Perspektive Bibliothek 5.1 (2016), S. 49-73 10.11588/pb.2016.1.33691

Text und Data Mining: Herausforderungen und Möglichkeiten für Bibliotheken

Bastian Drees

Technische Informationsbibliothek / Bibliotheksakademie Bayern

Bastian.Drees@tib.eu

1. Einleitung

Alleine in den ersten sechs Wochen des Jahres 2016 wurden über 50.000 bestätigte oder vermutete Fälle von Zika-Infektionen in Südamerika dokumentiert.[1] Dies entspricht im Schnitt mehr als tausend neuen Infektionen pro Tag und macht deutlich, dass Zeit bei der Untersuchung des Virus eine entscheidende Rolle spielt. Gleichzeitig gibt es einige hundert Fachartikel, die sich mit dem Zika-Virus beschäftigen[2]. Das Sichten und Zusammentragen des bereits vorhandenen Wissens und besonders dessen Analyse nimmt somit, manuell betrieben, viel Zeit und Personalressourcen in Anspruch. Eine Automatisierung mithilfe von Text und Data Mining-Methoden liefert diese Ergebnisse hingegen innerhalb weniger Minuten.[3] Dieses Beispiel zeigt den enormen Nutzen von Text und Data Mining (TDM), insbesondere bei der Analyse großer Text- oder Datenmengen. Mittlerweile wird TDM für eine große Fülle von Anwendungen gebraucht: TDM wurde in Barack Obamas Wahlkampf verwendet[4] und zur Analyse und Visualisierung der Wikileaks Reporte,[5] TDM kann helfen, chinesische Geisterstädte zu identifizieren[6] und die Tweets wissenschaftlicher

[1] 48.656 vermutete und 1.812 bestätigte Fälle. Vgl. Pan American Health Organization & World Health Organization (2016).
[2] Eine einfache Suche nach "Zika" liefert bei PubMed 345 Treffer (Stand: 1. März 2016).
[3] Murray-Rust (2016).
[4] Issenberg (2012).
[5] Stray (2010).
[6] Chi et al. (2015).

Bibliotheken zu analysieren,[7] Unternehmen verwenden TDM zur Warenkorbanalyse ihrer Kunden[8] usw.; die Liste ließe sich beliebig fortsetzen.

Mit zunehmendem, digital verfügbarem Publikationsaufkommen wächst auch die Bedeutung von TDM als wissenschaftlicher Methode. Wissenschaftlichen Bibliotheken kommt als Informationsinfrastrukturen die Aufgabe zu, Wissenschaftlerinnen und Forschern[9] Informationsquellen so zur Verfügung zu stellen, dass TDM-Anwendungen möglichst problemlos realisierbar sind. Um diese Aufgabe erfüllen zu können, müssen auf Seiten der Bibliothek gewisse Kenntnisse darüber vorhanden sein, was TDM ist (Abschnitt 2.), wie TDM funktioniert (Abschnitt 2.2), wofür TDM verwendet wird (Abschnitt 3.1) und welche Probleme es dabei gibt (Abschnitt 3.2). Auf diese Weise können Bibliotheken die Wissenschaft hinsichtlich TDM optimal unterstützen und sogar selbst TDM zur Verbesserung ihrer Angebote nutzen (Abschnitt 4.). Eine im Rahmen dieser Arbeit durchgeführte Befragung der Betreiber von Repositorien an Universitäten, Hochschulen und Forschungseinrichtungen in Deutschland zu Erfahrungen mit TDM zeigt jedoch, dass dieses Thema bislang nur wenig Beachtung findet (Abschnitt 4.2[10]).

Die Methoden und Anwendungsmöglichkeiten sowie die technischen und rechtlichen Hindernisse von TDM sind vielfältig und breit gefächert oder, wie es die Geschäftsführerin von LIBER, Susan Reilly, in einem Tweet formulierte: „(...) capturing the #TDM landscape is extremely difficult. It's so diverse and there are so many possibilities!"[11] Dennoch soll hier ein möglichst breiter Überblick über Methoden, Anwendungen und Probleme des TDM gegeben und daraus resultierende Herausforderungen und Handlungsfelder für (wissenschaftliche) Bibliotheken aufgezeigt werden. Ein besonderer Fokus liegt dabei auf TDM im wissenschaftlichen und bibliothekarischen Kontext, während das ebenfalls sehr weite Feld des TDM zu kommerziellen Zwecken lediglich am Rande erwähnt wird.

2. Was ist Text und Data Mining?

Die Frage, was Text und Data Mining ist, kann in zweierlei Hinsicht erörtert werden. Zum einen existieren verschiedene Begriffe, die z.T. synonym verwendet werden, z.T. unterschiedliche Aspekte des TDM betonen oder in engerem oder weiterem Sinne als

[7] Al-Daihani & Abrahams (2016).

[8] Besonders häufig wird hier das "Windeln und Bier-Beispiel" verwendet, welches besagt, dass diese beiden Artikel häufig zusammen gekauft werden. Vgl. z.B. D. T. Larose & C. D. Larose (2014, S.247f).

[9] Aus Gründen der besseren Lesbarkeit werden Bezeichnungen von Personengruppen entweder in der weiblichen oder männlichen Form verwendet. Personen anderen Geschlechts sind aber grundsätzlich mit gemeint.

[10] Die vollständigen Ergebnisse der Umfrage finden sich in Drees (2016).

[11] https://twitter.com/skreilly/status/699983231148490752 (besucht am 09. 10. 2016).

Ober- bzw. Unterbegriffe verstanden werden können. Zum anderen kann TDM durch die Prozesse und Methoden, welche unter diesen Begriff fallen, charakterisiert werden.

2.1 Begriffsbestimmung

Die Prozesse, für die sich in der Fachliteratur der Begriff *Text und Data Mining* weitgehend durchgesetzt hat, sind bezüglich ihrer Methoden, Ziele und Anwendungsgebiete zum Teil sehr unterschiedlich. Entsprechend vielfältig sind auch die in der Literatur verwendeten Begriffe. Zum Beispiel finden sich die Begriffe Text Mining, Data Mining, Text Data Mining, Textual Data Mining, Text Knowledge Engineering, Knowledge Discovery in Texts und Knowledge Discovery in Databases[12] sowie Web Mining, Web Content Mining, Web Structure Mining und Web Usage Mining,[13] Content Mining,[14] Literature Mining[15] und sogar Bibliomining („data mining techniques used (...) in libraries are called bibliomining“[16]). Diese werden teilweise synonym verwendet, betonen teilweise aber auch unterschiedliche Aspekte oder implizieren eine engere oder weitere Definition des Begriffs. Auch wenn einige Begriffe heute nur noch selten verwendet werden, z.B. *Knowledge Discovery in Databases* (KDD),[17] ist es sinnvoll, einige der am weitesten verbreiteten Begriffsdefinitionen und -abgrenzungen detaillierter zu betrachten.

Eine häufig zitierte Definition beschreibt KDD als einen „nicht-trivialen Prozess der Identifizierung valider, neuer, potentiell nützlicher und letztlich verständlicher Muster in Daten“ („KDD is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data“[18]). Im Sinne dieser Definition beschreibt KDD den gesamten Prozess, an dessen Anfang eine unprozessierte Datenmenge und an dessen Ende das extrahierte bzw. entdeckte Wissen steht. Dieser Prozess besteht laut Fayyad et al. aus mehreren Schritten, nämlich der Auswahl, Vorprozessierung und Transformation der Daten, gefolgt vom eigentlichen Data Mining-Schritt, in dem Muster identifiziert werden, sowie der abschließenden Interpretation, die erst zum Wissen führt.[19] Data Mining ist demnach ein Teilschritt im Gesamtprozess des KDD, der Datenanalyse und Discoveryalgorithmen verwendet, um Muster in Daten zu identifizieren („Data mining is a step in the KDD process that consists of applying data analysis and discovery algorithms

[12] Mehler & Wolff (2005, S.2).
[13] Mehler & Wolff (2005, S.7-8).
[14] Murray-Rust (o.D.).
[15] Kumar & Tipney (2014).
[16] Siguenza-Guzman et al. (2015, S.500).
[17] So finden sich Knowledge (Platz 6), Kdd (Platz 8) und Discovery (Platz 10) unter den Top 10 Wörtern in KDnuggets News, einem „leading forum for data mining community“, von 1996, während keines dieser Wörter in der entsprechenden Liste für 2005 auftaucht. Siehe Table 1 in Piatetsky-Shapiro (2007, S.102).
[18] Fayyad et al. (1996, S.40-41).
[19] Vgl. figure 1 in Fayyad et al. (1996, S.41).

that (...) produce a particular enumeration of patterns (or models) over the data“[20]). Eine weitere Definition versteht unter Data Mining auf Software basierende Mechanismen und Techniken, die versteckte Informationen aus Daten extrahieren („Broadly, data mining can be defined as a set of mechanisms and techniques, realized in software, to extract hidden information from data“[21]). Hier ist die Einschränkung auf (semi-)automatische, durch Software realisierte Prozesse, welche in der vorherigen Definition nur implizit enthalten war, und die Betonung der *versteckten* Information wichtig. Der letzte Aspekt macht deutlich, dass z.B. eine einfache Datenbankabfrage noch kein Data Mining ist.

Diese Extraktion versteckter oder bislang unbekannter Informationen, wird von einigen Autoren als ein wesentlicher Unterschied zwischen Data Mining und Text Mining betrachtet:

> „(...) it is helpful to distinguish between text mining as the extraction of semantic logic from text, and data mining which is the discovery of new insights. The knowledge that is extracted during text mining is not new and it is not hidden. That information was already known to the author of the text, otherwise they could not have written it down.“[22]

Text Mining im Sinne dieses Ansatzes extrahiert somit die Bedeutung von Texten[23] und findet insbesondere in Bereichen des Information Retrieval, der Informationsextraktion oder Textzusammenfassung Anwendung. Diese Auffassung wird von Mehler und Wolff als methodenorientierter Ansatz bezeichnet.[24] Demgegenüber stehen wissensorientierte Ansätze, die z.B. den „automatischen Aufbau von so genannten Ontologien und ihre Nutzbarmachung im Zusammenhang des Semantic Web“[25] zum Ziel haben. Hier werden also nicht nur die Bedeutungen der Texte extrahiert sondern auch neue Informationen geschaffen.

Unabhängig von diesen beiden Ansätzen liegt jedoch der zentrale Unterschied zwischen Data Mining und Text Mining in den analysierten Datentypen, nämlich (mehr oder weniger) strukturierten Daten einerseits und (natürlichsprachigen) unstrukturierten Daten andererseits. Darüber hinaus werden auch andere Inhalte (bspw. Bilder, Ton, Filme, etc.) mit ähnlichen Methoden analysiert. Um all diese verschiedenen Verfahren zusammenfassend zu beschreiben – dies ist insbesondere in der Diskussion um Urheberrechtsfragen von Bedeutung – führte der britische Chemiker Peter Murray-Rust den Begriff des *Content*

[20] Fayyad et al. (1996, S.41).
[21] Coenen (2011, S.25).
[22] Clark (2013, S.6).
[23] „This is what text mining aims to do: to extract the meaning of a passage of text and to store it as a database of facts about the content and not simply a list of words.“ Clark (2013, S.5).
[24] Mehler & Wolff (2005, S.5-6).
[25] Mehler & Wolff (2005, S.6).

Mining ein.[26] In diesem Sinne kann Content Mining als Prozess beschrieben werden, bei dem mithilfe von Software Informationen aus maschinenlesbarem Material gewonnen werden.[27] Hier soll der Begriff *Text und Data Mining (TDM)* verwendet werden, da dieser in der aktuellen Literatur der am weitesten verbreitete ist. TDM soll dabei aber in dem weitgefassten Sinne des gerade beschriebenen Content Mining und damit deutlich über das reine Text Mining und Data Mining, wie in den vorherigen Absätzen beschrieben, hinausgehend verstanden werden. TDM bezeichnet hier also die (semi-)automatische Analyse digitaler Inhalte jeglicher Form.

Neben der weiten Verbreitung des Begriffs ist es auch deshalb sinnvoll, von Text und Data Mining zu sprechen, da dieser Begriff, durch die eigenständige Nennung von Text *und* Daten, auf die Unterschiede zwischen Text Mining und Data Mining, die trotz aller Gemeinsamkeiten vorhanden sind, hinweist. Die strukturiertere Datengrundlage beim Data Mining (z.B. tabellarische Daten) im Gegensatz zu weitgehend unstrukturierten Daten beim Text Mining (natürlichsprachiger Text; Strukturelemente sind hier z.B. Überschriften) hat Auswirkungen auf die Anwendungsgebiete und die angewendeten Techniken. Wie erwähnt werden beim Data Mining neue Erkenntnisse gewonnen, während dies aufgrund der unstrukturierten Datengrundlage beim Text Mining häufig nicht ohne weiteres möglich ist. Text Mining beschränkt sich daher häufig darauf, aus dem unstrukturierten Text strukturierte Daten zu gewinnen.[28] Daher ist Text Mining eine Anwendung, die häufig im Zusammenhang mit Suche und Retrieval eingesetzt wird. Auch werden hier, gemäß der zugrunde liegenden (Text-)Daten, Methoden und Techniken des *Natural Language Processings* und der *Computerlinguistik* verwendet, die beim klassischen Data Mining keine Anwendung finden. Im Folgenden sollen die wichtigsten Aufgabengebiete und Techniken des TDM vorgestellt werden.

2.2 Ziele, Aufgaben und Methoden des Text und Data Minings

Ziele, Aufgaben und Methoden des TDM unterscheiden sich nach Anwendungsgebiet und individueller Fragestellung, dennoch lassen sie sich nach verschiedenen Kriterien kategorisieren. Allerdings ist auch diese Kategorisierung einer gewissen Willkür unterworfen und hängt von subjektiven Standpunkten ab, da eine klare Trennung nicht immer möglich ist und Unterschiede und Gemeinsamkeiten in verschiedenen Fachgebieten unterschiedlich beurteilt werden. Ferner ist die Unterscheidung zwischen Ziel und Aufgabe

[26] „When speaking about TDM, Mr Murray-Rust prefers to speak about 'content' mining" OpenForum Academy (2015, S.8).

[27] In der Hague Declaration wird Content Mining beschrieben als „the process of deriving information from machine-readable material" und als „computer analysis of content in all formats" LIBER (2015, S.1).

[28] „In simple terms, text mining is the process that turns text into data that can be analysed" Clark (2013, S.5).

bzw. Aufgabe und Methode nicht immer eindeutig. So betrachten Fayyad et al. *Vorhersage* ("prediction") und *Beschreibung* ("description") als Ziele ("goals")[29] während D. und C. Larose darunter Aufgaben ("tasks") des TDM verstehen.[30] Gleichwohl sollen hier die wichtigsten und häufigsten Ziele, Aufgaben und Methoden systematisch dargestellt werden. Die Ziele sollen dabei die Frage beantworten, *warum* TDM verwendet wird, die Aufgaben beschreiben, *was* gemacht wird und die Methoden, *wie* es gemacht wird.

Ziele des TDM

Das zentrale Ziel von TDM ist immer das Finden von Informationen im weitesten Sinne. Dieses kann, wie bereits erwähnt, unterschieden werden zwischen einem *Finden von Information* als Auffinden oder Wiederfinden bekannter Informationen einerseits oder als Entdeckung neuer, unbekannter Informationen andererseits ("Finding what is already known" und "Finding what was not obvious").[31] Wo TDM als (Teilschritt von) Knowledge Discovery verstanden wird, steht im Allgemeinen letzteres im Vordergrund. Das Wiederauffinden bekannter Informationen wird dagegen häufig als Ziel des Text Minings genannt und wird zu Zwecken der Anreicherung der Inhalte, der Suche und dem Information Retrieval verwendet.[32] Typische Ziele sind daher „Search and Retrieval",[33] „Enriching Content" oder „Systematic Literature Review".[34] Darüber hinaus nennt Clark noch die Forschung im Bereich der „Computational Linguistics", die TDM selbst als Forschungsobjekt hat.[35] Ein weiteres Ziel kann laut Saffer und Burnett die Analyse von Forschungsnetzwerken, z.B. Zitations- oder Kollaborationsnetzwerken, sein.[36] Solche Analysen können u.a. helfen, potentielle Kollaborationspartner zu ermitteln. Diese Ziele können, je nach konkretem Anwendungsfall, mithilfe verschiedener TDM *tasks* oder Aufgaben erreicht werden.

Aufgaben des TDM

Auch bei den Aufgaben des TDM können je nach Fachgebiet, Kontext oder Problemstellung speziellere oder allgemeinere Einteilungen sinnvoll sein. Eine allgemeingültige Einteilung ist daher nicht möglich, auch weil die Abgrenzungen oft unscharf sind. Eine der allgemeinsten Einteilungen liefert Coenen, der zwischen *Musterextraktion* („pattern

[29] Vgl. Fayyad et al. (1996, S.43).
[30] Vgl. D. T. Larose & C. D. Larose (2014, S.8-10).
[31] Saffer & Burnett (2014, S.3).
[32] So bezeichnet z.B. Clark Text Mining als "smart indexing" („text mining is smart indexing" Clark (2013, S.5).
[33] Saffer & Burnett (2014, S.2).
[34] Clark (2013, S.7).
[35] Clark (2013, S.7).
[36] Saffer & Burnett (2014, S.3-4).

extraction“), *Segmentierung* („Clustering“) und *Klassifikation* („Classification“) unterscheidet.[37] Fasst man diese drei Aufgaben im weitesten Sinne auf, so lassen sich nahezu alle TDM-Probleme einer der drei Aufgaben zuteilen. Segmentierung und Klassifikation ähneln sich dahingehend, dass hierbei Daten gruppiert, d.h. in Gruppen (Cluster oder Klassen) eingeteilt werden. Der Unterschied liegt in der Vorgehensweise: Klassifikation ist ein prognostisches Verfahren mit im Voraus bestimmten Klassen während Segmentierung ein beschreibendes Verfahren ohne vordefinierte Klassen ist. Dies wird auch als *supervised learning* (Klassifikation) und *unsupervised learning* (Segmentierung) bezeichnet.[38]

Neben *Extraktion/Identifikation*, *Clustering/Segmentierung* und *Klassifikation/Kategorisierung* sind weitere, häufig eigenständig genannte Aufgaben: Regressionsanalyse, Zusammenfassung, Abhängigkeitsanalyse und Abweichungsanalyse,[39] Beschreibung, Schätzung, Prognose, Assoziation[40] u.v.m. Diese weitere Untergliederung der Aufgaben ist abhängig vom jeweiligen Anwendungsgebiet häufig sinnvoll oder notwendig, soll hier aber nicht weiter vertieft werden. Für die Lösung dieser Aufgaben steht eine Vielzahl gut erprobter Methoden und Techniken vor allem aus den Bereichen Statistik, Maschinelles Lernen und Computerlinguistik zur Verfügung. Im Folgenden sollen einige der wichtigsten Methoden überblicksartig dargestellt werden.

Methoden des TDM

Musterextraktion: Eine weitverbreitete Methode der Musterextraktion ist die Assoziationsanalyse,[41] deren bekannteste Anwendung die Warenkorbanalyse ist. Bei der Assoziationsanalyse werden Attribute identifiziert, die häufig gemeinsam auftreten; d.h. sie liefert Assoziationsregeln der Form “Wenn Attribut A vorhanden ist, ist häufig auch Attribut B vorhanden”. Solche Analysen werden u.a. für Empfehlungssysteme genutzt, die Assoziationsregeln wie “Kunden, die Produkt X kauften, kauften auch Produkt Y” generieren.[42]

Clustering Algorithmen: Beim Clustering werden Daten in der Weise gruppiert, dass die Daten innerhalb eines Clusters möglichst ähnlich und zwischen verschiedenen Clustern

[37] Vgl. Coenen (2011, S.26-27).
[38] Vgl. Coenen (2011, S.27).
[39] Fayyad et al. sprechen von “Regression”, “Summarization”, “Dependency modeling”, “Change and deviation detection”. Fayyad et al. (1996, S.44-45), und auf der guten Einführungsseite wissensexploration.de heißt es: „Die Aufgaben des Data Mining sind Klassifikation, Segmentierung, Prognose, Abhängigkeitsanalyse und Abweichungsanalyse.“ Gotter (o.D.).
[40] D. und C. Larose sprechen von “Description”, “Estimation”, “Prediction”, “Association”. D. T. Larose & C. D. Larose (2014, S.8-14).
[41] Agrawal et al. (1993).
[42] Vgl. Petersohn (2005, S.101ff).

möglichst unterschiedlich sind. Die "Ähnlichkeit" wird dabei mit einer geeignet definierten "Distanz" der Daten gemessen.[43] Weit verbreitet sind das hierarchische und das *k*-means Clustering. Beim *hierarchischen Clustering* wird zunächst jedes Datum als einzelnes Cluster betrachtet. Dann werden durch Verschmelzung von Clustern mit geringer Distanz sukzessive größere Cluster gebildet. Man spricht hier vom agglomerativen hierarchischen Clustering im Gegensatz zum divisiven hierarchischen Clustering, das mit einem großen, alle Daten enthaltenden Cluster startet und dieses sukzessive teilt.[44] Beim k-*means Clustering* wird die Anzahl der Cluster (*k*) vorgegeben und diese dann so verteilt, dass die intra-Cluster-Varianz minimiert und die inter-Cluster-Varianz maximiert wird.[45] Letzteres Verfahren eignet sich insbesondere dann, wenn die Zahl der zu erwartenden Cluster im Voraus bekannt ist. Andernfalls kann dieses Verfahren zu Problemen und sinnlosen Ergebnissen führen.

Klassifikationsalgorithmen: Beim Klassifizieren werden Daten verschiedenen, im Voraus definierten Kategorien zugeordnet. Zum Beispiel könnten in der Biomedizinischen Forschung Faktoren danach klassifiziert werden, ob sie einen positiven, negativen oder gar keinen Einfluss auf eine bestimmte Krankheit haben. Häufig verwendete, einfache Techniken sind z.B. Entscheidungsbäume oder Nächste-Nachbarn-Klassifikation. Bei der *Nächste-Nachbarn-Klassifikation* wird ein Set von Trainingsdaten verwendet und das zu klassifizierende Datum mit den "nächsten Nachbarn" verglichen. Es wird dann der Klasse zugeordnet, mit der die größte Übereinstimmung, d.h. Ähnlichkeit besteht.[46] *Entscheidungsbäume* bestehen aus Knoten ("nodes") die durch Äste ("branches") verbunden sind. Ausgehend vom Wurzelknoten ("root node") werden Attribute an den Knoten geprüft und auf dieser Grundlage eine Entscheidung für einen Ast getroffen. Auf diese Weise wird der Baum bis zu einem der (zwei oder mehr) Blattknoten ("leaf nodes") durchlaufen, welche das Klassifikationsergebnis bilden.[47] Weitere Methoden sind u.a. die Regressionsanalyse und neuronale Netze.[48]

Computerlinguistische Ansätze: Um Information aus natürlichsprachigen Texten zu extrahieren, ist es notwendig, neben statistischen Methoden auch Methoden des *natural language processing* zu verwenden. Zur automatischen semantischen Analyse von Texten ist der Einsatz von Ontologien ein wichtiges Hilfsmittel. *Ontologien* sind kontrollierte,

[43] Vgl. D. T. Larose & C. D. Larose (2014, S.209ff).
[44] Vgl. Piegorsch (2015, S.376ff), und D. T. Larose & C. D. Larose (2014, S.212ff).
[45] Vgl. Piegorsch (2015, S.384ff), und D. T. Larose & C. D. Larose (2014, S.215ff).
[46] Vgl. Hotho et al. (2005, S.33f), Piegorsch (2015, S.308ff), und D. T. Larose & C. D. Larose (2014, S150ff).
[47] Vgl. Hotho et al. (2005, S.34), und D. T. Larose & C. D. Larose (2014, S165ff).
[48] Vgl. z.B. D. T. Larose & C. D. Larose (2014, S.118ff), und D. T. Larose & C. D. Larose (2014, S.188ff).

strukturierte Vokabulare, die Entitäten einer Domäne und ihre Beziehungen zueinander enthalten. Sie sind nötig, um Texte für den Computer interpretierbar zu machen.[49] Ontologien oder kontrollierte Vokabulare werden zum Beispiel zur *Named Entity Recognition* verwendet, bei der Eigennamen von Entitäten in Texten erkannt werden.[50] Dies kann z.B. zur automatischen Verschlagwortung von Texten genutzt werden. Andererseits können aber auch Text Mining-Methoden angewendet werden, um semi- oder vollautomatisch Ontologien zu erstellen.[51]

3. TDM in Wissenschaft und Forschung: Anwendungen und Hindernisse

In einer aktuellen Umfrage[52] unter 177 Wissenschaftlern aus den Bereichen der Geistes-, Sozial-, Lebens-, Natur- und Ingenieurswissenschaften zu „Bedarf und Anforderungen an Ressourcen für Text und Data Mining" gibt eine Mehrheit der Befragten an, TDM sei zwar nur "eine Forschungsmethode unter anderen",[53] habe aber einen großen oder sogar ausschlaggebenden Nutzen für ihre Forschung.[54] Außerdem erwartet die Mehrheit der Befragten, "dass in Zukunft der Anteil der TDM an der eigenen Tätigkeit größer werden wird".[55] Dies zeigt die große Bedeutung von TDM über alle Fachgrenzen hinweg und den Bedarf der Wissenschaft, Rahmenbedingungen vorzufinden, die TDM ermöglichen. Hierfür sind Rechtssicherheit und leichter Zugang zu maschinenlesbaren Inhalten die entscheidenden Aspekte. Im Folgenden wird der Nutzen sowie die Bandbreite von TDM exemplarisch anhand einiger Anwendungsbeispiele in verschiedenen Disziplinen dargelegt (Abschnitt 3.1). Ferner wird aufgezeigt, welche Faktoren und Rahmenbedingungen mögliche TDM-Vorhaben behindern (Abschnitt 3.2).

3.1 TDM in Wissenschaft und Forschung: Anwendungen

TDM-Techniken finden in nahezu jeder wissenschaftlichen Disziplin, von Geowissenschaften über Materialwissenschaften bis hin zu Neurowissenschaft, Psychologie oder Sozialwissenschaften, Anwendung.[56] In der Medizin werden TDM-Techniken angewen-

[49] Vgl. Blöhdorn & Hotho (2009, S.640ff), und Carstensen et al. (2010, S.532ff).
[50] Carstensen et al. (2010, S.596ff).
[51] Vgl. Haarmann (2014).
[52] Sens et al. (2015a), und Sens et al. (2015b).
[53] Sens et al. (2015a, S.6).
[54] Sens et al. (2015a, S.7).
[55] Sens et al. (2015a, S.8).
[56] So diskutieren Liao et al. Publikationen zu Data Mining-Anwendungen, welche zwischen 2000 und 2011 veröffentlicht wurden und aus einer Vielzahl von Disziplinen stammen: „In this paper, the articles discussed were sourced from different discipline areas, including computer science, engineering, medici-

det, um z.B. Risikogruppen für Krankheiten zu identifizieren,[57] in den Digital Humanities z.B. für "inhaltsanalytische Studien auf großen Textsammlungen",[58] in der Geographie z.B. zur geographischen Segmentierung, um Landnutzung oder Orte hoher Kriminalität zu analysieren[59] und in der Astronomie zur Segmentierung und Klassifizierung astronomischer Daten.[60]

Eine typische Anwendung von TDM-Methoden geht auf Swansons ABC-Modell[61] zurück, mit dem komplementäres aber getrennt vorliegendes Wissen[62] identifiziert werden soll. Die Idee ist dabei, dass zwei Aussagen der Form "wenn A, dann B" und "wenn B, dann C", aus der wissenschaftlichen Literatur extrahiert werden und zu der Aussage (bzw. zu beweisenden Hypothese) "wenn A, dann C" kombiniert werden. Auf diese Weise gelang es z.B. Weeber et al.[63] mehrere Krankheiten als mögliche Kandidaten zu identifizieren, die mit Thalidomid behandelt werden könnten. Hierzu verwendeten sie ein konzeptbasiertes Suchsystem, welches Konzepte in wissenschaftlichen Artikeln auf der Grundlage des *Unified Medical Language System (UMLS)* Thesaurus erkennt. Die konzeptbasierte Suche ermöglicht dabei insbesondere die Einschränkung auf semantische Kategorien, wie z.B. "Disease or Syndrome", "Gene or Genome" oder "Amino Acid, Peptide, or Protein".[64] Beginnend mit Thalidomid als Konzept A, suchten sie nach Konzepten der Kategorie "Immunologic Factor", die durch Thalidomid beeinflusst wurden (Konzept B im ABC-Modell). In einem zweiten Schritt identifizierten sie zu den Konzepten B – analog zum ersten Schritt – Konzepte C der Kategorie "Disease or Syndrome". So gelangten sie zu Hypothesen, welche Krankheiten mit Thalidomid behandelt werden könnten. Der Vorgang wurde semi-automatisch durchgeführt, indem Expertinnen die automatisch gefundenen Konzepte (B und C) weiter einschränkten und geeignete Kandidaten auswählten.

Eine weitere typische TDM-Anwendung ist die Vorhersage von Proteinfunktionen. Da durch die zunehmende Automatisierung der DNA-Sequenzierung die entspre-

---

ne, mathematics, earth and planetary sciences, biochemistry, genetics and molecular biology, business, management and accounting, social sciences, decision sciences, multidisciplinary, environmental science, energy, agricultural and biological sciences, nursing, materials science, pharmacology, toxicology and pharmaceutics, chemistry, health professions, physics and astronomy, economics, econometrics and finance, psychology, neuroscience, chemical engineering and veterinary" Liao et al. (2012, S.11307).

[57] Eine Anwendung, die neben dem offensichtlichen Nutzen für Prophylaxe, Prävention und Früherkennung auch von Krankenversicherungen zur Risikoanalyse genutzt wird. Vgl. Koh & Tan (2005, S.68).

[58] Abschnitt 2 in Blessing et al. (2015).

[59] Vgl. Han et al. (2009, S.152ff).

[60] Borne (2009).

[61] Swanson (1986, 1988).

[62] Eine Veröffentlichung Swansons trägt den Titel „Complementary Structures in Disjoint Science Literatures" Swanson (1991).

[63] Weeber et al. (2003).

[64] Weeber et al. (2003, S.253).

chenden Datenmengen exponentiell anwachsen, ist eine gleichermaßen automatisierte Datenanalyse oft wünschenswert, wenn nicht gar notwendig. Daher werden TDM-Methoden vielfach verwendet, um die Funktionen von Proteinen vorherzusagen.[65] Frühe Arbeiten zu diesem Thema stammen z.B. von King et al.,[66] die eine Kombination aus TDM und maschinellem Lernen verwenden. Anhand der *open reading frames*[67], die für Proteine bekannter Funktionen kodieren, werden zunächst Regeln abgeleitet, die Proteinfunktionalitäten ausschließlich auf die DNA-Sequenz zurückführen. Mithilfe dieser Regeln werden dann solche open reading frames, die für Proteine unbekannter Funktion kodieren, klassifiziert. Auf diese Weise konnten zwischen 24% und 65% der open reading frames dreier Organismen[68] mit einer Genauigkeit von 60%-80% bestimmt werden.

Diese beispielhaften Anwendungsszenarien illustrieren den großen potentiellen Nutzen von TDM als wissenschaftlicher Methode. Obwohl also bereits viele erfolgreiche und vielversprechende TDM-Anwendungen existieren, sind die aktuellen Rahmenbedingungen nicht immer ideal für TDM. Im Folgenden sollen die wesentlichsten Hindernisse, insbesondere rechtlicher und technischer Natur, beschrieben werden.

3.2 TDM in Wissenschaft und Forschung: Hindernisse

In der oben zitierten Umfrage[69] wurden die Wissenschaftler auch nach Hindernissen für TDM befragt. Die größten Probleme sind danach: 1. problematische rechtliche Rahmenbedingungen,[70] 2. die Verteilung der relevanten Texte und Daten auf zu viele Anbieter[71] und 3. der fehlende Zugang zu relevanten Texten und Daten.[72] Ferner wurde in mehreren Freitextantworten das Fehlen computerlesbarer Dateiformate angegeben.[73] Im Folgenden wird ein kurzer Überblick über die wichtigsten Probleme und Hindernisse gegeben.

[65] Ein Review zu diesem Thema findet sich bei Sleator (2012).
[66] Ross D. King et al. (2000), Clare & R. D. King (2003).
[67] Open reading frames – oder zu deutsch offene Leserahmen – sind die Bereiche der DNA, die für die Proteine kodieren, d.h. die Information über den "Bauplan" der Proteine enthalten.
[68] Bereits im Jahr 2000 konnten so die Proteinfunktionen für *Mycobacterium tuberculosis* und *Escherichia coli* bestimmt werden und 2003 für die Bäckerhefe *Saccharomyces cerevisiae*. Ross D. King et al. (2000), Clare & R. D. King (2003).
[69] Sens et al. (2015a).
[70] 52% der Befragten beantworten diese Frage mit stark (4) bis sehr stark (5) zutreffend. Sens et al. (2015a, S.28).
[71] 38% der Befragten beantworten diese Frage mit stark (4) bis sehr stark (5) zutreffend. Sens et al. (2015a, S.24).
[72] 28% der Befragten beantworten diese Frage mit stark (4) bis sehr stark (5) zutreffend. Sens et al. (2015a, S.25).
[73] Insbesondere wurde auf Probleme mit dem weitverbreiteten PDF-Format hingewiesen: „PDF Format von Volltext-Artikeln hindert Textextraktion", Sens et al. (2015a, S.29).

Rechtliche Hindernisse

Ein wesentliches Hindernis von TDM-Vorhaben und ein viel diskutiertes Thema ist die oft undursichtige und restriktive Urheberrechtslage bezüglich TDM. Eine Wissenschaftlerin, die legalen Zugang zu Texten und Daten hat, darf diese lesen und die enthaltenen Informationen und Fakten analysieren, um zu neuen Erkenntnissen zu gelangen. Dies ist die Grundlage jeder wissenschaftlichen Forschung. Weder das Lesen noch die Fakten selbst sind urheberrechtlich geschützt.[74] Die gleichen Vorgänge finden beim TDM statt, lediglich automatisiert. Da für das automatisierte Lesen im Regelfall Kopien der Inhalte erstellt werden, berufen sich die Rechteinhaber auf das vom Urheberrecht geschützte Vervielfältigungsrecht. Während in den USA die sogenannte *Fair Use*-Regelung TDM zu nicht-kommerziellen Zwecken erlaubt[75] ist die Situation in Europa unklarer. Das aktuelle europäische Urheberrecht basiert auf der *Richtlinie 2001/29/EG zur Harmonisierung bestimmter Aspekte des Urheberrechts und der verwandten Schutzrechte in der Informationsgesellschaft*,[76] die in Artikel 2 dem Urheber das ausschließliche Recht zuspricht, „die unmittelbare oder mittelbare, vorübergehende oder dauerhafte Vervielfältigung auf jede Art und Weise und in jeder Form ganz oder teilweise zu erlauben oder zu verbieten".[77] Eine Wissenschaftlerin in Europa, die TDM an urheberrechtlich geschützten Arbeiten durchführen möchte, benötigt also entweder eine entsprechende Lizenz des Rechteinhabers oder eine grundsätzliche Ausnahmeregelung vom Urheberrecht für TDM. Letzteres, eine sogenannte Urheberrechtsschranke, existiert z.B. seit dem 1. Juni 2014 in Großbritannien[78] und wird aktuell in Frankreich[79] sowie auf EU-Ebene[80] diskutiert.

Die Richtlinie 2001/29/EG erlaubt in Artikel 5 zwar „Vervielfältigungshandlungen, die flüchtig oder begleitend sind und einen integralen und wesentlichen Teil eines technischen Verfahrens darstellen und deren alleiniger Zweck es ist, (...) eine rechtmäßige Nutzung eines Werks oder sonstigen Schutzgegenstands zu ermöglichen, und die keine eigenständige wirtschaftliche Bedeutung haben",[81] also TDM zu nicht-kommerziellen Zwecken und insofern keine permanenten Kopien erstellt werden, allerdings kann diese Ausnahme von anderslautenden Lizenzbestimmungen überschrieben werden. Auch Lizenzen, die limitieren, wieviel Inhalt auf einmal heruntergeladen werden darf, können TDM-Vorhaben behindern.[82] So berichtet z.B. Chris Hartgerink davon, wie er sowohl von

[74] Vgl. Helmholtz Open Access Koordinationsbüro (2013, S.1).
[75] Vgl. z.B. Association of Research Libraries (2015).
[76] Europäisches Parlament (2015).
[77] Europäisches Parlament (2015, Artikel 2: Vervielfältigungsrecht).
[78] Intellectual Property Office (2014).
[79] Vgl. Langlais (2016).
[80] Vgl. z.B. LIBER (2015).
[81] Europäisches Parlament (2015, Artikel 5: Ausnahmen und Beschränkungen).
[82] Vgl. Science Europe (2015, S.5).

Elsevier als auch von Wiley daran gehindert wurde, Artikel zu TDM-Zwecken herunterzuladen. Dies geschah, obwohl er legalen Zugang zu den Artikeln hatte und seine Downloadrate mit weniger als neun bzw. fünf Artikeln pro Minute weit davon entfernt war, die Serverkapazitäten zu überlasten.[83] Wie die Beispiele Großbritannien und Frankreich zeigen, gibt es darüber hinaus eine Vielzahl nationaler Sonderregelungen, die Kooperationen von Forschern aus verschiedenen EU-Staaten ebenso erschweren, wie eine Beschränkung auf nicht-kommerzielles TDM das Zustandekommen von Public-Private-Partnerships verkompliziert.

Die zentralen urheberrechtlichen Probleme sowie Forderungen, wie diese im Sinne der Wissenschaft zu lösen wären, formuliert die Hague Declaration.[84] Deren Kern ist die Forderung nach einer europaweiten Ausnahmeregelung vom Urheberrecht für TDM. Diese soll TDM für alle Dokumente, auf die legaler Lesezugriff besteht, sowohl für kommerzielle als auch für nicht-kommerzielle Zwecke erlauben und nicht durch anderslautende Verträge überschrieben werden können.[85] Die grundsätzliche Argumentation folgt dabei den Überlegungen, dass „intellektuelles Eigentum nicht dazu gemacht wurde, den freien Fluss von Wissen zu behindern".[86] Außerdem sollte, wenn eine manuelle Datenanalyse erlaubt ist (d.h. legaler Lesezugriff besteht), auch eine computergestützte Analyse möglich sein („The right to read is the right to mine!"[87]). Auch eine aktuelle Studie,[88] laut der ein restriktives Urheberrecht die Durchführung von TDM hemmt, zeigt die Probleme der aktuellen Urheberrechtsregelungen bezüglich TDM. Zwar werden von Seiten der Verlage mittlerweile einige TDM-Plattformen und Lösungen angeboten,[89] die eine einfache Rechteklärung ermöglichen und in vielen Fällen eine gute Option für Wissenschaftler darstellen können, dennoch müssen diese Angebote kritisch betrachtet werden. Da TDM-Vorhaben sehr unterschiedlich ausfallen können, sind standardisierte Lizenzen, wenn sie nicht sehr liberal und weitgefasst sind, nicht immer hilfreich. Außerdem sind Wissenschaftler bei der Nutzung solcher Plattformen vom jeweiligen Anbieter und seinen TDM-Lösungen abhängig, wodurch Innovationen (bspw. eigene TDM-Lösungen) verhindert werden und die Unabhängigkeit der Wissenschaft sowie die Reproduzierbarkeit der Ergebnisse gefährdet sein können.[90]

[83] Hartgerink (2015, 2016).
[84] LIBER (2015).
[85] LIBER (2015).
[86] Prinzip 1 der Hague Declaration: „Intellectual property was not designed to regulate the free flow of facts, data and ideas, (...)" LIBER (2015).
[87] Murray-Rust (o.D.).
[88] Handke et al. (2015).
[89] Zwei der bekanntesten Angebote stammen von CrossRef (2015), und vom Copyright Clearing Center (2015).
[90] Vgl. Science Europe (2015, S.6f).

Technisch-strukturelle Hindernisse

Auch wenn alle rechtlichen Fragen geklärt sind und TDM legal durchführbar ist, können technische oder strukturelle Hindernisse TDM behindern, verlangsamen oder ganz unmöglich machen. Die relevanten Inhalte müssen erstens maschinenlesbar sein und zweitens muss ein automatisierter Zugang zu großen Mengen von Dokumenten möglich sein. Wie schon bei den Lizenzen und nationalen Urheberrechtsregelungen besteht auch hier ein großes Problem in fehlender Einheitlichkeit. In den meisten Fällen liegen die relevanten Inhalte in unterschiedlichen (strukturierten und unstrukturierten) Formaten, verteilt auf viele verschiedene Plattformen, welche unterschiedliche Schnittstellen zum Download anbieten, vor. Darüber hinaus werden insbesondere Textdokumente häufig ausschließlich als PDF angeboten. Zum Teil wird zwar eine zusätzliche HTML-Version aber nur selten auch eine XML-Version angeboten. Dies ist insofern problematisch, da strukturierte Formate (z.B. XML, HTML), die automatische Informationsextraktion gegenüber unstrukturierten Formaten (z.B. PDF) deutlich vereinfachen.[91] So empfiehlt bspw. Clark den Verlegern wissenschaftlicher Texte: „It would be more useful for the researcher the HTML is made available alongside the PDF. More useful still is content that can be delivered in a more structured format such as XML."[92]

Zusammenfassend können zwei Arten von Hindernissen unterschieden werden:[93] *Hindernisse durch Einschränkung* können rechtliche Verbote oder nicht-maschinenlesbare Formate sein. *Hindernisse durch Fragmentierung* sind z.B. Uneinheitlichkeiten bei Urheberrechtsregelungen verschiedener Länder oder Lizenzbestimmungen verschiedener Anbieter, auf vielen Plattformen verteilte Inhalte sowie Inhalte in unterschiedlichen Formaten.

4. TDM und Bibliotheken: Möglichkeiten und Herausforderungen

Für wissenschaftliche Bibliotheken ist TDM unter drei Aspekten relevant. Erstens nimmt die Bibliothek gleichsam eine Vermittlerrolle zwischen den Anbietern von Inhalten und den Wissenschaftlern ein und sollte hier dafür sorgen, dass TDM auf den angebotenen Inhalten technisch und rechtlich möglich ist. Zweitens sind Bibliotheken häufig selbst Anbieter von Inhalten und sollten daher ihrerseits darauf achten, dass die in Repositorien und auf Publikationsservern angebotenen Dokumente für TDM-Anwendungen zugänglich sind. Drittens sind Bibliotheken auch immer häufiger Anwender von TDM, z.B. zur

[91] „The more structure there is, the more meaning can be extracted by the machine." Clark (2013, S.10).
[92] Clark (2013, S.14).
[93] Molloy teilt die rechtlichen Beschränkungen in drei Kategorien „restriction by restrictiveness", „restriction by fragmentation" und „restriction by uncertainty"; vgl. Molloy (2016).

Anreicherung der angebotenen Inhalte, zum Information Retrieval oder beim Einsatz von Empfehlungssystemen.

4.1 Bibliotheken als Vermittler

Bibliotheken sind Dienstleister, die ihren Nutzerinnen den Zugang zu Informationen ermöglichen. Eine zentrale Aufgabe, die den Bibliotheken in dieser Vermittlerposition zwischen Informationsquelle und Nutzerin zukommt, ist es, die (Nach-)Nutzbarkeit dieser Informationen sicherzustellen. Bezogen auf TDM und vor dem Hintergrund der im letzten Abschnitt aufgezeigten Hindernisse bedeutet dies, dass es die Aufgabe der Bibliothek ist, rechtliche Fragen zu klären, unnötige technische Hindernisse zu beseitigen und Wissenschaftler bei der Nutzung zu unterstützen.[94] Solange keine Urheberrechtsschranke für TDM existiert, ist es daher die Aufgabe der Bibliotheken, bei Lizenzverhandlungen entsprechende Klauseln zu vereinbaren, die das TDM des lizenzierten Materials ermöglichen.

Zu diesem Zweck ermittelte eine Ad-hoc-Arbeitsgruppe "Text und Data Mining"[95] der Allianzinitiative im Rahmen einer Umfrage[96] unter Wissenschaftlern verschiedenster Fachgebiete die tatsächlichen Bedarfe der Wissenschaft bezüglich TDM. Auf Grundlage der Umfrageergebnisse wurde eine TDM-Klausel erarbeitet, die zum Musterlizenzvertrag der Allianzinitiative hinzugefügt wurde.[97] Diese Klausel erlaubt das TDM des lizenzierten Materials, um Angebote (z.B. Information Retrieval) zu verbessern, sowie für Lehre und Forschung („The Licensed Material may be used for text and data mining to enhance services, to encourage scholarship, teaching and learning and to conduct research (...)"[98]). Ferner dürfen Rohdaten extrahiert und analysiert werden und diese, um Reproduzierbarkeit zu gewährleisten, auch gespeichert, publiziert und verbreitet werden, sofern das lizenzierte Material nicht daraus reproduziert werden kann („Raw data may be extracted (...) stored, published and distributed in any medium or form under any license in order to ensure reproducibility and sustainability, as long as the Licensed Material cannot be reconstructed in its original, human readable form."[99]). Außerdem verpflichtet sich der Lizenzgeber dem Nutzer das lizenzierte Material in einer möglichst nützlichen Art und Weise zur Verfügung zu stellen („The Licensor will cooperate with Licensee and Authorised Users as reasonably necessary in making the Licensed Material available in a

[94] Vgl. Mumenthaler (2015).
[95] http://www.allianzinitiative.de/handlungsfelder/querschnittsthemen/nutzungsrechte/ad-hoc-arbeitsgruppe.html
[96] Sens et al. (2015a,b).
[97] Allianzinitiative (2016, §3 Abs.1 c).
[98] Allianzinitiative (2016, §3 Abs.1 c).
[99] Allianzinitiative (2016, §3 Abs.1 c).

manner and form most useful to the Licensee and Authorised Users."[100]). Bibliotheken können durch die Verhandlung solcher Klauseln ihren Nutzern das TDM der lizenzierten Inhalte ermöglichen.

4.2 Bibliotheken als Anbieter

Bibliotheken treten nicht nur als Vermittler zwischen Anbietern und Nutzern auf, sondern sind häufig selbst Anbieter von Inhalten: als Betreiber von Univeritätsverlagen, Dokumenten-, Hochschulschriften- und Publikationsservern, Datenrepositorien und Multimediaplattformen. Hier können Bibliotheken unmittelbar dafür sorgen, dass die angebotenen Inhalte für TDM-Anwendungen möglichst einfach zugänglich gemacht werden. Das heißt einheitliche, maschinenlesbare und möglichst strukturierte Formate, einheitliche und möglichst liberale Lizenzen, die Möglichkeit, Inhalte nach diesen Kriterien auszuwählen sowie Schnittstellen zum automatisierten Download sind hier wichtige Aspekte, die TDM-Vorhaben ermöglichen oder erleichtern können. Eine im Rahmen dieser Arbeit durchgeführte Befragung[101] der Betreiber von (vorwiegend Text-)Dokumentenservern in Deutschland zum Thema TDM zeigt, dass dieses bislang nur wenig Beachtung findet. Für 82,1% der Befragten spielte TDM noch überhaupt keine, für 10,3% nur eine untergeordnete oder zukünftige und nur für 7,7% bereits jetzt eine Rolle bei Konzeption und Betrieb des Repositoriums. Auch Anfragen von Wissenschaftlerinnen, die die angebotenen Inhalte für TDM-Vorhaben nutzen wollten, sind eher selten. Bei 69,0% gab es bisher noch keine diesbezüglichen Anfragen von Wissenschaftlerinnen und auch die Übrigen (28,2%[102]) wurden weniger als fünfmal angefragt. Auf der anderen Seite ist auch der Umgang mit dieser Art von Anfragen bei den Betreibern der Repositorien sehr unterschiedlich. Während in einem Fall alle TDM-Anfragen negativ beschieden wurden, wurden an anderer Stelle auch TDM-Aktivitäten ohne vorherige Anfrage beobachtet und toleriert.

Das vorwiegende und zum Teil ausschließliche Format, in dem die (Text-)Inhalte zur Verfügung gestellt werden, ist PDF. Somit ist die Anforderung einheitlicher Formate hier in den meisten Fällen erfüllt. Als unstrukturiertes Format ist PDF jedoch für automatisierte Verfahren wie TDM häufig problematisch und wird von vielen TDM-Anwendern sehr negativ bewertet. Dies belegen Aussagen wie: „Das PDF-File Format ist der Fluch der Chemie, weil automatisierte Auswertung schwierig bis nahezu unmöglich!"[103] Auch wenn viele Repositorien (61,5%) andere Formate grundsätzlich ermöglichen, ist PDF praktisch

[100] Allianzinitiative (2016, §3 Abs.1 c).
[101] Für die Details und vollständigen Ergebnisse der Umfrage siehe Drees (2016).
[102] 2,6% machten keine Angaben.
[103] Sens et al. (2015a, S.29).

überall das mit Abstand dominierende Format und die Vorteile strukturierter Formate werden gegenüber Autoren nicht aktiv beworben. Ebenso ist die Vergabe von CC-BY-Lizenzen zwar fast immer möglich, wird jedoch ebenfalls nur selten aktiv beworben oder als bevorzugte Lizenz vorgegeben. Daher überwiegen in vielen Repositorien Dokumente, die unter keiner Open Access-Lizenz veröffentlicht sind und ausschließlich im Rahmen des deutschen Urheberrechts genutzt werden können. Eine Verwendung von CC-BY-Lizenzen ist häufig jedoch auch deshalb problematisch, da es sich bei vielen Dokumenten um Zweitveröffentlichungen (Green Open Access) handelt und diese somit nicht unbedingt unter einer freien Lizenz angeboten werden dürfen. Auch eine Sucheinschränkung nach Datenformat oder Lizenz wird nur selten (17,9%[104]) angeboten. Immerhin bieten 89,7% der ausgewerteten Repositorien eine OAI-Schnittstelle zum Download der Metadaten an; 17,9% bieten darüber hinaus noch weitere Schnittstellen an. Einige Befragte äußerten sich dahingehend, dass der Fokus beim Aufbau der Plattform auf Sichtbarkeit und dauerhafter Auffindbarkeit lag und TDM erst in einem weiteren, zukünftigen Schritt Berücksichtigung fände. Auch wurde die Vermutung geäußert, der betriebene Publikationsserver biete für TDM-Anwendungen zu wenige und zu unbedeutende Inhalte.

4.3 Bibliotheken als Anwender

TDM-Methoden sind nicht nur im wissenschaftlichen Kontext von Interesse, sie können auch aktiv von Bibliotheken genutzt werden, um damit ihre Dienstleistungen zu verbessern oder neue Services zu entwickeln. Rudolf Mumenthaler nennt zwei Hauptanwendungsgebiete für TDM in Bibliotheken: „die Verbesserung der Suche sowie die Unterstützung bei der Beschlagwortung."[105] So hat beispielsweise die Deutsche Nationalbibliothek bereits im Jahr 2009 im Rahmen des Projektes PETRUS begonnen „die automatische Anreicherung von deutschsprachigen Netzpublikationen mit Schlagwörtern eines kontrollierten Vokabulars"[106] umzusetzen. Dies geschah mithilfe der Averbis Extraction Platform auf der Grundlage der GND als kontrolliertem Vokabular. Auch im AV-Portal[107] der TIB findet, neben weiteren Analysen, eine automatische Verschlagwortung statt. Die audiovisuellen Inhalte werden anhand von Bildmerkmalen automatisch segmentiert, geschriebener und gesprochener Text als OCR- und Audiotranskript extrahiert und diese durch Named Entity Recognition mit Sachbegriffen der GND verschlagwortet.[108] Zudem wurden durch Zuordnung der GND-Begriffe zu anderen Normdaten (u.a. DBpedia und Library of Congress Subject Headings) englischsprachige Bezeichner

[104] 5,1% bieten eine Einschränkung nach Format und Lizenz, 12,1% nur nach einem der beiden Kriterien an.
[105] Mumenthaler (2015).
[106] Uhlmann (2013, S.27).
[107] https://av.tib.eu/
[108] Vgl. Strobel & Plank (2014, S.254f).

ermittelt, wodurch eine zweisprachige (deutsch, englisch) Verschlagwortung ermöglicht wurde.[109] Die automatisch generierten Metadaten komplementieren dabei die manuell erstellten Metadaten dergestalt, dass letztere eine zuverlässige aber grobkörnigere, erstere eine weniger zuverlässige aber feingranularere Beschreibung des Videos liefern, wodurch „eine zielgenaue, segmentbasierte Suche“ ermöglicht wird.[110]

Eine weitere Anwendung, die in der Literatur manchmal als Bibliomining[111] bezeichnet wird, ist die Analyse von Daten aus Bibliothekssystemen (z.B. Nutzungs- oder Bestandsdaten), um Bibliotheksangebote zu verbessern. Siguenza-Guzman et al. bieten einen Überblick über die Bibliomininglitteratur zwischen 1998 und 2014, wonach die meisten Anwendungen zur Analyse des Nutzerverhaltens verwendet werden.[112] Hierunter fallen unter anderem sogenannte bibliothekarische Empfehlungsdienste, die dem Nutzer Empfehlungen der Art “Andere Nutzer interessierten sich auch für: ...” im Bibliothekskatalog anzeigen. Ein solcher Dienst ist das kommerzielle Produkt Bibtip,[113] das an vielen Bibliotheken eingesetzt wird und aus einem DFG-geförderten Projekt an der Universität Karlsruhe hervorging. Empfehlungsdienste sind Kataloganreicherungen, die die Sacherschließung ergänzen und dem Nutzer die Möglichkeit von Serendipitätseffekten bzw. assoziativer Literaturrecherche bieten.[114] Bibtip ist ein verhaltensbasierter Recommender, der analysiert, welche Titel von Nutzern während einer Recherche aufgerufen werden. Titel die häufig gemeinsam aufgerufen werden, werden dann wechselseitig empfohlen.[115]

5. Fazit

TDM gewinnt als wissenschaftliche Methode zunehmend an Bedeutung. Die automatisierte Gewinnung von Forschungsdaten und das damit verbundene Aufkommen riesiger Datenmengen machen in vielen Bereichen automatisierte Datenanalysen notwendig. Gleichermaßen macht das stetige Anwachsen wissenschaftlicher Publikationen und Informationen computergestützte Methoden des Information Retrieval, der Informationserschließung und -klassifizierung sowie der Informationsextraktion immer wichtiger. Nicht nur Wissenschaftler betonen den Nutzen von TDM[116] auch Verleger erwarten laut einer

[109] Vgl. Strobel & Plank (2014, S.257), und Strobel (2014).
[110] Strobel & Plank (2014, S.257).
[111] „Bibliomining is the application of statistical and pattern-recognition tools to large amounts of data associated with library systems in order to aid decision-making or justify services.“ Nicholson (2003, S.1).
[112] Siguenza-Guzman et al. (2015, S.502).
[113] http://www.bibtip.com/de
[114] Vgl. Mönnich & Spiering (2008, S.54).
[115] Vgl. Mönnich & Spiering (2008, S.55f).
[116] 63% der befragten Wissenschaftler messen TDM einen mittleren, großen oder ausschlaggebenden Nutzen bei. Sens et al. (2015a, S.7).

Studie von 2011 mehrheitlich, dass TDM sich als Forschungsmethode rasant ausbreiten wird.[117]

Vor dem Hintergrund dieses allseits artikulierten Bedarfs ist die geringe Anzahl der TDM bezogenen Anfragen bei den Anbietern erstaunlich.[118] Der tatsächliche Bedarf an TDM zugänglichen Ressourcen kann und muss jedoch deutlich höher eingeschätzt werden. So schreibt Okerson:

> „The reported numbers of requests for data mining are extremely small. (...) that explicit demand is quite a bit higher, and potential demand is VERY much higher. I surmise that there are interested researchers who are simply doing their own work in ways that don't get on the radar. Let me just say again that in my view demand will rise much higher very soon."[119]

Diese Einschätzung wird unter anderem durch die Beobachtungen unangemeldeter TDM-Aktivitäten – sowohl bei Verlagsangeboten[120] als auch auf institutionellen Dokumentenservern und Repositorien [121]– bekräftigt.

Wenn diese Einschätzung eines hohen aber nur selten artikulierten Bedarfs zutreffend ist, stellt sich die Frage, was Bibliotheken unternehmen können und sollten, um hier die Wissenschaft adäquat zu unterstützen? Vor dem Hintergrund der oben aufgezeigten Hindernisse sollten sie dazu beitragen, eine Standardisierung und Flexibilisierung zu schaffen! Das heißt, der Idealzustand wäre eine Plattform, die alle Publikationen und Daten im Open Access und in möglichst strukturiertem und untereinander einheitlichem Format anbietet (Standardisierung). Statt einer Einheitslizenz, die z.B. kommerzielles TDM ausschließt oder nur bestimmte Arten des TDM erlaubt, sollten möglichst liberale und weitreichende TDM-Rechte eingeräumt[122] und verschiedene parallele technische Zugangs- und Downloadmöglichkeiten ermöglicht werden, evtl. ergänzt (aber nicht ersetzt) durch vorgefertigte TDM-Tools (Flexibilisierung); eine Vision ähnlich der von Björn Brembs geforderten "World Library of Science": „All that would be required is some sort of standard, which would make the databases in each library interoperable with each other. Why isn't there a 'World Library of Science' which contains all the scientific literature and primary data?"[123] Konkret sollten Bibliotheken darauf achten, alle angebotenen Inhalte –

[117] „80.6% of the respondents is (somewhat or very much) in agreement with the statement that content mining will rapidly expand into new areas, new applications and new directions." Smit & van der Graaf (2011, S.82).

[118] vgl. Abschnitt 4.2 und Smit & van der Graaf, bei deren Umfrage lediglich 21% der befragten Verleger von mehr als 10 Anfragen berichteten. Smit & van der Graaf (2011, S.105).

[119] Okerson (2013, S.6).

[120] „Respondents also note a fair amount of illegal crawling and downloading that suggest unreported mining activities." Smit & van der Graaf (2011, S.1).

[121] vgl. Abschnitt 4.2

[122] Für den Fall einer europaweiten Urheberrechtsschranke für TDM im Sinne der Hague Declaration wäre die rechtliche Frage geklärt und entsprechende Lizenzen nicht notwendig

[123] Brembs (2011).

die eigenen sowie bei Verlagen lizenzierte – den Wissenschaftlerinnen derart zugänglich zu machen, dass Hindernisse für TDM minimiert werden.

Diese Aufgabe, die Bereitstellung und Nutzbarmachung von Informationen, ist die Kernaufgabe und Kernkompetenz wissenschaftlicher Bibliotheken. Text und Data Mining als eine neue, automatisierte Form der Informationsnutzung stellt somit neue Herausforderungen, aber auch neue Möglichkeiten für Bibliotheken dar. Dabei ist TDM keine Methode, die auf einige wenige Fachdisziplinen beschränkt ist, sondern eine Fächergrenzen überschreitende, weiter wachsende Forschungsmethode. Daher ist es wichtig, auf Seiten der Bibliothek entsprechende Kompetenzen auszubilden, um TDM-Bedarfe der Wissenschaft sowohl zu erkennen als auch zu verstehen. Nur so können Bibliotheken, in enger Zusammenarbeit mit Wissenschaftlern, TDM durch günstige Rahmenbedingungen ermöglichen, Wissenschaftler bei TDM-Vorhaben aktiv unterstützen und darüber hinaus eigene, neue Angebote entwickeln. Wie Bernard Reilly es formuliert: „Librarians can play a critical role in this process but only if they fully understand the practices of their constituents and integrate that understanding into their licensing and resource development work.“[124]

[124] Reilly (2012, S.76).

Literatur

Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining Association Rules Between Sets of Items in Large Databases. *SIGMOD Rec.* 22(2), 207–216. doi:10.1145/170036.170072.

Allianzinitiative. (2016). Musterlizenzvertrag. http://www.allianzinitiative.de/fileadmin/user_upload/www.allianzinitiative.de/Allianz_Musterlizenz_2016.pdf (abgerufen am 09. 10. 2016).

Association of Research Libraries. (2015). Issue brief : Text and data mining and fair use in the united states. http://www.arl.org/storage/documents/TDM-5JUNE2015.pdf (abgerufen am 09. 10. 2016).

Blessing, A., Kliche, F., Heid, U., Kantner, C., & Kuhn, J. (2015). Computerlinguistische Werkzeuge zur Erschließung und Exploration großer Textsammlungen aus der Perspektive fachspezifischer Theorien. In C. Baum & T. Stäcker (Hrsg.), *Grenzen und Möglichkeiten der Digital Humanities*. doi:10.17175/sb001_013.

Blöhdorn, S. & Hotho, A. (2009). Ontologies for Machine Learning. In S. Staab & R. Studer (Hrsg.), *Handbook on Ontologies* (S. 637–662). Dordrecht, u.a.: Springer. doi:10.1007/978-3-540-92673-3.

Borne, K. (2009). Scientific data mining in astronomy. arXiv: 0911.0505v1.

Brembs, B. (2011). A proposal for the library of the future. http://blogarchive.brembs.net/comment-n717.html (abgerufen am 09. 10. 2016).

Carstensen, K.-U., Ebert, C., Jekat, S., Klabunde, R., & Langer, H. (Hrsg.). (2010). *Computerlinguistik und Sprachtechnologie: Eine Einführung*. Heidelberg: Spektrum Akademischer Verlag, Springer.

Chi, G., Liu, Y., Wu, Z., & Wu, H. (2015). Ghost cities analysis based on positioning data in China. arXiv: 1510.08505v2.

Clare, A. & King, R. D. [R. D.]. (2003). Predicting gene function in Saccharomyces cerevisiae. *Bioinformatics*, 19(suppl 2), ii42–ii49. doi:10.1093/bioinformatics/btg1058.

Clark, J. (2013). Text mining and scholarly publishing. http://publishingresearchconsortium.com/index.php/prc-guides-main-menu/158-prc-guide-text-mining-and-scholarly-publishing (abgerufen am 09. 10. 2016).

Coenen, F. (2011). Data mining: past, present and future. *The Knowledge Engineering Review*, 26, 25–29. doi:10.1017/S0269888910000378.

Copyright Clearing Center. (2015). RightFind™XML for Mining Service. http://www.copyright.com/rightsholders/rightfind-xml-for-mining-service/ (abgerufen am 09. 10. 2016).

CrossRef. (2015). Crossref Text and Data Mining Services. http://tdmsupport.crossref.org/ (abgerufen am 09. 10. 2016).

Al-Daihani, S. M. & Abrahams, A. (2016). A Text Mining Analysis of Academic Libraries' Tweets. *The Journal of Academic Librarianship*. doi:10.1016/j.acalib.2015.12.014.

Drees, B. (2016). Text und Data Mining an wissenschaftlichen Repositorien und Publikationsservern in Deutschland – Zusammenfassung der Ergebnisse einer Umfrage im Februar und März 2016.

doi:10.11588/data/10090.

Europäisches Parlament. (2015). Richtlinie 2001/29/EG des Europäischen Parlaments und des Rates vom 22. Mai 2001 zur Harmonisierung bestimmter Aspekte des Urheberrechts und der verwandten Schutzrechte in der Informationsgesellschaft. http://eur-lex.europa.eu/legal-content/DE/TXT/PDF/?uri=CELEX:32001L0029&from=DE (abgerufen am 09. 10. 2016).

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, 17, 37–54. doi:10.1609/aimag.v17i3.1230.

Gotter, L. (o.D.). Wissensexploration: Aufgaben und Methoden des Data Mining. http://wissensexploration.de/datamining-kdd-aufgaben-methoden.php (abgerufen am 09. 10. 2016).

Haarmann, B. (2014). *Ontology on demand: Volautomatische Ontologieerstellung aus deutschen Texten mithilfe moderner Textmining-Prozesse*. Berlin: epubli.

Han, J., Lee, J.-G., & Kamber, M. (2009). An overview of clustering methods in geographic data analysis. In H. J. Miller & J. Han (Hrsg.), *Geographic data mining and knowledge discovery* (S. 149–188). Boca Raton: Taylor & Francis.

Handke, C., Guibault, L., & Vallbé, J.-J. (2015). Is Europe falling behind in data mining? Copyright's impact on data mining in academic research. *SSRN*. doi:10.2139/ssrn.2608513.

Hartgerink, C. H. J. (2015). Elsevier stopped me doing my research. http://onsnetwork.org/chartgerink/2015/11/16/elsevier-stopped-me-doing-my-research/ (abgerufen am 09. 10. 2016).

Hartgerink, C. H. J. (2016). Wiley also stopped me doing my research. http://onsnetwork.org/chartgerink/2016/02/23/wiley-also-stopped-my-doing-my-research/ (abgerufen am 09. 10. 2016).

Helmholtz Open Access Koordinationsbüro. (2013). Rechtliche Aspekte von Text und Data Mining. Helmholtz Open Science Briefing. http://os.helmholtz.de/fileadmin/user_upload/os.helmholtz.de/Dokumente/helmholtz_osb_tdm.pdf (abgerufen am 09. 10. 2016).

Hotho, A., Nürnberger, A., & Paaß, G. (2005). A brief survey of text mining. *LDV-Forum*, 20(1), 19–62.

Intellectual Property Office. (2014). Exceptions to copyright: Text and data mining for non-commercial research. https://www.gov.uk/guidance/exceptions-to-copyright#text-and-data-mining-for-non-commercial-research (abgerufen am 09. 10. 2016).

Issenberg, S. (2012). How president Obama's campaign used big data to rally individual voters. http://www.technologyreview.com/featuredstory/509026/how-obamas-team-used-big-data-to-rally-voters (abgerufen am 09. 10. 2016).

King, R. D. [Ross D.], Karwath, A., Clare, A., & Dehaspe, L. (2000). Accurate prediction of protein functional class from sequence in the Mycobacterium tuberculosis and Escherichia coli genomes using data mining. *Yeast*, 17(4), 283–293. doi:10.1002/1097-0061(200012)17:4<283::AID-YEA52>3.0.CO;2-F.

Koh, H. C. & Tan, G. (2005). Data mining applications in healthcare. *Journal of Healthcare Information*

Management, 19(2), 64–72.

Kumar, V. D. & Tipney, H. J. (Hrsg.). (2014). *Biomedical Literature Mining*. New York: Humana Press, Springer.

Langlais, P.-C. (2016). Will France get its text and data mining exception? http://scoms.hypotheses.org/602? (abgerufen am 09. 10. 2016).

Larose, D. T. & Larose, C. D. (2014). *Discovering knowledge in data* (second edition). Hoboken, New Jersey: John Wiley & Sons, Inc.

Liao, S.-H., Chu, P.-H., & Hsiao, P.-Y. (2012). Data mining techniques and applications – A decade review from 2000 to 2011. *Expert Systems with Applications*, 39(12), 11303–11311. doi:10.1016/j.eswa.2012.02.063.

LIBER. (2015). The Hague Declaration. http://thehaguedeclaration.com/wp-content/uploads/sites/2/2015/04/Liber_DeclarationA4_2015.pdf (abgerufen am 09. 10. 2016).

Mehler, A. & Wolff, C. (2005). Einleitung: Perspektiven und Positionen des Text Mining. *LDV-Forum*, 20(1), 1–18.

Molloy, J. (2016). FutureTDM is mapping the legal barriers to TDM in Europe. http://project.futuretdm.eu/2016/02/19/futuretdm-is-mapping-the-legal-barriers-to-tdm-in-europe/ (abgerufen am 09. 10. 2016).

Mönnich, M. & Spiering, M. (2008). Erschließung. Einsatz von BibTip als Recommendersystem im Bibliothekskatalog. *Bibliotheksdienst*, 1(42), 54–59. doi:10.1515/bd.2008.42.1.54.

Mumenthaler, R. (2015). Trend und Herausforderung: Text and Data Mining. http://ruedimumenthaler.ch/2015/06/09/trend-und-herausforderung-text-and-data-mining/ (abgerufen am 09. 10. 2016).

Murray-Rust, P. (o.D.). Contentmine. http://contentmine.org/ (abgerufen am 09. 10. 2016).

Murray-Rust, P. (2016). Zika in scientific literature. https://www.youtube.com/watch?v=5lYzOZ2Cv_I (abgerufen am 09. 10. 2016).

Nicholson, S. (2003). The Bibliomining Process: Data Warehousing and Data Mining for Library Decision-Making. *Information Technology and Libraries*, 22(4). http://hdl.handle.net/10150/106392 (abgerufen am 09. 10. 2016).

Okerson, A. (2013). Text & data mining - A librarian overview. http://library.ifla.org/252/1/165-okerson-en.pdf (abgerufen am 09. 10. 2016).

OpenForum Academy. (2015). OFA White Paper – How to unleash the innovative potential of text and data mining in the EU. http://www.openforumeurope.org/wp-content/uploads/2015/12/White-Paper-TDM-1.pdf (abgerufen am 09. 10. 2016).

Pan American Health Organization & World Health Organization. (2016). Suspected and confirmed Zika cases reported by countries and territories in the Americas, 2015-2016. Updated as of 25 February 2016, with data received by 24 February 2016. http://ais.paho.org/phip/viz/ed_zika_epicurve.asp (abgerufen am 06. 03. 2016).

Petersohn, H. (2005). *Data Mining: Verfahren, Prozesse, Anwendungsarchitektur*. München: Oldenbourg Wissenschaftsverlag.

Piatetsky-Shapiro, G. (2007). Data mining and knowledge discovery 1996 to 2005: overcoming the hype and moving from "university" to "business" and "analytics". *Data Mining and Knowledge Discovery*, 15(1), 99–105. doi:10.1007/s10618-006-0058-2.

Piegorsch, W. W. (2015). *Statistical data analysis: Foundations for data mining, informatics, and knowledge discovery*. Chichester: Wiley.

Reilly, B. F. (2012). When machines do research, part 2: Text-Mining and libraries. *Charleston Advisor*, 75–76. doi:10.5260/chara.14.2.75.

Saffer, J. D. & Burnett, V. L. (2014). Introduction to biomedical literature text mining: context and objectives. In V. D. Kumar & H. J. Tipney (Hrsg.), *Biomedical Literature Mining* (S. 1–7). New York: Humana Press, Springer.

Science Europe. (2015). Briefing Paper: Text and data mining and the need for a science-friendly EU copyright reform. http://www.scienceeurope.org/wp-content/uploads/2015/04/SE_Briefing_Paper_textand_Data_web.pdf (abgerufen am 09. 10. 2016).

Sens, I., Katerbow, M., Schöch, C., & Mittermaier, B. (2015a). Umfrageergebnisse „Bedarf und Anforderungen an Ressourcen für Text und Data Mining". doi:10.5281/zenodo.32583.

Sens, I., Katerbow, M., Schöch, C., & Mittermaier, B. (2015b). Zusammenfassung Workshop und Umfrageergebnisse „Bedarf und Anforderungen an Ressourcen für Text und Data Mining". doi:10.5281/zenodo.32584.

Siguenza-Guzman, L., Saquicela, V., Avila-Ordonez, E., Vandewalle, J., & Cattrysse, D. (2015). Literature review of data mining applications in academic libraries. *The Journal of Academic Librarianship*, 41, 499–510. doi:10.1016/j.acalib.2015.06.007.

Sleator, R. D. (2012). Functional Genomics: Methods and Protocols. In M. Kaufmann & C. Klinger (Hrsg.), (Kap. Prediction of Protein Functions, S. 15–24). New York, NY: Springer. doi:10.1007/978-1-61779-424-7_2.

Smit, E. & van der Graaf, M. (2011). Journal Article Mining : A research study into practices, policies, plans ... and promises. Amsterdam: Commissioned by the Publishing Research Consortium. http://publishingresearchconsortium.com/index.php/128-prc-projects/research-reports/journal-article-mining-research-report/160-journal-article-mining (abgerufen am 09. 10. 2016).

Stray, J. (2010). Wikileaks Iraq: how to visualise the text. *The Guardian*. http://www.theguardian.com/news/datablog/2010/dec/16/wikileaks-iraq-visualisation (abgerufen am 09. 10. 2016).

Strobel, S. (2014). Englischsprachige Erweiterung des TIB|AV-Portals. Ein GND/DBpedia-Mapping zur Gewinnung eines englischen Begriffssystems. *o-bib. Das offene Bibliotheksjournal / herausgegeben vom VDB*, 1(1), 197–204. doi:10.5282/o-bib/2014H1S197-204.

Strobel, S. & Plank, M. (2014). Semantische Suche nach wissenschaftlichen Videos – Automatische Verschlagwortung durch Named Entity Recognition. *ZfBB*, 4/5(61), 254–258. doi:10.3196/1864295014614515 4.

Swanson, D. R. (1986). Fish oil, raynaud's syndrome, and undiscovered public knowledge. *Perspectives in biology and medicine*, 30(1), 7–18.

Swanson, D. R. (1988). Migraine and magnesium : Eleven neglected connections. *Perspectives in biology and medicine*, 31(4), 526–557. doi:10.1353/pbm.1988.0009.

Swanson, D. R. (1991). Complementary Structures in Disjoint Science Literatures. In *Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (S. 280–289). SIGIR '91. Chicago, Illinois, USA: ACM. doi:10.1145/122860.122889.

Uhlmann, S. (2013). Automatische Beschlagwortung von deutschsprachigen Netzpublikationen mit dem Vokabular der Gemeinsamen Normdatei (GND). *Dialog mit Bibliotheken*, 2(1), 26–36. http://nbn-resolving.de/urn:nbn:de:101-20140305238 (abgerufen am 09. 10. 2016).

Weeber, M., Vos, R., Klein, H., de Jong-van den Berg, L. T. W., Aronson, A. R., & Molema, G. (2003). Generating Hypotheses by Discovering Implicit Associations in the Literature: A Case Report of a Search for New Potential Therapeutic Uses for Thalidomide. *Journal of the American Medical Informatics Association*, 10(3), 252–259. doi:10.1197/jamia.M1158.