Small, large, medium – wie viele Informationen braucht die Wissenschaft wirklich?

Christian Hänger

This document appeared in

Bernhard Mittermaier (Eds.):
eLibrary - den Wandel gestalten
5. Konferenz der Zentralbibliothek
Proceedings of the WissKom 2010: 5. Konferenz der Zentralbibliothek, 08.-11. November 2010, Jülich
Schriften des Forschungszentrums Jülich / Reihe Bibliothek/Library, Vol. 20
Zentralbibliothek (ZB)
Forschungszentrum Jülich GmbH, Zentralbibliothek, Verlag, 2010
ISBN: 978-3-89336-668-2

Small, large, medium – wie viele Informationen braucht die Wissenschaft wirklich?

Christian Hänger

Zusammenfassung

Die Universitätsbibliothek Mannheim hat als erste deutsche Bibliothek die von der Firma Ex Libris entwickelte Suchmaschinensoftware Primo eingeführt. Sie präsentiert damit einen Großteil ihrer elektronischen und gedruckten Ressourcen unter einer Oberfläche und bietet einen direkten Zugang zur Ressource an. Es wurden allein die Datenquellen ausgewählt, die im Rahmen von freien und kostenpflichtigen Lizenzen erworben wurden und einen unmittelbarer Zugriff auf die Ressourcen erlauben. Zusätzlich zu den Bibliotheksbeständen steht mit "Primo Central" ein umfangreicher Index zur Verfügung, der die Recherche in den Titeldaten von Volltextarchiven erlaubt.

Die Nutzerinnen und Nutzer von Primo stehen vor der Herausforderung, in diesen großen Datenmengen zu navigieren und sich zu orientieren. Die Universitätsbibliothek Mannheim hat sich daher dafür entschieden, die Auswahl der Datenquellen auf relevante Informationen zu beschränken und damit die Treffermengen zu reduzieren. Ein anderer Ansatz wäre, alle verfügbaren Datenquellen in Primo zu indexieren und die Suchergebnisse danach durch die Anwendung von Algorithmen wie z.B. der Zählung der Häufigkeit des Titelzugriffs zu ordnen.

Abstract

Mannheim University Library is the first German academic library to introduce the search and delivery system Primo and is now presenting a good part of its electronic and printed resources in a single integrated user interface with direct access to the resource. Only those data sources have been integrated that have been licensed or are freely accessible and allow unhindered access to the resource. In addition to the library's holdings searching the extensive index of scholarly materials "Primo Central" allows searching full text archives.

Library patrons using Primo have to meet the challenge of navigating and orientating theirselves within huge result sets. Mannheim University Library opted for integrating only relevant data sources and thus reducing the size of result sets. Another option would be the indexing of all available data sources in Primo and then ranking results through an algorithm using click-through statistics among other criteria.

Ausgangslage

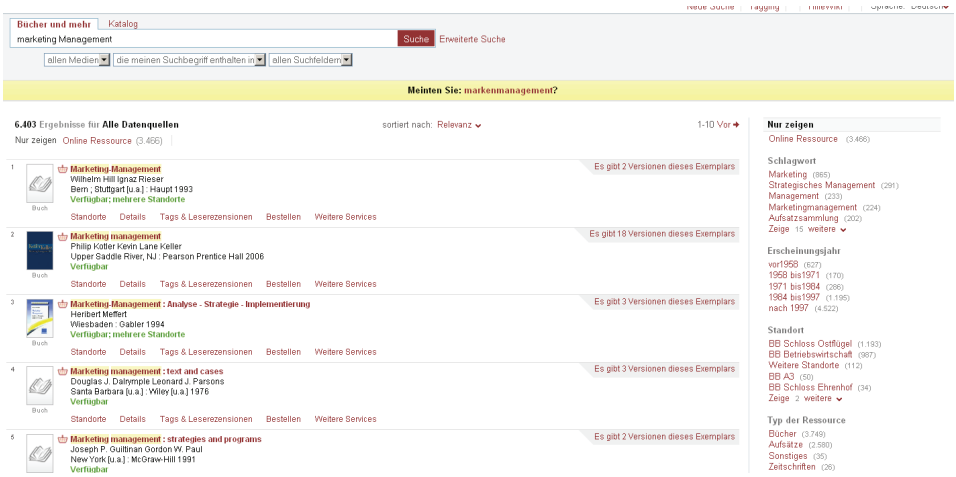
Die Universitätsbibliothek Mannheim verfolgt seit Jahren den konsequenten Ausbau der digitalen Bibliothek, um die Recherche und den Zugriff auf wissenschaftliche Informationen für die eigenen Nutzer zu optimieren und komfortabel zu gestalten. Im Rahmen regionaler und nationaler Konsortien sowie eigener Initiativen wurde ein umfangreiches digitales Angebot akkumuliert. Dazu gehören u. a. Business Source Premier (BSP) von EBSCO, SocINDEX with Full Text, Social Sciences Citation Index (Web of Science), diverse E-Book-Pakete und die von der Deutschen Forschungsgemeinschaft finanzierten Nationallizenzen. Die Informationsversorgung folgt der fachlichen Ausrichtung der Universität, die von den renommierten sozial- und wirtschaftswissenschaftlichen Fakultäten geprägt ist, die mit leistungsstarken Geisteswissenschaften, Rechtswissenschaft sowie Mathematik und Informatik vernetzt sind. Die Nutzungszahlen und die alltägliche bibliothekarische Praxis belegen allerdings, dass die E-Books und die Angebote der Nationallizenzen von den Nutzern oft nicht gefunden werden, da diese dezentral präsentiert werden und ein Zugang oft nur über die Homepage der Universitätsbibliothek möglich ist. Eine Integration dieser Datenquellen in den Aleph-Katalog ist aus technischer Sicht nicht möglich, da die Daten in heterogenen Formaten vorliegen (MAB2, MARC21, Dublin Core usw.) und der Katalog nur ein einziges Format abbilden kann.

Abbildung von heterogenen Metadaten in Primo

Es ist also der Einsatz einer Software notwendig, die das Laden von bibliografischen Metadaten in heterogenen Formaten sowie auch das Löschen von Titeldaten unterstützt. Diese Funktionen bieten verschiedene lokale Suchmaschinen, deren technische Basis in der Regel eine quelloffene Suchsoftware wie z.B. Lucene ist und die mit den lokalen Bibliothekssystemen allein über Webschnittstellen kommunizieren, ansonsten aber eigenständige Softwaresysteme sind.

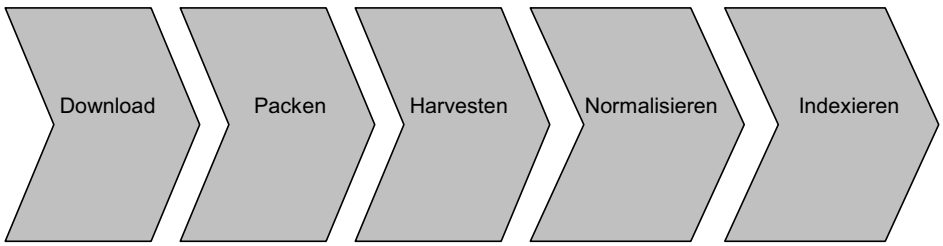
Die Universitätsbibliothek Mannheim setzt seit November 2009 als erste deutsche Bibliothek das von der Firma Ex Libris entwickelte und vertriebene Produkt Primo produktiv ein. Zentraler Einstiegspunkt für eine Recherche ist die Suche über alle indexierten Felder (der „Google-Schlitz“). Die Suchtreffer werden standardmäßig nach Relevanz sortiert, die sich u.a. aus der Häufigkeit und Positionierung der Suchbegriffe im jeweiligen Datensatz errechnet. Weitere Sortiermöglichkeiten sind das Erscheinungsjahr, Titel, Autor und die Popularität, die sich u.a. aus der Klickhäufigkeit des Treffers errechnet. Auf der rechten Seite des Bildschirms sind Facetten zu

sehen, die eine schrittweise Einschränkung der Treffermenge nach konfigurierbaren inhaltlichen und formalen Kriterien erlauben.



(Abb. 1: Suchfenster von Primo)

Alle Daten werden nach dem in Abbildung 2 dargestellten Schema in Primo bearbeitet: Die MAB2- oder MARC-Dateien werden vom Server des Anbieters geladen, in eine Archivdatei gepackt und auf dem Primo-Server abgelegt. Dieses Archiv wird von Primo geharvestet, d.h. im Ursprungsformat geladen und für die weitere Verarbeitung vorbereitet. Anschließend werden die Quelldaten in ein einheitliches Datenformat überführt und angeglichen („normalisiert“). Im letzten Schritt indexiert die Suchsoftware Lucene die manipulierten Daten, die danach für die Recherche in Primo zur Verfügung stehen.



(Abb. 2:Workflow für die Einbindung von Metadaten in Primo)

Dreh- und Angelpunkt des oben beschriebenen Prozesses sind die Normalisierungsregeln, die die Quelldaten bei der Transformation manipulieren, und das dabei zugrunde gelegte Datenformat „Primo Normalized XML“ (PNX). Dieses Format enthält für jeden funktionalen Aspekt, der durch die Suchsoftware realisiert werden soll, eine separate Menge von Feldern. Unterschieden werden u.a. die Bereiche Darstellung, Suche, Sortierung, Verlinkung, Facettierung, Deduplizierung und FRBR. Die Unterfelder sind jeweils an die Anforderungen der Teilfunktion angepasst. So finden sich im Bereich "Darstellung" Felder analog zu der Empfehlung der Dublin Core Metadata Initiative oder im Bereich "Suche" Felder für jeden vom Suchinterface vorgesehenen Schlüssel. Alle Bereiche lassen sich um eigene Felder ergänzen, um lokale Besonderheiten abzubilden und Zusatzfunktionen zu implementieren.
Die Normalisierungsregeln werden mit Hilfe der Primo Publishing Platform festgelegt, die über eine Weboberfläche zugänglich ist. Jeder Datenquelle werden Normalisierungsregeln zugeordnet, die, wenn nötig, individuell für jede Datenquelle gestaltet werden können. Muster für die gängigen bibliografischen Metadatenformate werden vom Anbieter mitgeliefert und dienen als Basis für eigene Anpassungen. Nach den bisherigen Erfahrungen können diese Anpassungen durchaus aufwändig sein, denn die Musterregeln bilden nicht den vollen Umfang der Ausgangsformate ab und können nicht alle Variationen berücksichtigen, die die Auslegung der Katalogisierungsregeln und die Nutzung des Datenformates erlauben. So hat die Universitätsbibliothek Mannheim einen erheblichen Zeitaufwand investiert, um ihre lokalen Titeldaten, die im MAB2-Format vorliegen, vollständig und korrekt nach PNX zu wandeln. Analog zum Laden der Metadaten gibt es in der Publishing Platform die Möglichkeit, Datenquellen mit einem Mausclick temporär von der Suche auszuschließen oder ganz aus dem Index zu löschen.

Primo Central als zentraler Datenspeicher und die Navigation in großen Treffermengen

Bei Primo Central handelt es sich um einen zentralen Datenspeicher für bibliographische Metadaten. Dazu gehören u. a. die Aufsatz- und Buchtitel von EBSCO Business Source Premier, JSTOR, Oxford University Press, Source OECD und Springer. Die Primo-Anwender können Primo Central als zusätzlichen Service abonnieren, in ihren lokalen Systemen auf den externen Suchindex von Primo Central zugreifen und damit ihren Nutzerinnen und Nutzern den Zugriff auf einen umfangreichen Datenpool mit mehr als 100 Millionen Titeln ermöglichen.

Bei einer Suche nach den Stichwörtern "Marketing Management" erhält man in Primo Central ca. 145.000 Treffer. Obwohl sich mit Hilfe der Schlagwort- oder Erscheinungsjahr-Facetten die Suchmenge reduzieren lässt, bleiben immer noch sehr große Treffermengen von mehreren tausend Titeln übrig. Die Nutzerinnen und Nutzer haben bereits mehrfach diese Tatsache bemängelt und um Abhilfe gebeten. Insbesondere Verfasser von Bachelor-Arbeiten haben bei großen Treffermengen Schwierigkeiten, die für sie relevanten Informationen zu finden. Aber auch erfahrene Wissenschaftler benötigen Verfahren, um die für sie relevante Literatur zu filtern und für die eigenen Publikationen zu verwenden.

Wie lässt sich dieses Problem lösen? Ein Ansatz besteht darin, den Nutzerinnen und Nutzern nur die für die jeweilige Fragestellung relevante Literatur anzubieten und redundante Information zu eliminieren. Ein anderer Ansatz ist, den Nutzerinnen und Nutzern jegliche verfügbare wissenschaftliche Publikation anzubieten und über Möglichkeiten nachzudenken, diese Titel nach dem jeweiligen Fokus einzuschränken und zu ranken. Beide Möglichkeiten werde ich im Folgenden am Beispiel des von der Universitätsbibliothek Mannheim eingesetzten Produkts Primo diskutieren.

Die Bedeutung von ECONIS *und der Online Contents für die Recherche*

Folgt man dem ersten Ansatz, stehen mit ECONIS und den Online Contents zwei für die sozial- und wirtschaftswissenschaftliche Ausrichtung der Universität Mannheim wichtige Datenquellen zur Verfügung. Bei ECONIS handelt es sich um den Katalog der Zentralbibliothek für Wirtschaftswissenschaften in Kiel.[1] Dort finden sich 4,4 Millionen Titeldaten von gedruckter und elektronischer wirtschaftswissenschaftlicher Literatur aus aller Welt. Die Datenbank wächst jährlich um etwa 90.000 neue Einträge. ECONIS enthält Bücher und Zeitschriften aus den Bereichen Betriebswirtschaftslehre, Volkswirtschaftslehre und Wirtschaftspraxis. Darüber hinaus werden in ECONIS Zeitschriftenaufsätze und Aufsätze aus Sammelwerken nachgewiesen.

Der Kern der Online Contents (OLC) sind die eingescannten Inhaltsverzeichnisse der Zeitschriftenagentur Swets, die um weitere Aufsatzdaten durch die deutschen Sondersammelgebietsbibliotheken ergänzt werden. Die Online Contents werden vom gemeinsamen Bibliotheksverbund (GBV) gehostet und im Web nach der jeweiligen Fachdisziplin gegliedert angeboten. Die Online Contents umfassen mehr als 30,6 Mio. Aufsatztitel aus über 24.000 Zeitschriften.[2]

[1] http://www.zbw.eu/kataloge/econis.htm.
[2] http://www.gbv.de/vgm/info/benutzer/01datenbanken/01datenbanken_2522?lang=de#info3.

Für diese Datenbank werden seit dem Erscheinungsjahr 1993 Inhaltsverzeichnisse von Zeitschriften aller Fachrichtungen mit besonderem Schwerpunkt auf Naturwissenschaften erfasst.

Allerdings bietet die Universitätsbibliothek Mannheim ihren Nutzerinnen und Nutzern diese beiden Datenquellen nicht vollständig, sondern nur in Auswahl an, die sich nach dem tatsächlich lokal vorhandenen und lizenzierten physischen und elektronischen Bestand richtet. Auf diesem Weg werden die abonnierten Zeitschriften um die in Econis und den OLCs nachgewiesenen Aufsätze angereichert, womit den Nutzerinnen und Nutzern der Universitätsbibliothek Mannheim etwa 700.000 Titel aus Econis und 16.000.000 Titel aus den Online Contents zur Verfügung stehen. Eventuell durch inhaltliche Überschneidungen in beiden Datenquellen vorhandene Dubletten werden durch Primo eliminiert, so dass die Nutzerinnen und Nutzer nur einen Titel in der Trefferliste sehen. Insgesamt stehen in Primo ca. 22 Millionen Datensätze zur Verfügung.

Diese zusätzlichen Datenquellen wurden in Primo integriert, da die Nutzerinnen und Nutzer bei der Einführung im November 2009 monierten, dass die Abbildung der "klassischen" Titeldaten in Primo keinen Mehrwert im Vergleich zum bisher eingesetzten Aleph-Katalog darstelle und die Suche nach Aufsatztiteln ein dringendes Erfordernis sei. In diesem Zusammenhang wurde diskutiert, wie sich die fachliche Relevanz einer Datenquelle und damit das gemeinsame Interesse von Verfassern von Qualifizierungsarbeiten (Bachelorarbeiten, Masterarbeiten und Dissertationen) und Professorinnen und Professoren ermitteln lassen. Dabei ist die Universitätsbibliothek von der Grundannahme ausgegangen, dass die intellektuelle Erfassung der Titel und die Auswahl der Zeitschriften bei ECONIS bzw. deren Ergänzung bei den OLCs durch Erschließungsspezialisten auch mit einer fachlichen Normierung einhergehen und auf diese Weise ein großer Teil des Bedarfs abgedeckt wird.

Diese Annahme wird gestützt, wenn man für die jeweiligen Fachgebiete die Zeitschriften der Online Contents mit den Top-20-Zeitschriften des Journal Citation Index von Thomson Reuters vergleicht. Dabei wurde für die Fachgebiete Economics, Sociology und Political Sciences eine Übereinstimmung von jeweils 80%, 75% und 85% erzielt. Ein vergleichbares Bild ergibt sich, wenn man als Vergleichsbasis die von den Fachreferentinnen und Fachreferenten als wichtig eingeschätzten Zeitschriften nimmt: Bei der Betriebswirtschaftslehre ergibt sich eine Übereinstimmung von 86%, bei der Soziologie von 79%, bei der Politologie von 100% und der

Volkswirtschaftslehre von 80%. Vergleichbare Ergebnisse lassen sich bei der Auswertung des Handelsblattrankings für Betriebswirtschaftslehre und Volkswirtschaftslehre und der am meisten nachgefragten Zeitschriften an der Universität Mannheim ermitteln. Aufs Ganze gesehen decken die Online Contents etwa 80% der wichtigen Zeitschriften ab, die retrospektiv um die Aufsätze der letzten 30 Jahre ergänzt sind.

Verbesserung der Recherchequalität durch Ranking von Treffermengen

Möchte man die Navigation von Nutzerinnen und Nutzern in großen Treffermengen unterstützen, besteht ein zweiter Ansatz darin, das Ranking der Treffer zu optimieren. Primo kennt wie die "klassischen" Kataloge die Möglichkeit, die Suchergebnisse alphabetisch nach Autor und Titel sowie nach Jahr zu sortieren.

Zusätzlich wird ein Ranking der Treffer nach der Relevanz angeboten. Dabei wird berücksichtigt, wie häufig ein oder mehrere Suchwörter in den indexierten Metadaten vorhanden sind. Finden sich beispielsweise die Begriffe "Marketing" und "Management" mehrfach im Titel und den Schlagwörtern, wird der entsprechende Datensatz hoch gerankt. Finden sich dagegen die Begriffe nur einmal in den Metadaten, wird der entsprechende Datensatz niedrig gerankt und in der Treffermenge weit unten angezeigt.

Des Weiteren wird in Primo eine Sortierung nach Bekanntheitsgrad (Popularity) angeboten, die die Treffer nach einem spezifischen Algorithmus anordnet. Beim Klick auf die Vollanzeige eines Titels wird dieser gezählt und die Gesamtmenge mit dem Faktor "fünf" multipliziert. Ebenso werden die Zugriffe auf die Volltexte oder die Verfügbarkeitsanzeige des gedruckten Mediums gezählt und die Gesamtzahl mit dem Faktor "zehn" multipliziert. Wird ein Titel in den als E-Shelf bezeichneten Warenkorb aufgenommen, wird er mit dem Faktor "15" bewertet. Dadurch erhalten die in Primo indexierten Titel jeweils "Punkte", die das Ranking des Titels bei der Trefferanzeige nach dem Prinzip "Popularity" bestimmen.

Weitere Möglichkeiten zum Ranking von Treffern ergeben sich, wenn man in externen Quellen gespeicherte Informationen hinzuzieht, die beispielsweise Google Scholar oder Thomsen Scientific mit ihren Diensten anbieten. Google Scholar wird seit 2005 in einer Betaversion angeboten und hat einen inhaltlichen Schwerpunkt auf frei verfügbaren Volltexten, die Google uneingeschränkt für eine Indexierung zur Verfügung stehen. Bei der Anzeige werden die Treffer nach der Anzahl der jeweiligen Verweise anderer Autoren auf den Volltext gerankt. Sucht man beispielsweise in Google Scholar nach den Begriffen "Marketingmanagement" sind die Titel

"Marketing-Management. Analyse, Strategie, Implementierung" von Heribert Meffert und "Marketingmanagement. Strategie, Instrumente, Umsetzung, Unternehmensführung" von Christian Homburg mit den meisten Zitationen ganz oben sortiert. Bei beiden Werken korreliert das hohe Ranking in Google auch mit der wissenschaftlichen Bedeutung der Werke. Bei dem Autor Christian Homburg handelt es sich um einen Lehrstuhlinhaber der Universität Mannheim, der zu den weltweit führenden Forschern im Bereich Marketing zählt.

Google Scholar bietet eine API für eine externe Abfrage der eigenen Daten an. Damit kann beispielsweise Google Scholar als externer Dienst in eine lokale Suchmaschinen eingebunden werden und die Anzahl der Zitate pro Titel in Google Scholar für eine Ranking der Trefferanzeige in Google Scholar herangezogen werden. Beispielsweise greift die Desktopanwendung "Harzing's Publish or Perish" auf diese API zu und errechnet einen individuellen Zitationswert für jeden Autor. Gezählt werden u. a. die Anzahl der Artikel, der Zitationen und der Zeitraum der Veröffentlichungen. Daraus errechnet sich der sogenannte h-index (auch Hirsch-Index): Ein Wissenschaftler hat einen Index h, wenn h von seinen insgesamt N Veröffentlichungen mindestens jeweils h Zitierungen haben und die anderen (N-h) Publikationen weniger als h Zitierungen.

Explizit für den wissenschaftlichen Bereich bietet Thomson Reuters Dienste an, die Verweise in Publikationen auswerten und nach dieser Auswertung ranken. Journal Citation Index wertet über 8.000 wissenschaftliche peer-reviewed Zeitschriften aus und erstellt gemäß der Anzahl der Verweise eine Rangfolge. Dabei werden die Zeitschriften nach dem sogenannten Impact Factor gerankt, der sich nach der folgenden Formel errechnet: Zahl der Zitate der Artikel geteilt durch die Anzahl der veröffentlichten Artikel einer individuellen Person oder der jeweiligen Zeitschrift. Thomson Reuters bietet für seine Dienste meines Wissens keine API an, die für externe Dienste einen Zugriff auf die eigenen Daten bietet und damit ein Mashup mit den dort hinterlegten Rankinginformationen für wissenschaftliche Zeitschriften bietet. Eine solche API würde ermöglichen, dass eine lokale Suchmaschine wie Primo die Suchergebnisse von Aufsätzen gemäß des Zeitschriftenrankings im Journal Citation Index sortiert.

An dieser Stelle wird deutlich, dass durch den Einzug der Suchmaschinentechnologie in die bibliothekarische Welt und die damit verbundenen Möglichkeiten, große Menge von Titeldaten zu indexieren, ein unbedingtes Erfordernis besteht, die

erzielten Treffermengen zu sortieren und damit den Nutzerinnen und Nutzern die Navigation und Orientierung zu erleichtern. Diese Sortierung kann gemäß der in Primo gezählten Nutzungshäufigkeit der einzelnen Titeldaten oder gemäß der in externen Systemen wie Google Scholar gespeicherten Informationen erfolgen. Es lässt sich auf jeden Fall bilanzieren, dass die großen Treffermengen die praktische Relevanz von bibliometrischen Verfahren wie dem Impact Factor für die Navigation in großen Treffermengen deutlich machen und die unbedingte Erfordernis zeigen, die Daten unterschiedlicher Datenmengen miteinander zu vernetzen.

Fazit

Die Universitätsbibliothek Mannheim hat mit Hilfe der Software Primo eine adäquate Lösung gefunden, um bibliografische Metadaten mit heterogenen Datenaustauschformaten zu indexieren und für die Nutzerinnen und Nutzer unter einer Oberfläche durchsuchbar zu machen. Dazu zählen elektronische und gedruckte Aufsätze aus den unterschiedlichen Datenquellen (eigener Katalog, Econis, Online Contents usw.). In der praktischen Arbeit hat sich herausgestellt, dass die Nutzerinnen und Nutzer häufig vor der Herausforderung stehen, sich in diesen umfangreichen Datensammlungen zu orientieren. Ein Lösungsansatz besteht darin, die angebotenen Titeldaten auf Datenquellen zu beschränken, die einen sehr hohen Anteil von fachlich relevanter Literatur beinhalten und zusätzlich noch intellektuell erschlossen werden. Dies trifft vor allem für die Online Contents und Econis zu. Ein anderer Ansatz sieht vor, den Nutzerinnen und Nutzern einen sehr umfangreichen Datenpool wie Primo Central anzubieten und die Suchergebnisse durch die Anwendung von Algorithmen zu ranken und dadurch die individuelle Recherche zu unterstützen.