

intelligentCAPTURE und dandelon.com: Collaborative Catalog Enrichment

Manfred Hauer

AGI – Information Management Consultants
Neustadt/Weinstrasse, Deutschland

manfred.hauer@agi-imc.de

<http://www.dandelon.com>

Abstract

Wissenschaftliche Bibliotheken können insbesondere papierbasierte Medien durch Digitalisierung, OCR, maschinelle Indexierung, Dokumentenanalyse, Informationsextraktion und modernes Information Retrieval zurückholen in den Wahrnehmungsbereich ihrer Klientel. Die Indexierung von Bibliothekskatalogen wird anhand von drei Messreihen problematisiert und in einem Retrieval-Test dandelon.com und Google Scholar gegenübergestellt. Dabei fallen OPACs ohne Catalog Enrichment, sprich Inhaltsverzeichnisse, maschinell generierte Deskriptoren und semantisches Retrieval, hinter die neuen Ansätze deutlich zurück. Durch Grundsatzentscheidungen bei Bibliotheksverbänden wird eine starke Expansion derartiger Inhalte vorhergesagt.

Da viele Bibliotheken die gleichen Medien sammeln, bietet sich ein kollaborativer Ansatz an. Im deutschsprachigen Raum ging der Anstoß in diese Richtung nicht unwesentlich von AGI und ihrem Programm intelligentCAPTURE als Software zu Erfassung, Aufbereitung, Konvertierung, Indexierung und Verteilung und dem wissenschaftlichen Suchdienst "dandelon.com" aus. Bei der Produktion werden zusätzlich Collaboration Tools wie Mail, Chat, IP-Telefonie, Application Sharing, Blog und Webservices genutzt.

1 Digitales Inhaltsverzeichnis als Brücke zum Buch

Artikel und Bücher sollten möglichst komplett online sein, sonst werden sie von den heutigen Studenten und zukünftigen Wissenschaftlern kaum noch wahrgenommen, gelesen und zitiert. Durch den Medienträgerwandel von Papier zu digitalem Dokument droht eine erhebliche Menge bisher

gesammelten Wissens zu versenden. Diese Lücke betrifft insbesondere eine Zeitspanne von ca. 68 Jahren: Medien älter als 70 Jahre sind meist urheberrechtsfrei, also digitalisierbar ohne juristische Klärung und nur die letzten 2-5 Jahre sind von einigen Verlagen bisher zusätzlich online verfügbar, zumindest bei den ganz großen Verlagen.

Etablierte Reputationsverfahren (Impact Factor) hinken hinter den technischen Möglichkeiten her und präferieren klassisch gedruckte Medien. Zitationslisten widerspiegeln nur noch begrenzt das tatsächliche Lese- und Rezeptionsverhalten der Autoren. Gewiss, der Lesekomfort und die Rezeptionsintensität spricht auch weiterhin für gedruckte Medien, doch der zunehmende Wahrnehmungsverlust nicht. Neues Wissen ist nicht weniger betroffen. Dies gilt auch für jeden Artikel in diesem gedruckten Tagungsband. Nicht ausreichende Budgets zum Kauf digitalen Contents stehen dem entgegen und mittels des Urheberrechts schützen sich Verlage vor rigoroser Digitalisierung ohne Entschädigung.

Ein Mittelweg ist die digitale Erschließung von Inhaltsverzeichnissen, urheberrechtsfrei, weit aussagekräftiger als allein die intellektuelle Erschließung klassischer Bibliothekskataloge, räumlich nicht beschränkt: digital finden, analog lesen.

Als AGI 1996 erstmals einen Buchhandelskatalog mit Inhaltsverzeichnissen erweiterte, war die Idee noch jung. Doch die praktische Umsetzung im großen Stil ist für derzeitige wissenschaftliche Bibliotheken im Jahr 2007 noch immer die Ausnahme. Doch die Lage ändert sich derzeit rasch. Ein gemeinsamer Beschluss der Bibliotheksverbände in Deutschland und Österreich brach das Eis und die Deutsche Nationalbibliothek plant Aktivitäten. AGI ist in dieser Entwicklung nicht nur Software-Entwickler und Content-Provider, sondern mittlerweile ein führender Dienstleister mit eigenen, dezentralen Scan-Teams.

2 Methode: Zusammenspiel von Menschen, Organisationen, Programmen und Servern

Es entwickelte sich oberhalb des Bodensees eine erste Community von intelligentCAPTURE-Anwendern, die seit Sommer 2003 ihren Content nicht nur in die lokalen Bibliothekssysteme integrieren, sondern über „dandelon.com“ zusätzlich zentral speichern und gezielt dezentral distribuieren. Dandelon.com ist mehr als ein Dokumenten-Management-

System, für die Öffentlichkeit erscheint es primär als Suchmaschine für wissenschaftliche Literatur mit einer semantischen, crosslingualen Suche.

Mittels Scanning – von bisher rund 1 Million Papierseiten – über Fujitsu-Flachbettscanner (ein Weltmarktführer bei Scannern mit 40 % Marktanteil - http://www.fujitsu.com/us/news/pr/fcpa_20061213-01.html) wurden bisher rund 280.000 Bücherinhaltsverzeichnisse in 35 Sprachen in bisher 13 Bibliotheken in vier deutschsprachigen Staaten gescannt und auch ein wenig in Italien. Für Ende 2007 prognostiziert der Autor eine Menge von 500.000 Inhaltsverzeichnissen. Die Mehrheit ist in dandelon.com such- und tauschbar. Dazu kommen 470.000 Aufsatztitel. Noch wenige zeigen den kompletten Volltext (vor allem IWP – Informationswissenschaft- und Praxis der DGI), einige zeigen ihn nur in der Vorarlberg Landesbibliothek. Hier greift das Digital Rights Management der Zeitschriftenagentur Swets. Beim Rest führt der Standortnachweis zum jeweiligen Regal.

Die gesamte Lösung ist ein komplexes Zusammenspiel von

- * immer mehr Bibliotheken, Bibliothekssystemen und Verbundzentren
- * mehreren Entwicklungs-, Hostings- und Projektstandorten bisher vorwiegend in Deutschland und Indien – eine virtualisierte Struktur
- * der Integration von Thesauri von den Vereinten Nationen, der Europäischen Union, mehreren Dokumentationszentren, Projekten und aus eigener Produktion,
- * von Basisprogrammen verschiedener Hersteller, vor allem
 - * IBM mit Lotus Notes & Domino als zentrale multimediale Datenbankumgebung und Entwicklungsplattform für Workflows und mit Lotus Sametime für die realtime Collaboration und darin der GTR als Retrieval-Kernel
 - * Abbyy's OCR mit zwei FineReader-Versionen (mobile und Engine)
 - * Adobe mit Acrobat für Formatierung und Bearbeitung
 - * IAI mit der maschinellen Indexierung, genannt CAI- Engine (=Autindex),
 - * Z39.50-Client und XML-Konnektoren,
 - * Kofax Image Controls für bestmögliche Images und schnelles Scanning, ISIS-Scanner-Treibern für Fujitsu-Scanner (andere möglich), TWAIN-Treiber werden noch unterstützt
 - * Servern von SUN und Cients unter Microsofts Windows
- * Programmierung und Zusammenstellung von AGI.

14 Thesaurus-Datenbanken mit 1,6 Millionen Begriffen, mehrere Medien-Datenbanken, eine Image-Datenbank für Cover-Pages, Logging und mehrere kleinere Datenbanken für Konfiguration, Monitoring, Buchkaufabwicklung, Kundenverwaltung etc, gehören zusammen. Mittlerweile ca. 11 Millionen Dokumente/Datensätze insgesamt.

Social Software wird als Zusammenspiel beliebiger Menschen mit zumeist einfachen Editoren verstanden (z.B. Wikis, Weblogs). Bei dandelon.com spielen ausgewählte Menschen - nur Information Professionals, welche Medien gezielt selektieren und Urheberrechte achten - und zahlreiche wissensbasierte Programme, die automatisch über Workflows miteinander interagieren, zusammen. Der Aufwand zum Editieren und Kommunizieren durch Menschen ist gering und hoch zwischen Computern und Programmen.

3 Benchmarking: Indexierung- und Retrievaltests

Viel Aufwand, aber lohnt sich dieser? Studenten und Wissenschaftler greifen prozentual, wenn sie etwas suchen nur noch selten auf Bibliothekskataloge zu - um die 3 % laut DNB. Andererseits zeigen die Anzahl von Fragen, die Bibliothekskataloge monatlich beantworten, dass sie nahe bei kommerziellen Informationsanbietern wie FIZ Technik, STN, GENIOS liegen - also rund 20.000 bis 80.000 Abfragen pro Monat, die Allermeisten davon beziehen sich auf Sachthemen, die Minderheit auf Namen und bibliografischen Angaben. Auch wenn insbesondere Google in Europa den Suchmarkt zu monopolisieren versucht und wir alle helfen jeden Tag mit. Google hat den Suchraum eher erweitert, klassische und auch mehrere neue Angebote aber noch nicht ganz ersetzt, der Angriff läuft. Es bleiben Nischen, dazu zählen die OPACs der wissenschaftlichen Bibliotheken und die verschiedenen, spezialisierten Fachinformationsdatenbanken.

Mit drei Messreihen hat der Autor, die Situation beleuchtet:

1. Varianz von Suchergebnissen durch grammatische Varianten, Synonyme und Übersetzungen

Eine Gruppe von 20 Nachdiplomstudenten der HTW in Chur, teils Mitarbeiter der ETH-Bibliothek und anderer Bibliotheken in der Schweiz, alle mit Erfahrungen im Umgang mit dem Schweizer Bibliotheksverbundkatalog

NEBIS, sollten im Sommer 2006 Begriffsvarianten und die Trefferzahlen notieren für Google Scholar, zu der Zeit vermutlich um die 7,5 Millionen Dokumente, NEBIS mit ca. 4,5 Millionen Katalogdatensätzen auf Basis von ALEPH und mit Retrieval-Unterstützung durch OSIRIS und dandelon.com mit damals 130.000 Dokumenten (siehe Abb. 1).

Varianz von Suchergebnissen					
	2. Juni 2006 - NDL-Studium HTW Chur, 20 Studenten				
	Google ca. 7.2 Mio	Nebis 4,5 Mio	Faktor x 1,6	dandelon 130.000 Dokumente	Faktor 54
Suchwort	Scholar	Nebis	x 1,6	dandelon	x 54
prävention	26100	2581	4130	1413	76303
praevention	9690	2581	4130	1283	69282
prevention	1290000	9714	15542	1348	72792
Vorsorge	11400	443	709	10263	554202
Unfall	32700	1434	2294	911	49194
Unfälle	6210	3737	5979	911	49194
Unfaelle	6210	3737	5979	911	49194
accident	616100	4287	6859	642	34668
accidents	435000	8204	13126	142	7668
injury	2020000	503	804	2118	114372
injuries	828000	1280	2048	104	5616
blume	81700	748	1197	661	35694
flower	415000	1518	2429	793	42822
fleur	17000	199	318	786	42444
Blumen	22600	1368	2189	910	49140
flowers	462000	1571	2514	101	5454
fleurs	18800	1297	2075	21	1134
museum	883000	35143	56229	2340	126360
museums	159000	7713	12341	2461	132894
museen	11800	6594	10550	2340	126360
musée	31300	12195	19512	2055	110970
musee	23800	12195	19512	2055	110970
musseum	46	0	0	2152	116208
Territorialisierung	174	3	5	21	1139
Staatsbildung	684	109	174	84	4536
Territorialherrschaft	59	2	3	14	756
Landesherrschaft	327	23	27	63	3402
Landesherr	466	10	16	79	4566
Territorialherr	25	0	0	7	378
Landesherrlich	17	0	0	3	162
"state-building"	22100	762		19	1026
"emergence of territories"	4	0	0	0	0

Abbildung 1

Zum Vergleich wurde in den markierten Spalten die Ergebniszahl von NEBIS und dandelon.com auf die Menge von Google Scholar hochgerechnet, davor die Absolutzahlen. Anbei ein Ausschnitt aus dem Anfang der Messreihe. Sie zeigt in der Spalte, dass z.B. Google Scholar beim Begriff „Blume“ zwischen 17.000 und 462.000 schwankt, NEBIS zwischen 318 und 2500 und

dandelon.com zwischen 1100 und 49.000 schwankt, aber insgesamt weit weniger streut, immer wenn, die Varianten in den Thesauri erkannt werden. Geringfügige Varianten eines Suchbegriffs führen zu teils extrem verschiedenen Ergebnissen sowohl innerhalb des jeweiligen Suchsystems als auch zwischen den Suchsystemen. Addiert man die drei Spalten zusammen, so fällt NEBIS als Repräsentant von Bibliothekskatalogen hinter dandelon.com und Google Scholar zurück.

Man kann einwenden, diese Testreihe ist zu kurz, die Begriffsauswahl durch die Studenten zu zufällig, nicht repräsentativ, die Dokumentenmenge von dandelon.com durch die Hochrechnung um den Faktor 54 evtl. falsch bewertet und zu ungenau. Diese Einwände sind richtig. Dennoch, die komplette Messreihe zeigt noch deutlicher die Schwächen des Bibliothekssystems im Vergleich zu Google Scholar, lässt aber im Vergleich von dandelon.com und Scholar keine Aussage zu, die eindeutig für oder gegen dandelon.com spricht, da 130.000 Dokumente angesichts der Anzahl möglicher Themen noch keine statistisch validen Angaben erlaubt. Dennoch, Google liegt wohl keineswegs immer vorne.

Die Messung überraschte die Studenten sehr, solch große Unterschiede waren niemandem bewusst, da man sich meist mit einer Frage und einer Antwort bei den meisten Suchsystemen zufrieden gibt.

Ziel eines Information Retrieval-Systems muss es sein, derartige Schwankungen beim gleichen Begriff möglichst gering zu halten.

2. Indexierungstest: Varianz von Autor, Indexierer, Leser

Ist es nötig, so viele Sprachvarianten zu testen und zu vergleichen? Ist unsere Sprache denn nicht hinreichend klar um eine Sache zu beschreiben?

Üblicherweise werden Indexierungstests zwischen zwei menschlichen Indexierern oder zwischen Mensch und maschineller Indexierung gemacht. Zur Verbesserung der Indexierungsqualität sind solche Messungen nützlich, sie greifen aber zu kurz, wenn es um Retrieval geht – die Frage muß vielmehr lauten: Kommen Autor und Suchender/Leser zusammen?

Eine Messreihe (Hauer 2005) mit 99 Studenten und 33 Aufsätzen zeigte, dass nur 21 % der zur Beschreibung verwendeten freien Indexierungsbegriffe zwischen drei Personen übereinstimmten. Drei steht für das Wechselspiel von Autor / Indexierer / Sucher. Die Brücke von Autor zu Sucher ist also schwer zu schlagen. Autoren orientieren sich genauso wenig an

Klassifikationssystemen wie ihre Leser und damit stehen Bibliothekare recht einsam auf weiter begrifflicher Flur.

3. Known-Item-Test: Genau das richtige Buch finden

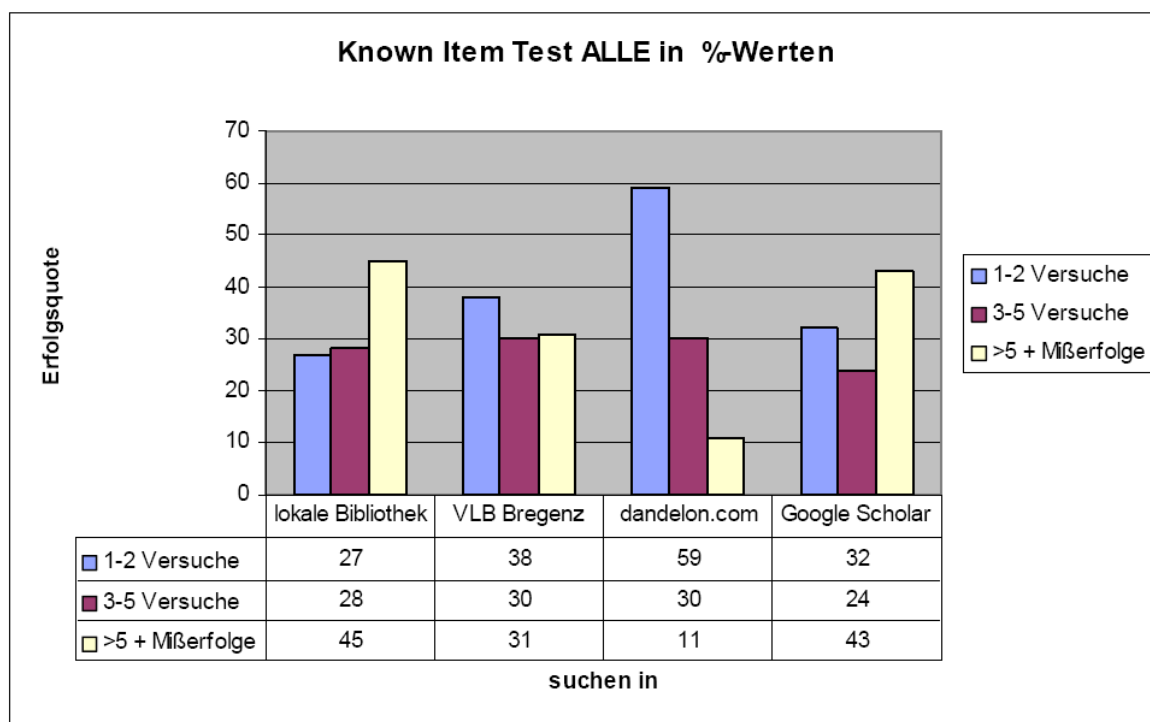


Abbildung 2

Ob dies gelingt, war Gegenstand der dritten hier vorgestellten Messreihe (Hauer: Vergleich der Retrievalleistungen von Bibliothekskatalogen gegen erweiterte und neue Konzepte. In: ABI-Technik Dez. 2005): 75 Studenten mussten 295 Bücher finden, deren Inhalt sie gut kannten, aber nichts über bibliografische Details wussten. Nur durch thematische Recherche konnten die Titel gefunden werden. Die Abbildung 2 zeigt einen Vergleich von lokalen Katalogen (FH Burgenland - 14.000 Titel , HTW-Chur 4000 Titel , NEBIS ca 4 Mio Titel), der Katalog der Vorarlberger Landesbibliothek (VLB), welche intelligentCAPTURE-Indexierung und Inhaltsverzeichnisse in den ALEPH-Katalog integriert, dandelon.com und Google Scholar.

Die Grafik zeigt, dass in Bibliothekskatalogen insgesamt 45 % der Studenten scheiterten – dies stimmt nicht für den sehr kleinen Katalog für Chur, während im NEBIS-Katalog die Erfolgsrate extrem niedrig war. Der VLB-Katalog konnte durch das Catalog-Enrichment via intelligentCAPTURE-Deskriptoren Google Scholar bereits übertreffen. Mit 59 % Erfolg mit 1 bis 2 Versuchen lag dandelon.com deutlich vorne. Damals, heute hat dandelon.com

nicht mehr 60.000 Titel, sondern über 210.000 öffentlich online, Google Scholar ist gewachsen, eine Wiederholung der Messung wäre spannend.

Scanning von Inhaltsverzeichnissen, die Übertragung von maschinellen Indexaten und PDFs der Inhaltsverzeichnisse an Bibliothekskataloge zahlt sich aus und in Kombination mit semantischen Ressourcen und Ranking-Verfahren führt es signifikant schneller auf jene Medien hin, welche der Anfrager meint.

4 Herausforderungen, Hindernisse und Barrieren

Bezogen auf Inhaltsverzeichnisse gibt es keine wesentlichen Hindernisse durch das Urheberrecht oder Digital Rights Management-Systeme.

Die Kosten für Personalkosten und die Subskriptionsgebühren für die Nutzung von intelligentCAPTURE und dandelon.com werden häufig als Hindernis von Bibliotheken bezeichnet - es ist eher eine Frage der Präferenzen, denn der Aufwand pro Buch kann im Schnitt mit 1,50 € gerechnet werden und dem stehen Einspareffekte bei überflüssigen Ausleihen entgegen. Nur 1 von 10 aus Magazinen ausgeliehenen Titeln wird wirklich gelesen, schätzen viele Bibliothekare. Bei oft entliehenen Titeln sind Mehrfachexemplare nötig - alles nicht zum Nulltarif, aber versteckt in alten Haushaltsposten. Demgegenüber „share“t dandelon.com und auch die Verbünde teilen digitalisierte Daten ohne wesentliche Mehrkosten, die Kosten sind also deutlich degressiv mit einer wachsenden Menge von Produzenten.

Hindernisse sind auch persönliche Einstellungen, Ängste von Sacherschließern, Ablehnung von komplexen maschinellen Verfahren und „Not-invented-here“-Haltungen. Jenseits dieser menschlichen Dimension ist zu berücksichtigen, dass Bibliotheken nicht in 100 oder 1000 Medientiteln denken, sondern oft in Hunderttausenden oder Millionen, das macht jede Kursänderung schwerfällig.

Erfreulicherweise haben die Verbünde und die Deutsche Nationalbibliothek 2006 Impulse gesetzt, die Richtung ist neu bestimmt, die erste Million digitale Inhaltsverzeichnisse im deutschsprachigen Bibliotheksraum prognostizieren wir für 2008. Dann fehlen nur noch 14 Millionen Sachtitel aus früheren Jahren, schätzen wir.

5 Best-Practice

intelligentCAPTURE entwickelte sich zunächst dort, wo die Bibliotheksverbände und zentrale Strukturen nicht oder nur schwach präsent waren. Innovationen kommen nicht aus den etablierten Zentren, sondern vom Rande her. Sie setzen sich aber erst durch, wenn die Zentren sie akzeptieren und fördern. Der GBV spielt hier für AGI die Rolle des Förderers und zieht mittlerweile daraus einen erheblichen eigenen Nutzen.

Staatliche Förderung von EU oder BMBF wurde bis heute verwehrt.

Heute ist intelligentCAPTURE ein sehr performantes und hochwertiges Verfahren zur Digitalisierung von Inhaltsverzeichnissen, deren maschineller Auswertung und über dandelon.com des Austausches mittlerweile über fünf Staaten. Es ist auch die erste mobile Version (siehe Abbildung 3) für Digitalisierung zwischen Regalen. Mit 42.000 Titeln gescannt, komplett verarbeitet und publiziert via GBV an drei Workstations in zwei Monaten an der SUB Hamburg (Mitte Februar bis Mitte April 2007) liegt die Messlatte für Qualität (Image, OCR, Indexierung) und Performance nicht gerade niedrig.



Abbildung 3: *intelligentCAPTURE mobile* wird direkt zwischen engen Regalen eingesetzt. Der Scanner fährt zum Buch, die hohen Medientransportkosten werden stark minimiert. Über WLAN kommuniziert die Scanstation mit dem Bibliothekssystem, dandelon.com, dem Domino-Server und dem Internet. Die Mitarbeiter kommunizieren über Chat, IP-Telefonie, Application Sharing und eMail mit der Projektleitung. Vom Einsatz gibt es ein Video auf der AGI-Homepage.

6 Entwicklungsperspektive

dandelon.com integriert bereits National-, Landes- und Hochschulbibliotheken, Verlage und Buchhandel Staaten übergreifend. Weitere Vernetzungen, noch mehr Content, noch mehr Fokussierung, noch mehr Verfahren aus dem Information Retrieval werden die weitere Entwicklung kennzeichnen. Die Dokumentenstrukturanalyse und Informationsextraktion werden an Bedeutung weiter stark zunehmen. Geld- und Personalmangel bremsen noch immer das bereits Mögliche, doch wo ist es anders?

7 Literaturverzeichnis

Hauer, Manfred: Vergleich der Retrievalleistungen von Bibliothekskatalogen gegen erweiterte und neue Konzept. In ABI-Technik Dez. 2005, S 295-301