budgets erfordern die Modelle

- eigene Produktion,
- Projekte in den eigenen Räumen durch externe Partner oder
- ganz externe Dienstleistungen.

Jedes Modell wird am Markt angeboten.

Fazit

Die „Bibliothek als Funktion" kehrt „virtuell" zunehmend zum Wissenschaftler zurück. Viel umfassender denn je, weit besser nutzbar und zunehmend maschinell auswertbar. Technisch rückt sie der Antwortmaschine näher – ist aber noch lange nicht am Ziel. In Teilbereichen funktioniert es schon ganz gut, wie WolframAlpha lehrt (www.wolframalpha.com).

Diese Bibliothek aus Sicht des Wissenschaftlers verteilt sich auf viele verschiedene Speicherorte, Dienste (wie Zeitschrift, Mail/Feed, Webservices) und Programme.

„Bibliotheken als Institution" sind mit unterschiedlichem Tempo auf dem Weg zum „Rich Catalog". Über eigene Produktion (Digitalisieren, Importieren, Indexieren), Crosslinking, modernes Volltextretrieval und neue Interfaces und Funktionen kommen Sie den Benutzererwartungen langsam näher, die sich mehr an Google, Amazon, eBay, Wikipedia, YouTube, Delicious, Twitter, Skype und sicherlich bald weiteren neuen Hypes orientieren. Nur einholen werden sie die Erwartungen wohl kaum, sind sie doch eher reaktiv, sammelnd, bewahrend, mit viel „Altlast" behaftet als an vorderster Front des globalen Internet- und Informationsmarktes.

Seit Einführung der IT in Bibliotheken gehen sie den Weg des Zusammenschlusses. Waren in den letzten 25 Jahren die regionalen Bibliotheksverbünde die erste Wahl, so könnte OCLC bald global diese Rolle übernehmen: ein zentraler Katalog (WorldCat), nationale, regionale, thematische und lokale Sichten, lokale Medienadministration – und dazu das Crosslinking auf Googles digitale Bibliothek, wo sich auch die meisten Verlagsangebote und Open Archives wiederfinden. Shibboleth regelt die Rechte. Doch wie die Geschichte lehrt, alle großen Reiche zerfallen irgendwann – gerade Webtechnologien sind ein guter Nährboden für laufende Innovation und Diversifizierung.