Mit dem digitalen Wandel gewinnen kleinteilig vernetzte Informationen immer größere Bedeutung. Dieser Struktur verschließen sich die Angebote der Bibliotheken durch ihre Konzentration auf große Texte in Büchern und Zeitschriften noch weitgehend. Einen neuen Ansatz ermöglicht der semantische Discovery Service Yewno, der den Inhalt von elektronischen Dokumenten mit Methoden der künstlichen Intelligenz und maschinellen Lernens automatisch und auch in Details erschließt. Die Information darüber wird den Nutzern in einer grafischen Darstellung des Netzwerks relevanter Konzepte und ihrer inhaltlichen Zusammenhänge präsentiert. Darüber ist ein Navigieren über thematische Beziehungen bis hin zu den Fundstellen im Text möglich. Die Bayerische Staatsbibliothek stellt diese neue Recherchetechnologie in einem dreimonatigen Beta-Test ihrer Nutzerschaft zur Verfügung, um erste Erkenntnisse über mögliche Anwendungen und Weiterentwicklungsmöglichkeiten der neuen Technologie zu gewinnen. Noch stehen solche Entwicklungen ganz am Anfang, aber für eine tiefere Einbindung der Bibliotheken in die digitale Kultur könnten sie ein wichtiger Baustein sein.

The digital transformation is raising the significance of intricately interlinked information. Library services are still largely neglecting this structure by concentrating on longer texts in books and journals. The semantic discovery service Yewno, which offers automatic and detailed cataloguing of electronic documents using artificial intelligence and automatic learning methods, has opened up the possibility of a new approach. The information is presented to the user in a graphical representation of the network of relevant concepts and their substantive relationships. Furthermore, thematic searches can be made to locate particular points in a text. The Bavarian State Library is making this new search technology available to its users in a three-month beta test phase in order to obtain initial insights into possible applications and opportunities for further development of the new technology. Such developments are still in their infancy, but they could represent important building blocks for integrating libraries more closely into digital culture.

BERTHOLD GILLITZER

Vom Recherchesystem zum inferentiellen Service – ein Paradigmenwechsel?

Yewno, ein semantischer Discovery Service im Pilotversuch an der Bayerischen Staatsbibliothek

Das Auffinden von Informationen über Bücher und in Büchern und darüber, was in den Büchern zu finden ist, stellt fast immer eine schwierige Aufgabe dar, umso schwieriger, je größer die wenigstens theoretisch verfügbare Menge an Literatur ist, für die man diese Frage stellen kann. Bibliothekarinnen und Bibliothekare haben dafür eine ganze Reihe von Lösungen entwickelt, Katalogsysteme mit den zugehörigen Regeln, von Bandkatalogen über Kartensysteme bis hin zu den modernen Datenbanken und netzbasierten Recherchesystemen. Dabei haben sich in den letzten Jahren die Superlative überboten, die neue Dimensionen und Paradigmen des Informationsretrievals versprachen. Am Horizont und als Herausforderung standen hier stets die großen Suchmaschinen im Web, die die dort vorhandene und schon sprichwörtliche Informationsflut scheinbar gut in den Griff bekommen haben.

Zuletzt waren es die sogenannten Discoverysysteme, wie z. B. Primo Central von Exlibris, das auch im OPAC der Bayerischen Staatsbibliothek zum Einsatz kommt, die schon mit ihrer Bezeichnung als Instrumente zum »Entdecken« einen Paradigmenwechsel im Zugang zu Information beanspruchten. Dokumente, die bislang oftmals nur über unterschiedliche Systeme und Suchwege für die Nutzer zugänglich waren, werden hier integriert auffindbar: Aufsätze, Bücher, Zeitungen, Zeitschriften, unabhängig davon, ob als digitale Dokumente oder im Print. Ganz gewiss bedeutet dies eine enorme Verbesserung gegenüber einer Situation, in der eine Nutzerin oder ein Nutzer gezielt zwischen mehreren Bibliothekskatalogen, einer Aufsatz- oder Fachdatenbank wechseln muss, um die relevante Literatur zu finden.

Die Frage, wann man sinnvollerweise von einem Paradigmenwechsel sprechen kann, lässt sicherlich breiten Raum für Diskussionen, und vermutlich kann man den Übergang vom Katalog in Papierform zum elektronischen Katalog auch so betrachten. In einem recht fundamentalen Sinn ist aber auch bei den Discoverysystemen etwas noch ganz beim Alten geblieben: Immer basiert die Suche auf dem Vergleich von Zeichenfolgen. Im Katalog sind Beschreibungen des gesuchten Mediums festgehalten, und die Nutzer müssen diese Beschreibungen in der

Weise auffinden, wie sie im Katalog enthalten sind. Der Computer erleichtert dies sicherlich durch einen schnellen Vergleich der eingegebenen Zeichenfolge und der gespeicherten Zeichenfolge, aber das Prinzip bleibt dasselbe. So gesehen, gab es einen echten Paradigmenwechsel, als nicht mehr primär Bibliotheksbeschäftigte als Mittler zwischen der Nutzerschaft und ihren Literaturwünschen standen, sondern ein schriftlicher oder elektronischer Katalog, den der Nutzer oder die Nutzerin selbst bedient. Mit dem Bibliothekspersonal konnten die Nutzer einen echten Dialog in menschlicher Sprache führen, sich über das Gemeinte verständigen und das bibliothekarische Hintergrundwissen über den Kontext des Buchwunsches nutzen. Es wird noch eine Weile dauern, bis Computer zu einem derartig komplexen Vorgang in der Lage sein werden, auch wenn es sich jetzt nicht genau prognostizieren lässt, wie lange diese Zeit sein wird. Viele Fortschritte hat man noch vor nicht allzu langer Zeit für kaum möglich gehalten, z. B. die nahezu vollständige Digitalisierung der urheberrechtsfreien Bestände der Bibliotheken, die inzwischen zum großen Teil Wirklichkeit geworden ist. Ebenso haben Verfahren der künstlichen Intelligenz in den letzten Jahren große Fortschritte gemacht, Bilderkennung und Bildähnlichkeitssuche hätte man sich vor einigen Jahren genauso wenig vorstellen können wie selbstfahrende Autos. Wann das Recherchesystem zu einem echten Dialogpartner der Nutzer werden wird, ist aber noch eine offene Frage.

Von der Schriftkultur zur digitalen Kultur

Zugleich ist damit aber auch schon ein anderer Paradigmenwechsel angesprochen, der bislang in Bibliotheken erst in Ansätzen mitvollzogen wurde und der beim Blick auf Bibliothekskataloge und Recherchetechnologien eine nicht unerhebliche Rolle spielt. Lobin[1] reklamiert einen solchen Wandel für den Übergang von der Schriftkultur zur digitalen Kultur. Dabei nennt er drei grundlegende technische Triebkräfte – Automatisierung, Datenintegration und Vernetzung –, welchen in gewisser Hinsicht drei Tendenzen im kulturellen Wandel beim Umgang mit geschriebener Information entsprechen: Hybridität, Multimedialität und Vernetzung. Die Schriftkultur, was Lobin mit McLuhan die Gutenberg-Galaxis nennt, wird abgelöst durch eine digitale Kultur, die Turing-Galaxis, benannt nach Alan Turing, dem ersten Theoretiker eines universalisierten Computers.[2]

Die Ablösung der Schriftkultur bedeutet aber nicht einfach den Ersatz von etwas Altem – der schriftlichen Information – durch etwas vollkommen anderes. Der Blick allein auf die Vielzahl von Menschen, die unablässig in mobile Endgeräte tippen und darin lesen, genügt, um zu zeigen, dass geschriebenes Wort nach wie vor eine immens große Rolle in unserer Kultur spielt. Andere Elemente sind hier neu, von denen nur zwei hervorgehoben seien: Schrift spielt nicht mehr in erster Linie im Printformat (oder gar handschriftlich) eine Rolle, sondern als digitale Medien, bei welchen der Computer sowohl in der Entstehung als auch bei der Rezeption eine aktive Rolle spielt. Lobin spricht in diesem Zusammenhang von hybridem Schreiben und hybridem Lesen.[3] Dies hat mehrere Dimensionen, wobei die Abhängigkeit von Programmen und Hardware, die zunächst digitale Texte grundlegend ermöglichen (Schreibprogramme zur Erstellung von Texten und deren Anzeige, Browser usw.), nur einen geringen Faktor darstellt, der noch keinen grundlegenden Kulturwandel markiert. Lobin macht eindringlich darauf aufmerksam, dass erst mit dem Hypertext, der Möglichkeit der Verlinkung von Textteilen und noch mehr von Texten untereinander, ein neues Lesen entsteht, in dem der Leser oder die Leserin nicht mehr linear einem Text folgt, sondern selbst entscheidet, welchem Link gefolgt wird und welche Informationen damit rezipiert werden.[4] Das »Aufbrechen dieser Linearität«[5], wie Ceynowa diese Veränderung bezeichnet, hat dabei nicht nur die Dimension, dass Textteile und Texte verknüpft werden und ein Springen zwischen diesen Elementen möglich wird. Vielmehr werden auf diese Weise Texte auch mit nicht textuellen Elementen wie Bildern, (interaktiven) Grafiken, Forschungsdaten, Tabellen usw. verknüpft, die ihrerseits Information und Wissen vermitteln, aber auf eine andere Weise als linear zu lesende Texte.[6]

Die Konsequenzen einer sich in dieser Art wandelnden Kultur und der Rolle von Texten darin, die häufig kontrovers diskutiert und nicht selten mit Skepsis betrachtet werden, bringt Ceynowa auf den Punkt, wenn er schreibt, dass der Text in diesem System nur noch ein Element ist, »und nicht einmal das Wichtigste: Im Grunde fungiert er lediglich als dokumentierende Momentaufnahme in einem vernetzten, dynamischen Wissensraum.«[7]

Fragmentierung als Merkmal der digitalen Kultur

Wichtig scheint an der Stelle Folgendes: Wenn wir in einem solchen vernetzten Wissensraum Texte als eine Art von Knotenpunkten zwischen einer Vielzahl von diversen Informationseinheiten betrachten, so sprechen wir nicht von den Einheiten, die wir derzeit z. B. in unseren Bibliothekskatalogen als Werke verzeichnet haben. Ein wesentliches Element unserer digitalen Kultur, das mit dem zuvor Dargestellten einhergeht, scheint dem Verfasser eine Fragmentierung zu sein, bei der Information und Wissen in wesentlich kleineren Einheiten wahrgenommen und verarbeitet werden, als das früher der Fall war oder zumindest der Fall zu sein schien.[8] Viele traditionell größere Texteinheiten werden entweder gleich in kleineren Teilen verfügbar oder auch situativ »aufgebrochen«, wie Ceynowa dies ausdrückt, was besonders für per se portionierbare Inhalte, wie Reiseführer, Rezeptbücher usw. zutrifft.[9] Dies gilt m. E. in

gewisser Hinsicht auch für die Formate, die sich laut Ceynowa der Verwandlung von Texten in multimedialen Content widersetzen, weil sie narrativ angelegt sind: Erzählungen, Romane, geisteswissenschaftliche Sachbücher.[10] Für die Entstehung dieser Werke, ihre Anlage und Gesamtkomposition ist ihr Verständnis als große ganzheitliche Texte gewiss essentiell, nicht mehr aber in jedem Fall für ihre Rezeption und Verarbeitung in anderen Texten und alternativen Formen der Wissensrepräsentation. Freilich sind diese Werke auf eine ganzheitliche Rezeption angelegt, und oftmals werden die Bücher als Ganze rezipiert werden. Aber auch hier gibt es den fragmentierten Zugang und die fragmentierte Auseinandersetzung: Romanteile, einzelne Szenen, Figuren werden in anderen Texten, Videosequenzen oder bildlicher Darstellung aufgegriffen, einzelne Thesen aus größeren geisteswissenschaftlichen Werken zur Geschichte, Philologie oder Philosophie herausgegriffen und in neue Kontexte eingebettet, dort einzeln diskutiert und abgewandelt.

Konsequenzen aus Fragmentierung und Diskontinuität für Bibliotheken

Diese Fragmentierung der Texte und die damit einhergehende Diskontinuität unserer Wissensrezeption ist immer wieder auch Teil einer kritischen Diskussion, die deren negative Folgen hervorhebt. Paradigmatisch mag hier der Artikel »Is Google Making Us Stupid?« von Nicholas Carr[11] aus dem Jahr 2012 genannt werden, der Ablenkung und Oberflächlichkeit in der Auseinandersetzung mit Inhalten als Folgen benennt. Auf diese Auseinandersetzung soll an der Stelle nicht näher eingegangen werden, da sich gegenwärtig hier auch noch keine abschließende Bewertung absehen lässt.[12] Die Möglichkeit negativer Folgen in bestimmten Kontexten, wie sie manche Untersuchungen nahelegen, die eine geringere Verweildauer bei Texten und ein vermindertes Erinnern der gelesenen Inhalte feststellen, möchte der Verfasser gar nicht bestreiten.[13] Die Abkehr von digitalen Medien in Bibliotheken zugunsten von Printmaterial, wie es Nancy McCormack nahelegt, scheint aber keine sinnvolle Reaktion auf die angenommenen Probleme darzustellen.[14] Es ist wohl kaum zu erwarten, dass Bibliotheken, würden sie sich aus der Digitalisierung zurückziehen, ein namhaftes Gegengewicht zu den angenommenen negativen Folgen des Internets und des digitalen Contents bilden könnten. Wenn wir tatsächlich die digitale Durchdringung unserer Gesellschaft als einen Kulturwandel, einen Paradigmenwechsel betrachten können, hätte ein Rückzug der Bibliotheken aus dem Prozess wohl schlicht zur Folge, dass sie und die von ihnen bewahrte Schriftkultur an Bedeutung verlieren würden und zum musealen Relikt verkämen.

Im vorliegenden Kontext möchte der Verfasser deshalb eine andere Konsequenz als möglicherweise sinnvollere Reaktion darstellen: In den vergangenen Jahren wurde vielfach der Wandel der Bibliotheken hin zur hybriden Bibliothek thematisiert, teilweise auch bereits die erfolgreiche Umsetzung beansprucht.[15] Vornehmlich digitale Texte (aber auch andere digitale Objekte, Bilder, Tondokumente usw.) werden heute tatsächlich weitgehend gleichberechtigt zu ihren analogen Gegenstücken in den Bibliotheken zur Nutzung angeboten, gewiss ein wichtiger und entscheidender Schritt hin zur digitalen Kultur. Gegenüber der durch Fragmentierung und zugleich durch stärkere Verknüpfung geprägten digitalen Lebenswelt bleibt aber auch dieser hybride Korpus in einem gewissen Ausmaß fremd. Die in den Bibliotheken verfügbaren Informationen werden nicht in der gleichen Weise verfügbar, wie es viele Informationen im Web sind, weil nur die großen Texte, die Bücher und Zeitschriftenartikel als Ganze zugänglich sind, nicht aber direkt die in ihnen enthaltenen viel kleinteiligeren Informationen, auf die es in der vernetzten digitalen Welt ankommt.

In diesem Sinn hat m. E. bislang noch kein Paradigmenwandel vom Katalogsystem zum Discovery Service stattgefunden. Noch immer verzeichnen die Systeme eben vornehmlich Bücher, Zeitschriften, Zeitungen und Aufsätze, aber sie gehen noch nicht in diese Medien hinein, mögen sie nun als Printmedien oder digitale Dokumente vorliegen. Erst wenn die Discovery Services die Informationen in den Dokumenten direkt auffindbar machen, die einzelnen Dokumente thematisch in ihren Teilen für die Nutzung verfügbar werden, können diese Teile auch in der stärker fragmentierten Form des Umgangs mit Informationen, wie er für die digitale Kultur charakteristisch ist, eine Rolle spielen. Die oftmals nur bruchstückhaft vorhandene Volltextsuche schafft dabei nur eine geringe Abhilfe. Es ist mehr ein Zufall, ob eine Nutzerin oder ein Nutzer mit dem eingegebenen Suchbegriff eine Textstelle auffindet, die für das thematische Interesse wirklich relevant ist. Eine tatsächlich wichtige Textstelle könnte auch nur mit anderen Wörtern – anderen Zeichenfolgen also – auffindbar sein. Werden die Textkorpora sehr groß, kommt darüber hinaus das Mengenproblem zum Tragen, weil nicht nur eventuell Relevantes nicht gefunden wird, sondern auch zu viele nicht relevante Treffer das Auffinden des Gewünschten behindern. Die Lösung des Problems von Synonymie und Polysemie bleibt der intellektuellen Kapazität der Nutzer überlassen, die mit oftmals vielen hunderten oder tausenden Treffern in jedem Fall überfordert sind. Ein effektives Relevanzranking steht in Bibliothekskatalogen oder Discovery Services bislang ebenfalls nicht zur Verfügung und erscheint auch speziell problematisch bei gemischten Treffermengen aus Volltexttreffern und Treffern allein auf der Basis von Metadaten, wie sie bisher üblich sind.

Semantik als Lösungsansatz – Yewno, ein semantischer Discovery Service

Erst wenn Discovery Services semantisch werden, also wissen, an welcher Stelle es in einem Werk um ein

bestimmtes Thema geht, können diese Defizite ausgeglichen werden. Dieser fundamental neue Schritt wird nun mit dem »semantischen Discovery Service Yewno« des gleichnamigen kalifornischen Startup-Unternehmens versucht.[16]

Yewno hat zwei Grundpfeiler, die den angesprochenen Paradigmenwechsel markieren. Der erste ist die Extraktion von Konzepten aus elektronischen Volltexten. Ganz grundsätzlich gibt es dazu auch an anderer Stelle schon Ansätze. Die Verfahren werden mit dem Stichwort Datamining bezeichnet, wobei versucht wird, sinntragende Begriffe in Texten zu identifizieren und gewissermaßen als Schlagwörter zur Verfügung zu stellen.[17] Meist werden hier aber nur Begriffe bestimmter Kategorien ermittelt, wie Personen, Orte und Ereignisse. Zum anderen basieren die bislang tatsächlich eingesetzten Verfahren üblicherweise auf Ontologien, Thesauri oder Wörterbüchern, einer statischen Datenbasis also, deren Begriffe dann mit computerlinguistischen Verfahren bestimmten Textstellen oder sogar nur ganzen Werken zugeordnet werden.[18] Von solchen Projekten unterscheidet sich Yewno dadurch, dass nicht zuerst auf eine systematische und oftmals auch hierarchisch organisierte Datenbasis zurückgegriffen wird. Vielmehr basiert es auf linguistischer Analyse der Texte mittels künstlicher Intelligenz und maschinellem Lernen.[19]

Yewno arbeitet mit sogenannten Konzepten, die definiert sind als Mengen von Wörtern mit gleicher Bedeutung, gewissermaßen also reine Bedeutungsentitäten im Unterschied zu ihren verbalen Expressionen. Diese Konzepte haben so betrachtet Ähnlichkeit mit klassischen Schlagworten, denen ja auch unterschiedliche Schreibweisen oder Synonyme zugeordnet sein können. Anders als Schlagworte sind sie aber gerade nicht in einer Normdatei oder einer Datenbasis vorgegeben. Konzepte werden vielmehr durch diese Verfahren der statistischen Semantik extrahiert und bestimmten Textstellen in digitalen Dokument, wo es um diese Konzepte geht, zugeordnet. Die statistisch-semantischen Verfahren beruhen auf der sogenannten »distributional hypotheses«, wonach Wörter/Konzepte mit ähnlicher Bedeutung in ähnlichen Kontexten vorkommen.[20] Das heißt, dass es eine Korrelation zwischen ähnlicher Verteilung von Wörtern in begrenzten Kontexten und ähnlicher Bedeutung gibt, die es gestattet vom einen auf das andere zu schließen. Erfolgreich wurde dieses Verfahren durch den Gründer von Yewno, Ruggero Gramatica, schon im biomedizinischen Kontext zur Anwendung gebracht, wo es darum ging, noch unbekannte Beziehungen zwischen Peptiden auf der Basis von deren Erwähnung in biomedizinischen wissenschaftlichen Artikeln aufzudecken, mit dem Ziel, Zweitanwendungsmöglichkeiten der Verwendung von Medikamenten für seltene und wenig erforschte Krankheiten zu finden.[21]

Auf die Feinheiten und technischen Details kann hier nicht eingegangen werden, wie z. B. Probleme der Polysemie und Synonymie gemeistert werden. Letztlich liegt in diesen technischen Lösungen ja auch das Betriebskapital von Yewno, ähnlich wie bei den Relevanzalgorithmen der Suchmaschinenbetreiber, die deshalb nicht einfach öffentlich zugänglich gemacht werden können. Umgesetzt werden diese theoretischen Ansätze mit einer auf künstlichen neuronalen Netzwerken basierenden Struktur, die auf den lokalen Kontext eines Wortes trainiert wird. Wörter werden in einen Vektorraum mit fixer Kardinalität übersetzt, bei der semantisch verwandte Wörter nahe zusammen gruppiert sind. Anders als in anderen Ansätzen stehen bei Yewno die Vektoren allerdings nicht für Wörter, sondern für »wohlgeformte Konzepte«.

Auf dieser technischen Basis gewinnt Yewno nicht nur die Konzepte in den verarbeiteten elektronischen Dokumenten, sondern auch Kenntnisse über die semantischen Relationen zwischen den Konzepten. In der Verknüpfung der Konzepte zeigt sich aber nochmals ein wichtiger Unterschied zu anderen Verfahren, da sich diese Konzeptbeziehungen nicht hierarchisch darstellen, sondern gewissermaßen assoziativ, multidimensional vernetzt. Zu einem Konzept gehören Eigenschaften (wiederum Konzepte), und unterschiedliche Konzepte teilen Eigenschaften, über die sie verknüpft sind. Gramatica erläutert dies an einem einfachen Beispiel: Konzepte können sehr eng verknüpft sein, wenn sie (gemäß der »distributional hypotheses«) sehr nahe zusammen liegen (z. B. Shiraz und Merlot, zwei Weine), sie können dann aber auch über abstraktere Eigenschaften und Abhängigkeiten verknüpft sein, z. B. darüber, dass beide eine ähnliche Farbe haben. Über diese Methode lässt sich auch eine Stärke der Verknüpfung zwischen Konzepten feststellen.

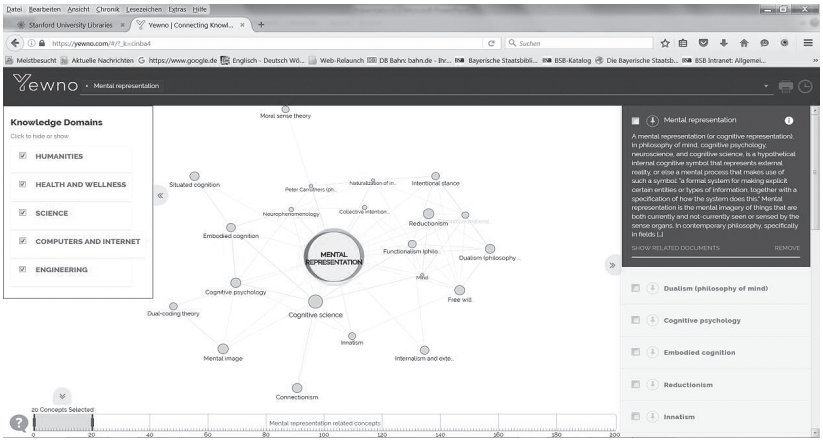
Eine weitere, auf das Verfahren der Konzeptextraktion bezogene Besonderheit ist die Einbeziehung der zeitlichen Dimension. Es werden nicht nur die Konzepte und ihre Beziehungen analysiert, sondern auch deren Entwicklung über die Zeit hinweg, insofern sich die Konzepte in unterschiedlichen zeitlichen Kontexten unterschiedlich darstellen: durch abweichende Eigenschaften oder Verknüpfungen in verschiedenen Kontexten. Noch ein weiterer damit in Zusammenhang stehender Vorteil dieses Vorgehens, das auf eine fixe Datenbasis als Ausgangspunkt für das Datamining verzichtet, ist die Tatsache, dass damit auch die Basis für dieses Verfahren nicht veralten kann. Wird das Verfahren auf Texte über neue Theorien mit entsprechend neuen Konzepten angewandt, werden diese automatisch Teil des Konzeptnetzwerks, unabhängig davon, ob solche Konzepte in einer Wissensbasis bereits vorhanden sind oder nicht.

Von der Suchmaschine zum inferentiellen Service

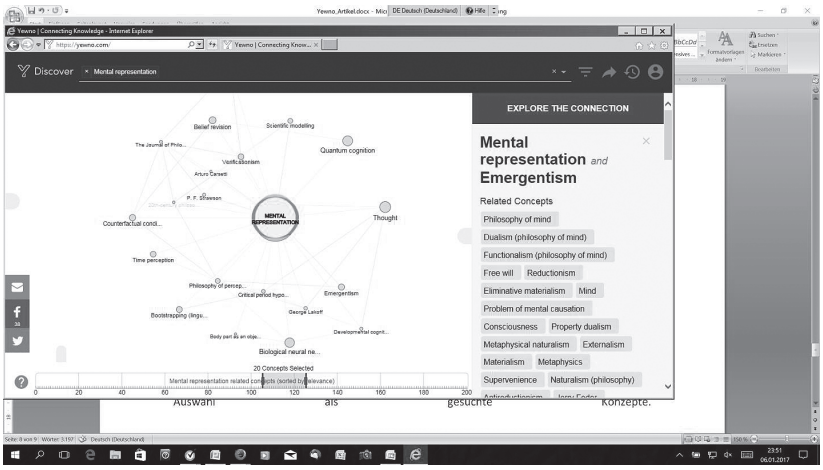
Der zweite Grundpfeiler ist die spezielle Weise, wie Yewno die Konzepte und ihre Relationen visualisiert. Die Konzepte werden zwar aus den elektronischen Texten gewonnen, aber die Recherche und ihre Darstellung

in einem semantischen Netz erfolgt zunächst unabhängig davon. Als Ergebnis der Recherche werden den Nutzern die Konzepte mit ihren semantischen Eigenschaften direkt visualisiert: Ein gefundenes Konzept wird im Netz mit verknüpften Konzepten als großer farblich orange markierter Punkt im Zentrum dargestellt. Am rechten Rand des Bildschirms erscheint eine kurze Erklärung zu dem Konzept, oftmals aus Wikipedia. Alle Konzepte sind mit ihren verknüpften Konzepten, die wiederum als Punkte dargestellt werden, über Linien verbunden. Die Größe der Punkte gibt dabei Aufschluss darüber, wie stark das Konzept mit dem Ausgangskonzept verbunden ist. Die Anordnung der Einheiten im Netzwerk lässt sich durch Klick auf einen Punkt und Ziehen des Punktes an eine andere Stelle verändern und so auf die Nutzerinteressen anpassen. Durch einen Schieberegler am unteren Rand des Bildschirms lässt sich nicht nur die Anzahl der angezeigten verknüpften Konzepte variieren, derzeit zwischen 5 und 20 je gesuchtem Konzept. Es lässt sich durch ein Verschieben des Reglers von links nach rechts auch bestimmen, ob man nur verknüpfte Konzepte angezeigt bekommen will, die sehr eng mit dem initial gesuchten Begriff verbunden sind (der Regler steht links), oder solche, die in einer weiteren und abstrakteren Beziehung zu diesem Ausgangsbegriff stehen (der Schieberegler steht weiter rechts).

Die bekannte Art auch der thematischen Suche wird damit grundsätzlich verändert: Es wird nicht eine lange Ergebnisliste primär präsentiert, sondern das gesuchte Thema vernetzt in seinen vielfältigen sachlichen Bezügen dargestellt. Nicht, dass es zuvor nicht auch schon Recherchemöglichkeiten mit thematischem Browsing gegeben hätte. Klassischerweise handelte es sich dabei aber um ein Navigieren über hierarchische Klassifikationen. Die multidimensionale Vernetzung und die Möglichkeit, gerade auch gezielt die Verknüpfung zu »weiter



1 Ein gesuchtes Konzept im Netz der damit verknüpften Konzepte mit Erklärung zur Bedeutung des Konzepts



2 Gezielt können im semantischen Netz Konzepte ausgewählt werden, deren Beziehung zum Ausgangskonzept schwächer ist, dafür aber vielleicht neue und unerwartete Zusammenhänge offenbart.

entfernten« Begriffen aufzusuchen, ermöglicht ein deutlich intuitiveres Navigieren und kann zum Entdecken von Zusammenhängen führen, die den Nutzern so nicht bekannt waren und die in einer Klassifikation beispielsweise auch gar nicht verfügbar wären. Das Browsen soll damit bewusst an intuitives menschliches Schlussfolgern angepasst werden. Eigentlich kann erst ein System mit dieser Architektur Anspruch auf die Bezeichnung Discovery Service erheben. Nach Ruggero Gramatica sollte deshalb die Anwendung nicht als Suchmaschine betrachtet und bezeichnet werden, sondern als schlussfolgernde Anwendung, als »inference engine«[22], die komplementär zu traditionellen Suchmaschinen und Bibliothekskatalogen konzipiert ist.

Entscheidender Aspekt dieser Form der Visualisierung und Nutzerführung ist die Möglichkeit, stets weitere Informationen zum Kontext des Suchergebnisses zu erhalten. Nicht nur Erläuterungen zum jeweils einzelnen Konzept sind verfügbar: Ein Klick auf die Linie zwischen zwei Konzepten führt zur Anzeige von Informationen über die Verknüpfung am rechten Bildschirmrand (definiert durch die gemeinsamen Konzepte). Das einzelne Konzept kann auch aus dem multipel verknüpften Netzwerk mit einem Klick herausgehoben werden. Es wird dann nur dieses und die ihm zugeordneten Konzepte angezeigt, ohne deren weitere Querverbindungen.

Wird mit Doppelklick auf ein verknüpftes Konzept dieses zusätzlich ausgewählt, wird die Suche expandiert. Beide Konzepte (oder auch mehrere) werden mit ihren verknüpften Konzepten quasi als Konzeptraum dazwischen groß dargestellt, es entsteht ein erweiterter semantischer Raum, mit einem erweiterten Kontext. Dieses Angebot kontextueller erweiternder Sachinformation entspricht zwar noch nicht dem Dialog mit einem sachverständigen menschlichen Gesprächspartner, aber es nimmt Teile davon auf und geht zugleich auch darüber hinaus, indem es ggf. Beziehungen sichtbar werden lässt, die aufgrund der Vielzahl an Dokumenten nur im maschinellen Verfahren aufgedeckt werden können.

Zu guter Letzt enthält Yewno noch einen mehr oder weniger konventionellen Teil, den Weg von den angezeigten Konzepten zu den damit verknüpften Dokumenten: Auf der rechten Seite des Bildschirms können unterhalb der Erklärung zu den gesuchten Konzepten die zugehörigen Dokumente aufgerufen werden, in denen es um die jeweiligen Konzepte geht. Die zugeordneten Dokumente werden untereinander mit Textteaser angezeigt, die direkt die Textstellen zeigen, welchen das jeweilige Konzept zugeordnet ist. Die Nutzer werden damit eben nicht nur auf ganze Werke geführt, sondern zu den Teilen, die für ihr thematisches Interesse relevant sind. Ein weiterer Link führt dann auf den jeweiligen Volltext, entweder einen im Open Access vorhandenen elektronischen Volltext oder ein von der Bibliothek lizenziertes Dokument. Um sicherzustellen, dass es sich wirklich um ein von der Bibliothek lizenziertes Dokument handelt, findet ein Abgleich mit den Lizenzinformationen der jeweiligen Bibliothek statt, z. B. über deren SFX Knowledgebase.

Pilotbetrieb an der Bayerischen Staatsbibliothek

An der Bayerischen Staatsbibliothek wird die Anwendung den Nutzern zunächst in einem dreimonatigen Beta-Test zur Verfügung gestellt, in dem momentan vorhandenen Standard. Durchsucht werden also die elektronischen Dokumente, die von Yewno bislang indexiert und verarbeitet wurden, eine große Menge von wissenschaftlichen Zeitschriftenartikeln, die im Open Access vorliegen, aber auch elektronische Volltexte aus lizenzpflichtigen E-Books und Zeitschriftenartikeln, für die Yewno eine entsprechende Vereinbarung mit den jeweiligen Verlagen abgeschlossen hat. Eigener Bestand an elektronischen Volltexten der Bibliotheken, z. B. aus Retrodigitalisierungsprojekten kann von Yewno zusätzlich einbezogen werden. Bislang sind es allein englischsprachige Dokumente, die von Yewno verarbeitet wurden und im System angeboten werden können. Weitere west- und osteuropäische Sprachen sollen aber folgen, die »inference engine« auch multilingual angeboten werden.

Ziel dieses Beta-Tests ist es, ein erstes Feedback von den Nutzern über diese neue Form von Recherche, dieses vollkommen andere Herangehen an die thematische Suche zu bekommen. Zugleich ist es notwendig, eine Vorstellung zu entwickeln, wie diese neue Technologie zukünftig vielleicht einmal ein Teil des Standardservices der Bibliothek werden könnte. Noch steht die Technik am Anfang, und es werden gewiss noch einige Entwicklungsschritte notwendig sein, um sie im Umfeld der Bibliotheken zu etablieren und vollständig nutzbar zu machen.

Selbst für den Beta-Test an der Bayerischen Staatsbibliothek mussten bereits Anpassungen vorgenommen werden, die über den Abgleich der Lizenzdaten aus SFX hinausgehen. Auch ohne eine persönliche Anmeldung bei Yewno wird bei der Recherche die IP-Adresse des Nutzers oder der Nutzerin vom Browser wie üblich übermittelt. Da Yewnos Server in den USA stehen, wird das Angebot an der Bayerischen Staatsbibliothek komplett über den bibliothekseigenen HAN Proxy[23] abgewickelt. Die Nutzer bleiben gegenüber Yewno also vollständig anonym, und der Service mit deutschem Datenschutzrecht vereinbar. Mindestens genauso wichtig ist die zweite Anwendung des Proxyservers beim Aufruf der elektronischen Volltexte. Hier verlinkt Yewno bei lizenzpflichtigen Dokumenten im Standardfall zum Angebot des Verlags, was bei externem Zugang (also außerhalb der Räume der Bayerischen Staatsbibliothek) für die meisten Nutzer der Bibliothek keinen Zugriff auf das Dokument ermöglicht. Für eine Nutzung direkt aus dem Netz der Bibliothek in deren Lesesälen ist dies kein

Problem, da die IP-Adressen der Bibliothek den Verlagen bekannt sind und diese, bei einer entsprechenden Lizenz, den Zugriff auf die elektronischen Volltexte dann erlauben. Für den externen Zugang musste Yewno seine Anwendung anpassen. Die Links führen jetzt zu einer Anmeldeseite der Bibliothek, bei der sich die Nutzer mit ihrer Bibliothekskennung authentifizieren können und dann den Zugriff auf die von der Bibliothek lizenzierten Dokumente über den Proxyserver der Bayerischen Staatsbibliothek erhalten.

Anwendungsperspektiven

Abschließend seien für die weitere Entwicklung und die Perspektiven eines produktiven Einsatzes drei Aspekte erwähnt:

Weiter oben wurde bereits angemerkt, dass Yewno bislang nur auf englischsprachige Texte Anwendung findet. Für europäische Bibliotheken ist damit die Datenbasis zu schmal, aber letztlich nicht nur für europäische Bibliotheken. Forschung ist prinzipiell international angelegt und multilingual. Keine große wissenschaftliche Forschungsbibliothek wird für ihre Nutzer nur mit englischsprachigen Dokumenten auskommen. Deshalb ist bei Yewno die Integration weiterer Sprachen in konkreter Planung. Erst dieser wichtige Entwicklungsschritt wird die Anwendung für europäische Bibliotheken voll nutzbar machen.

Der zweite Aspekt künftiger Weiterentwicklung ist weniger konkret, auch wenn der Bedarf offensichtlich ist. Yewno schließt eine Lücke, indem es den Paradigmenwechsel der digitalen Kultur, der eingangs dargelegt wurde, für Texte mitvollzieht. Dies kann aber nicht heißen, dass es beziehungslos neben die bisherigen Bibliotheksangebote tritt oder den Anspruch erhebt, diese zu ersetzen. Nicht alle elektronischen Volltexte werden von Yewno bislang verarbeitet und der riesige Bestand, der nur in Printform vorliegt, wird durch die neuartige »inference engine« auch nicht unbedeutend. Es stellt sich damit die Frage, wie Yewno mit den bereits vorhandenen Bibliotheks- und Katalogangeboten integriert wird. Ein erster Schritt besteht darin, das parallele Printangebot der Bibliothek beim Verweis auf die Dokumente über den Bibliothekskatalog zugänglich zu machen, wo bei Yewno zwar die elektronische Ressource verarbeitet ist, die Bibliothek aber keine Lizenz für den elektronischen Volltext besitzt, dafür aber die Printversion des entsprechenden Artikels oder Buches. Dazu muss Yewno die Katalogdaten der Bibliothek auf parallele Printformen hin mit den bei sich vorliegenden und verarbeiteten Volltexten vergleichen und für das Angebot eines Kataloglinks zuordnen. Die Nutzerin oder der Nutzer kommen dann zwar nicht vom Konzept mit einem Klick zum elektronischen Volltext, aber immerhin zum Nachweis der Printform im Katalog. Dieser Entwicklungsschritt könnte noch innerhalb der Testphase an der Bayerischen Staatsbibliothek vollzogen werden, Yewno arbeitet daran. Die Frage, wie Yewno im Verhältnis zu den dort gar nicht erfassten Inhalten angeboten wird, bleibt damit aber noch offen. Man könnte an eine eigene Rechercheoption und Präsentationsschicht, integriert in den OPAC, denken. Der Bruch zum großen reinen Printbestand wäre aber erst dann beseitigt, wenn dieser letztlich zumindest rudimentär in diese besondere Form der Suche und Präsentation einbezogen werden könnte. Ob es gelingen kann, wenigstens zu Zwecken der Indexierung und Recherche für den bisherigen Printbestand elektronische Volltextdaten zu gewinnen oder die Technik von Yewno in irgendeiner Weise auf reine Katalogdaten anzuwenden, ist bisher aber eine offene Frage. Man könnte an die Einbeziehung vorhandener sachlicher Erschließung im Katalog und die Anreicherung mit weiteren Informationen über Linked Open Data denken. Aber noch ist das ein Feld kommender Forschung.

Eine letzte Option ist in Yewno schon angelegt: Die Technik wurde bereits erfolgreich im biomedizinischen Bereich eingesetzt, und derzeit ist auch ein spezialisierter Einsatz im Finanzsektor in Vorbereitung.[24] Yewno könnte auch eines der Tools werden, das Bibliotheken als Partner von Forschung und Wissenschaft in spezialisierten Bereichen zum Einsatz bringen. Ein mögliches Szenario wäre die Anwendung von Yewno beispielsweise in den Fachinformationsdiensten nicht nur zu einer vertieften Erschließung von Spezialbeständen für die Forschung, sondern eben mit der besonderen Fähigkeit, mit dieser vertieften Erschließung durch die inferentiellen Funktionen in Yewno, auch neue und unerwartete Zusammenhänge für die Wissenschaft sichtbar zu machen. Für kleinere spezialisierte Sammlungsteile besteht im Kontext der Fachinformationsdienste auch eine größere Wahrscheinlichkeit, dass sich der Aufwand von Digitalisierung mit einer qualitativ hochwertigen OCR-Bearbeitung und ggf. noch Rechteeinwerbung bei deren Inhabern realisieren lässt. Damit könnte forschungsrelevante Literatur, die nicht per se als elektronischer Volltext vorliegt, für eine thematisch vertiefte Erschließung gewonnen werden, was aus Aufwandsgründen gewiss nur für kleine und hochspezialisierte Teile einer Bibliothek, derzeit nicht aber für ihren gesamten Bestand realisierbar wäre. Yewno könnte damit Teil einer spezialisierten Forschungsumgebung sein, ein Teil, den die Bibliotheken im Rahmen ihrer forschungsnahen Dienste beitragen.

Der Ausblick auf die Entwicklungsperspektiven zeigt, dass vermutlich noch ein guter Weg zu gehen ist, bis der Einsatz einer »inference engine«, wie Yewno, zum Standardservice der Bibliotheken werden wird. Skeptiker können infrage stellen, ob sich ein solcher Aufwand lohnen wird. Im Hinblick auf die eingangs angestellten Überlegungen zum Paradigmenwechsel von der Schriftkultur zur digitalen Kultur scheint es für Bibliotheken m. E. eher angeraten, diese Entwicklung aktiv mitzuverfolgen und mitzugestalten. Semantische und inferen-

tielle Rechercheangebote, ob sie nun von Yewno kommen oder zukünftig auch von anderen Anbietern bereitgestellt werden, mögen den Bibliotheken nicht alleine in der sich schnell wandelnden digitalen Welt die Zukunftsfähigkeit garantieren, aber sie könnten ein entscheidender Baustein dafür sein.

Anmerkungen

1 Vgl. Lobin, Henning, Engelbarts Traum, Frankfurt a. M. 2014, Kap. 4.4 und Kap. 4.5.
2 Vgl. Lobin, 2014, Kap. 4.2.
3 Vgl. Lobin, 2014, Kap. 5.2 und Kap. 6.2.
4 Vgl. Lobin 2014, S. 99.
5 Ceynowa, Klaus: Der Text ist tot – es lebe das Wissen. In: Hohe Luft: Philosophie-Zeitschrift (1), 2014, S. 53–57, S. 54.
6 Vgl. Ceynowa 2014, S. 54 f.
7 Ceynowa 2014, S. 54.
8 Schon früher wurden oft nur Teile von Werken be- und verarbeitet, gerade auch im wissenschaftlichen Kontext, doch waren diese Teile nicht separat verfügbar.
9 Ceynowa 2014, S. 55.
10 Vgl. Ceynowa 2014, S. 55.
11 Carr, Nicholas, Is Google Making Us Stupid? In: Atlantic Monthly, 301(6) 2012. Verfügbar unter: https://www.theatlantic.com/magazine/archive/2008/07/is-google-making-us-stupid/306868/
12 Fragmentierung und Diskontinuität werden zumeist als neue Eigenschaft bei der Rezeption und Auseinandersetzung mit geistigen Inhalten durch das »Medium« Internet betrachtet, mit der gleichzeitig ein Verfall einhergeht hinsichtlich der Tiefe und Intensität der Auseinandersetzung. Carr geht dabei so weit, dass er hier sogar eine Absicht der großen Player wie Google annimmt: »The last thing these companies want, is to encourage leisurely reading or slow, concentrated thought. It's their economic interest to drive us to distraction.« (Carr, 2012). Wenig Beachtung findet m. E. bislang, dass Fragmentierung und Diskontinuität schon immer Eigenschaften menschlicher Informationsübermittlung sind, wie sie in alltäglichen durch Gespräche, Gesten, kurze schriftliche Mitteilungen usw. stattfindenden Kommunikationsprozessen vorkommen. Es sind keine großen und umfassenden semantischen Einheiten aus einer einzigen Quelle, die hier übermittelt werden, und auch Kontinuität, wie man sie für eine idealen Dialog vielleicht sich wünschen möchte, wird hier selten gewahrt. Vor dem Aufkommen digitaler Informationsquellen und des Internets blieb der Bereich der Schriftkultur aus diesem fragmentierten Informationsstrom weitgehend ausgespart. Jetzt hat auch das alltägliche Gespräch mit dem Chat eine schriftliche Form bekommen, und durch das Netz und die dort fragmentiert abrufbare Information entsteht ein Kontinuum von Schriftkultur, Bildkultur und Gespräch. Aus dieser Perspektive kann das Internet auch so betrachtet werden, dass es nicht nur etwas vollkommen Neues ist, sondern auch an sehr alte, tief im Menschen sitzende Eigenschaften und Bedürfnisse anknüpft.
13 Vgl. dazu McCormack, Nancy, Are E-Books Making Us Stupid? In: International Journal of Digital Library Systems, 3 (2), 2012, S. 27–47, S. 39 f.
14 Vgl. McCormack 2012, S. 43 f.
15 Vgl. Griebel, Rolf, Die Bayerische Staatsbibliothek im digitalen Zeitalter. In: Bibliotheksforum Bayern, Sonderheft, 9. Jg. (2015).
16 http://yewno.com/about/
17 Vgl. Lobin, 2014, S. 159.
18 Als Beispiel dafür kann SLUB Semantics betrachtet werden, welches auf der Basis von Wikipedia Konzepten Katalogaufnahmen um diese Konzepte anreichert. Vgl. dazu auch Bonte, A., et al., Brillante Erweiterung des Horizonts: Eine multilinguale semantische Suche für den SLUB-Katalog. In: BIS 4, 4 (2011), S. 210–213.
19 Vgl. dazu Gilson, Tom und Katina Strauch, ATG Interviews Ruggero Gramatica. In: Against the Grain, 2016, S. 64 f.; weitere Informationen zum technischen Hintergrund von Yewno entstammen einem noch nicht veröffentlichten Papier von Ruggero Gramatica und direkter Nachfrage bei Yewno.
20 Vgl. Sahlgren, M., The distributional hypothesis. In: Italian Journal of Linguistics 20.1 (2008), S. 33–54.
21 Vgl. Gramatica, R., et al., Graph Theory Enables Drug Repurposing – How a Mathematical Model Can Drive the Discovery of Hidden Mechanisms of Action. In: PloS one 9.1 (2014): e84912.
22 Vgl. Gilson und Strauch, 2016, S. 66 f.
23 HAN steht für Hidden Automatic Navigator. Dabei handelt es sich um einen Revers Proxy Server der Firma H+H Software GmbH, der den kontrollierten externen Zugriff auf lizenzierte elektronische Medien ermöglicht.
24 Vgl. Gilson und Strauch, 2016, S. 66 f.

Der Verfasser
Dr. Berthold Gillitzer, Stellvertretender Leiter der Abteilung Benutzungsdienste, Bayerische Staatsbibliothek, Ludwigstraße 16, 80539 München,
berthold.gillitzer@bsb-muenchen.de
Foto: privat