# Tiefenindexierung im Bibliothekskatalog: 17 Jahre intelligentCAPTURE

**Manfred Hauer**

## Die Idee

„Bei uns soll niemand nach Hause gehen, ohne die richtige Literatur“ – so lautet der Anspruch der Vorarlberger Landesbibliothek (VLB) in Bregenz, Österreich. Die VLB war häufig der primus inter pares unter den Bibliotheken. Als eine der ersten setzte sie ein elektronisches Bibliothekssystem, DOBIS/LIBIS von IBM ein und wechselte als eine der ersten staatlichen Bibliotheken in Europa auf ALEPH von Ex Libris. Sie ist der europäische Pionier in Tiefenindexierung, der maschinellen Inhaltsanalyse auf Basis von Inhaltsverzeichnissen und auch Abstracts.

Das Eingangs-Statement könnte auch von Google stammen. Doch statt mit Google saß 2001 Karl Rädler, Bibliothekar der VLB mit Manfred Hauer von AGI zusammen. Die VLB setzte damals schon deren Thesaurus-Entwicklungsprogramm ein, das aktuell 154.938 Deskriptoren, Nichtdeskriptoren, Abkürzungen, Synonyme, Ober- und Unterbegriffe sowie 31.384 Klassen und deren Vernetzung enthält. Beiden war klar, dass die intellektuelle Erschließung nicht in der Lage ist, die thematische Tiefe vieler Werke adäquat und in einem für die VLB bezahlbaren Rahmen abzubilden (Rädler 2008). Die meisten Bibliothekare

### Maschinelle Indexierung im Vergleich zu intellektueller – ein Beispiel

**L’université en transition - L’évolution de son rôle et des défis à relever**

Bo Göransson, Claes Brundenius https://doi.org/10.1007/978-1-4614-1236-6

| | |
|---|---|
| Number of Terms | **Resultate maschinelle Übersetzung Fra->Deu und Indexierung**<br>In document 376 / extracted 73<br>*Worte in grüner Schrift kommen auch in der GND vor (hier manueller Abgleich)* |
| Descriptors | Hochschulreform[100]; wirtschaftliche Entwicklung[39]; Innovation[39]; Debatte[15]; Institution[12]; Forschung[7]; Leistung[7] - *Zahlen stehe für Termgewichtung (max ist 100)* |
| Free Descriptors | Hochschulbildung[14]; Spitzenleistung[7]; Neupositionierung[2]; Innovationssystem[0] |
| Noun Phrases | kleines Land; akademische Institution; aktuelle Debatte; brasilianische Universität; internationale Perspektive; nationales Innovationssystem; peripheres Land; starkes Wachstum; universitäre Forschung; verändernde Herausforderung; verändernde Rolle; wandelnde Rolle |
| Countries | China C9CHIN; Schweden C4EUSW; Kuba C5CUBA; Vietnam C9VIET; Uruguay C3URUG; Russland C4EXRU; Deutschland C4EUGE; Dänemark C4EUDE; Lettland C4EXLA; Tansania C6TANZ; Südafrika C6SOUT; Frankreich C4EUFR<br>Erkannte Namen weggelassen – Liste zu lang |

**Indexierung in Bibliotheken/Bibliotheksverbünden**

SWB: Economics; Education; Development Economics; Economic policy; Economics/Management Science

BVB: Enseignement supérieur / Aspect économique; Universités / Administration; EDUCATION / Higher; Education; Higher / Economic aspects; Universities and colleges / Administration; Education; HigherxEconomic aspects; Universities and collegesxAdministration

LoC: Education; Higher Cross-cultural studies; Universities and colleges Cross-cultural studies; Educational change Cross-cultural studies; Education and globalization Cross-cultural studies.

hatten sich auf ein bis zwei Klassen und/oder auf drei bis fünf Schlagworte pro Werk seit 200 Jahren eingependelt. Dies entsprach der Anzahl der möglichen Durchschläge bei Karteikarten als Katalog. Es war also eine technisch-organisatorische Restriktion.
Der Aufstieg der dokumentierenden Wissenschaftler und Dokumentare ist eine Kritik an dieser Erschließungsrestriktion. Sie erfanden angesichts der wachsenden Zahl wissenschaftlicher Zeitschriften im frühen 19. Jahrhundert Abstracting-Netzwerke und nutzten natürlichsprachige Deskriptoren zur Beschreibung, die mit der Zeit stärker kontrolliert zu Thesauri heranwuchsen. Es entstanden Netzwerkstrukturen statt nur hierarchischer Bäume. Dokumentation passte gut zu den Information Retrieval-Lösungen wie einst GOLEM von Siemens oder STAIRS von IBM (ab 60er/70er Jahre) und internationalen Online-Diensten wie „Dialog“. Ca. 10 Deskriptoren waren normal in deutschen Fachinformationszentren, den Produzenten hinter den Datenbanken und Hosts. Sie wurden intellektuell ermittelt, von Hand erfasst und mit Abstracts ergänzt. Der Abstract war über den Volltext-Index erschlossen. Ein Standard, den Verlage zunehmend übernahmen.
Ende der 80er Jahre im Rahmen einer Kooperation mit Prof. Harald Zimmermann von der Informationswissenschaft in Saarbrücken, gelingt es mit Primus IDX einen ersten Stein der maschinellen Erschließung in ein Bibliotheksfenster zu werfen. Doch die Textbasis der Kataloge war schlicht zu dünn für gute Resultate und die Komposita-Zerlegung verallgemeinerte zu sehr. Mehr als Titel und gelegentlich Untertitel war nicht verfügbar. IDX lief noch etliche Jahre an der UB Düsseldorf und in wenigen anderen Bibliotheken.

## Die Umsetzung

1999 greift AGI die Technologie der linguistisch basierten maschinellen Indexierung neu auf, wendet sie im Presse-Clipping als Erschließungsmethode nach dem Scannen und OCR der Artikel an. Über Scanning und OCR war jetzt ausreichend Text vorhanden. Diesmal ist das IAI in Saarbrücken der Partner, Prof. Zimmermann und Prof. Hans Haller dessen Gründer. AUTINDEX heißt das neue Produkt. AUTINDEX ist ein regelbasiertes KI-Verfahren auf Basis linguistischer Wörterbücher, Grammatikregeln, Ausnahme-Listen und statistischer Gewichtung.
Hauer und Rädler kamen 2001 überein, eine Applikation zu bauen, die Inhaltsverzeichnisse von Büchern einscannt, in Text wandelt, inhaltlich analysiert und die Indexierungsresultate in den Katalog schreibt. Für den Volltext war nur in den wenigsten Bibliothekssystemen Platz. Im Januar 2002 ging es los. Etwa zur gleichen Zeit startete Syndetics in den USA auf Basis von ISBN-Meldungen und Text-Dateien der Inhaltsverzeichnisse. Inzwischen ist es ein Dienst von ProQuest. Inhaltsverzeichnisse von Sachliteratur widerspiegeln die Terminologie und die thematische Logik der Autoren umfangreich und detailliert. Sie sind für viele Leser der primäre Ort bei der Literaturauswahl. Inhaltsverzeichnisse sind nicht ausreichend urheberrechtlich schutzwürdig wie mehrere Juristen bestätigten. Sechs Jahre später findet diese Einschätzung Niederschlag in einer Vereinbarung zwischen dem Börsenverein und den Bibliotheksverbänden. Dabei handelt es sich nur um referentielle Hinweise, nicht um zu schützenden Text.
Das gesamte Inhaltsverzeichnis zählt meist 200 bis 500 Worte (Tendenz wachsend) und die maschinelle Inhaltsanalyse erkennt darin ca. 10 % der Worte als relevant. Durch die Linguistik werden Wortvarianten auf ihre Grundform zurückgeführt (Wohnhäusern -> Wohnhaus), durch die Komposita-Zerlegung (aber nicht Ausgabe) wird erkannt, dass es um Gebäude geht, durch Satzstrukturanalyse wird erkannt, ob das Wort als Subjekt oder Objekt im Hauptsatz auftaucht oder in einem weniger wichtigen Nebensatz und durch Häufigkeiten rund um Gebäude werden weitere Termgewichtungen pro Satz und Text ermittelt. Ein integrierter Allgemein-Thesaurus mit 220.000 Einträgen – er deckt sich inhaltlich und hinsichtlich der Menge relativ weit mit der Gemeinsamen Normdatei – trägt zur Gewichtung und Normalisierung bei. Durch die verbal deutlich umfassendere Beschreibung, eingespielt in den Katalog, werden Titel erstmals über aktuelle Fachbegriffe findbar, die früher nicht erfasst und/oder in der Normterminologie nicht bekannt waren. (ausführlicher in Hauer 2017)
Im Januar 2002 wurde „icapture 1.0“ in Bregenz eingeführt und in b.i.t.online darüber berichtet. 2008 waren dort 70 % des Bestands ohne Fördermittel in Eigenleistung bearbeitet sowie sämtliche Neuerwerbungen und seit 2004 auch die Zeitschriftentitel. Es folgen bald die UB St. Gallen, die Liechtensteinische Landesbibliothek, die HTW Berlin und Westfälische Hochschule, die ULB Darmstadt als erste Universitätsbibliothek. 2003 stellt AGI auf der IFLA in Berlin das Produktionsverfahren – nun „intelligentCAPTURE“ – vor und dazu die gemeinsame Austausch-Plattform „intelligentSEARCH“, seit 2004 mit dem Label „dandelon.com“ – com für Community – bekannt. Google Scholar startet im gleichen Jahr. Die Bezeichnung Verbund sollte vermieden werden, denn schon früh zeichneten sich Ablehnungen durch einige Verbünde ab. 2005 schlossen AGI und GBV einen Kooperationsvertrag, der bis heute trägt.

## Migrationsstufen der Medienerschließung

| Indexierungs-verfahren | Hierarchische Klassifikation, Systematik | + Schlagworte | Polyhierarchischer Thesaurus | Maschinelle Indexierung und Retrieval mit Relevance Ranking und Facetting | Retrieval mit Relevance Ranking, Facetting, Tracking user behavior, Weighting cross references (citations) |
|---|---|---|---|---|---|
| **Abstraktionsgrad** | Sehr hoch | etwa wie Titel | deutlich konkreter | Textnah | Volltext |
| **Indexierungstiefe durchschnittliche Anzahl suchbarer Terme** | 1 Klasse Karteikarten sortiert nach Autor und zweite Sortierung nach Klasse/Nummer | bis zu 5 Schlagworte, Karteikarten-Limit durch Durchschlag und Sortier-aufwand | 12 Deskriptoren (Limits durch menschlichen Aufwand bei Lesen, Erfassen, Thesauruspflege), teils Klassen und meist Erstellung/Über-nahme von Abstracts | 20-50 frei und kontrollierte Deskriptoren, Ca. 10-20 % normalisierte, typisierte Wörter aus Quelltext für 100 Sprachen verfügbar mit Übersetzungsoption in Zielsprachen und Volltext/Übersetzung von Titel, TOC oder Abstract | 100 % des Textes suchbar |
| **Katalogtyp** | ab 80er Jahren OPAC | ab 80er Jahren OPAC | Retrieval-Systeme mit boolescher Logik und Feldsuche | Retrieval mit Relevance Ranking, Facetting, Volltextindexierung | Google |
| **Textbasis, Input** | Gebundene Werke | + weitere analoge Medientypen | primär Aufsätze | Print und ePublikationen: Inhaltsverzeichnis, Abstract u.a. Text digital/digitalisiert | Volltext + Katalogdaten |
| **Indexierer** | Bibliothekare | Bibliothekare | Dokumentare Spezialbibliotheken | Maschinelle linguistische und/oder statistische Verfahren | Google |
| **Zeitaufwand pro Medieneinheit für Indexierung** | 2 Minuten Erfassen/einst manuelles Sortieren | 4-10 Minuten | 10-60 Minuten | 1 Minute für Digitalisierung bei Papier pro TOC | Keine menschliche Arbeit |
| **Aufwand Software-Entwicklung** | Von Null bis hoher Aufwand und zunehmend modernere Informatik-Konzepte, wachsender Bedarf für Schnittstellen und Standardisierung | | | | |

AGI hatte nie eine „Kataloganreicherung“ im Sinne, sondern wollte von Anfang an eine neue Qualität von Katalogen. Mit begrenzten Mitteln – immer ohne Förderung – war dandelon.com, verbunden mit den jeweiligen Bibliothekskatalogen der Anwender von intelligentCAPTURE eine frühe „Discovery Engine“. Diesen Ball spielten die großen internationalen Anbieter erfolgreicher weiter.

### Aktueller Stand

intelligentCAPTURE läuft in der Version 9. Neuronale Netze nehmen darin an Bedeutung zu: Abbyy FineReader Engine 12 nutzt sie für die Zeichen- und Layout-Erkennung und Klassifikation. Die Indexierung verarbeitet Text aus über 100 Sprachen auf Basis der neuronalen Netze mit der Indexierungssprache Deutsch und liefert optional zusätzlich die Titel, die Sacherschließung der Bibliothekare, die maschinelle Erschließung und die Volltexte wiederum in über 100 Zielsprachen. Die neueste Auswertung von Zeitschriftenaufsätzen nutzt ein selbst trainiertes neuronales Netz zur Erkennung von Autoren und Titeln.

Viele Bibliothekare misstrauen dem breiteren Indexierungsansatz bis heute, wie der letzte Bibliothekartag in Berlin erneut zeigte. Den Vergleich zur klassischen Indexierung hält der Autor für einen Denkfehler, stattdessen ist die Accuracy beim Information Retrieval zu testen: Findet der Leser schnellstmöglich die für

*intelligentCAPTURE office*

ihn relevante Literatur? Bei genauerer Analyse wird offensichtlich, dass der klassische Ansatz keine ausreichende Retrievalqualität ermöglicht. Beide Verfahren können im intelligenten Zusammenwirken mit gewissen Funktionalitäten moderner Suchmaschinen erst die Synergien erzeugen, die in der Lage sind, die Retrievalqualität auf ein neues Niveau zu heben. Eine Voraussetzung dabei ist allerdings, dass die intellektuellen Verfahren die polyhierarchischen und polydimensionalen Zusammenhänge des Begriffs- bzw. Benennungsraumes auch wirklich abbilden, so dass diese implizite Information von Suchmaschinen in der Recherche entsprechend benutzerfreundlich in Szene gesetzt werden kann. Die diesbezüglichen Defizite sind offensichtlich. Dennoch sind inzwischen viele Millionen Inhaltsverzeichnisse in den Katalogen mit dem PDF des Inhaltsverzeichnisses verlinkt (und in wenigen Spezialbibliotheken teils auch händisch erfasst). Discovery Engines indexieren auch deren Text. Die Anwender von intelligentCAPTURE tragen im Regelbetrieb mit 4500 gescannten und 1100 eBook-Inhaltsverzeichnissen monatlich zum Wachstum bei (2018). Der Zuwachs insgesamt – mit einem Projekt – betrug in 2018 über 340.000 Titel und 34.540 Current-Content-Titel, teils mit Abstract. Insgesamt sind 3,2 Millionen Titel in dandelon.com für jeden recherchierbar mit Link auf die Bibliotheken, die den jeweiligen Titel mit intelligentCAPTURE angefasst haben. Nur einmal wird gescannt, die anderen nutzen nach. Zusätzlich kann der jeweilige Titel über eine Schnittstelle zum KVK in weiteren Bibliotheken und im Buchhandel gefunden werden. All dies kostet die Leser nichts. Es gibt keine Werbung, keine Weiterverwertung von Benutzerdaten und auch keine Einnahmen, denn der Austausch zwischen den Anwendern von intelligentCAPTURE ist ohne jede Gebühr: Wer viel gibt, kann auch viel nehmen.

Eine HEBIS-Studie zeigte 2011, dass die Leser Inhaltsverzeichnisse im Katalog als wichtigste Ergänzung einstufen, dies stimmt mit den Log-Statistiken von GBV und DNB bis heute überein. Die DNB nennt im Jahresbericht 2017 130.531 Zugriffe auf die 1,7 Mio. Inhaltsverzeichnisse pro Tag, die im DNB-Katalog, den Verbünden und Bibliotheken verlinkt sind. Mit intelligentCAPTURE wurden bislang über 2,6 Mio. Inhaltsverzeichnisse produziert, verteilt in vielen Katalogen bis hin zum WorldCat. Die wahrscheinliche Nutzung lässt sich mit einem Dreisatz schätzen. Dandelon.com zeigt davon einen großen Ausschnitt.

## Fazit

Die Idee der Inhaltsverzeichnisse kam gut an. Die Idee der maschinellen Indexierung hingegen steht fast 20 Jahre nach ihrem Start im Bibliothekswesen in der Mehrzahl der Bibliotheken noch am Anfang. Wie weit wir unser persönliches Ziel, die Lernenden, Lehrenden und Forschenden erfolgreicher zu machen durch besseren Zugang zur richtigen Literatur zur richtigen Zeit, ist leider nicht messbar. Doch die derzeit hohe awareness für das Thema „Artifical Intelligence“ in allen Medien könnte unserer Idee einen ungeahnten Schub geben.

**Manfred Hauer**
AGI – Information Management Consultants
67433 Neustadt an der Weinstraße
manfred.hauer@agi-imc.de
https://dandelon.com
https://agi-imc.de

**Literatur**

Hauer, Manfred: iCapture 1.0 bringt Inhaltsverzeichnisse in Bibliothekssysteme und verbessert die Recherche 2002, b.i.t.online, S. 49-50, https://www.dandelon.com/d/DE_AGI472524C6164FBEE6C125827F0043C522

Rädler, Karl: Im 7. Jahr der Digitalisierung von Inhaltsverzeichnissen aus der Vorarlberger Landesbibliothek, ABI-Technik 2008, Heft 2, S. 118-119. https://www.dandelon.com/d/DE_AGI4D983B51BE566998C125826E0047E671

Nienerza, Heike; Sunckel, Bettina: Kataloge und Portale im Web 2.0-Zeitalter – Online-Umfrage für den HeBIS-Verbund, 2011, https://www.agi-imc.de/d/DE_AGI8BAB31F5F1EF6E06C125826E00442831

DNB-Jahresbericht 2017, Seite 55 https://d-nb.info/1160486344/34

Hauer, Manfred: Indexing of ebooks with intelligentCAPTURE 9.0. Ausführungen zur maschinellen Indexierung fremdsprachiger Texte und Vergleich zur intellektuellen Indexierung durch Bibliotheken. 2017, https://www.dandelon.com/d/DEAGIB2D67B77DC696142C12581D400512C5A