Computerunterstützte Inhaltserschließung

Bericht über einen Workshop an der UB Stuttgart – mit einem Exkurs zum neuen Inhaltserschließungskonzept der DNB

Heidrun Wiesenmüller, Hochschule der Medien Stuttgart
Imma Hinrichs, IZUS/Universitätsbibliothek Stuttgart

Am 8. und 9. Mai 2017 fand an der Universitätsbibliothek Stuttgart ein Workshop zum Thema „Computerunterstützte Inhaltserschließung" mit über 50 Teilnehmenden statt.[1] Ausgangspunkt der Veranstaltung war die Einführung des sogenannten „Digitalen Assistenten" – eines Werkzeugs zur maschinellen Unterstützung der verbalen Sacherschließung – an der UB Stuttgart und weiteren Bibliotheken in Baden-Württemberg. Doch ging der Workshop auch darüber hinaus und bot die Gelegenheit, sich grundsätzlicher mit der Frage nach dem Wert von Inhaltserschließung und nach der aktuellen und künftigen Rolle von intellektuellen, halbautomatischen und vollautomatischen Verfahren auseinanderzusetzen. Alle Vortragsfolien stehen auf der Website der UB Stuttgart zur Verfügung.[2]

Keynote: „Erschließung in schwierigen Zeiten"

Die Keynote von Heidrun Wiesenmüller (HdM Stuttgart) stand unter dem Titel „Erschließung in schwierigen Zeiten – Ansichten und Einsichten" und begann mit der Erkenntnis, dass es nicht gerade „vergnügungssteuerpflichtig" sei, in der Erschließung tätig zu sein. Diese habe wenig Lobby; gerade die Sacherschließung werde primär als Kostenfaktor betrachtet, wohingegen ihr Nutzen kaum wahrgenommen werde. Als Strategien, um die Sacherschließung „aus der Defensive" zu bringen, führte Wiesenmüller zuerst das selbstkritische Hinterfragen der eigenen Aktivitäten an: Man müsse überlegen, was in der Vergangenheit „schief gelaufen" sei und weshalb naheliegende Verbesserungen so häufig scheiterten. Wichtig sei es aber auch, sich auf den Wert von Erschließung zu besinnen: Diesen dürfe man nicht für selbstverständlich halten, sondern müsse ihn aktiv bewerben – auch bei Entscheidungsträger/inne/n. Zu verbessern sei außerdem das Kosten-Nutzen-Verhältnis: Dafür müssten einerseits Potenziale für die Nutzung erkannt und konkrete Optimierungen angestoßen werden, andererseits sei aber auch der Aufwand zu verringern. Dies betreffe sowohl einen rationelleren Umgang mit der intellektuellen Erschließung (noch immer werden unter Umständen dieselben Ressourcen an mehreren Stellen nach der gleichen Methode sachlich erschlossen) als auch das Zusammenspiel zwischen intellektueller Erschließung und automatischen Verfahren.

Ausführlich beschäftigte sich Wiesenmüller mit der Frage, ob Daten und Suchsysteme „zusammenpassen". Vieles sei in heutigen Bibliothekskatalogen suboptimal gelöst, werde aber klaglos hingenommen. Beispielsweise erscheint die Anzeige von Sacherschließungsinformationen erst bei der Volltrefferanzeige, und auch dort in der Regel weit unten – anders als im Sachkatalog in Zettelform, wo diese

1 Das Programm des Workshops kann auf der Website der UB Stuttgart abgerufen werden: „Workshop ‚Computerunterstützte Inhaltserschließung' am 8./9. Mai 2017 in der UB Stuttgart," zuletzt geprüft am 14.08.2017, http://blog.ub.uni-stuttgart.de/2017/04/workshop-computerunterstuetzte-inhaltserschliessung/.

2 „Workshop ‚Computerunterstützte Inhaltserschließung' am 8./9. Mai 2017 in der UB Stuttgart: Vortragsfolien," zuletzt geprüft am 14.08.2017, http://blog.ub.uni-stuttgart.de/veranstaltungen/workshop-computerunterstuetzte-inhaltserschliessung/. Auf einen Einzelnachweis der jeweiligen Vortragsfolien wird im Folgenden verzichtet.

Angaben ganz oben (im Kopf der Karten) zu sehen waren. Unterbegriffe, auf die im Zettelkatalog mit Verweisungskarten hingewiesen wurde, stehen in den meisten Online-Katalogen entweder gar nicht oder nur auf sehr umständlichem Weg für die Recherche zur Verfügung. Wiesenmüller empfahl, den gesamten „Reichtum" der vorhandenen Daten in die Kataloge zu bringen und diese benutzerfreundlich und zeitgemäß anzubieten. Dazu gehöre es auch, sich von „zu vielen verwirrenden Angeboten zum Anklicken" zu verabschieden. Gefragt seien stattdessen überwiegend im Hintergrund laufende Assistenzsysteme und der Wechsel von einem Hol- zu einem Bring-Ansatz. Anstatt von Nutzerinnen und Nutzern zu erwarten, dass sie zum richtigen Zeitpunkt auf den richtigen Link klicken, müssten ihnen automatisch sinnvolle Angebote gemacht werden (z.B. „Möchten Sie auch Treffer zu den Themen … sehen?").

Zur Illustration erinnerte die Referentin an die vor einigen Jahren entwickelte und in den Katalogen der UB Heidelberg und UB Mannheim implementierte geografische Facette.[3] Diese beruht auf einer Auswertung der Ländercodes in den verknüpften Normdatensätzen und ergibt jeweils deutlich mehr relevante Treffer als eine direkte Suche nach dem gewünschten Geografikum. Obwohl die Technik einwandfrei funktioniert, hat sie sich nicht in weiteren Katalogen verbreitet. Dies liege vielleicht daran, dass die Umsetzung normale Nutzerinnen und Nutzer überfordere. Denn sie müssten zunächst bewusst auf die Eingabe des Geografikums verzichten und ihre Suche in einem zweiten Schritt über die Geo-Facette einschränken. Ein wirklich nutzerfreundliches System würde hingegen die Eingabe eines Geografikums im Suchfeld automatisch in eine Recherche über den Ländercode umsetzen. Darüber hinaus müsse auch an die heterogenen Daten in Resource Discovery Systemen (RDS) gedacht werden. Bei deren Aufbereitung biete sich – so Wiesenmüller – eine große Chance für die Erschließung.[4] Um die RDS-Daten selbst durch ein bibliothekarisches Metadatenmanagement verbessern zu können, seien allerdings neue Formen der Zusammenarbeit mit kommerziellen Anbietern nötig – oder man „emanzipiere" sich mit einem eigenen Angebot, wie dies die UB Leipzig vorgemacht habe.[5]

Derzeit definierten Bibliotheken ihre Rolle neu. Aus der Befürchtung heraus, dass das Kerngeschäft mehr und mehr verloren gehen könnte, halte man nach neuen Tätigkeitsfeldern Ausschau. Allerdings: Zwar könnten Bereiche wie Lernräume, Makerspaces, Publikationsdienste etc. von Bibliothekar/inn/en bespielt werden, doch sei durchaus auch eine Realisierung ohne die Beteiligung von Bibliotheken möglich. Deshalb dürfe man neue und alte Aktionsfelder nicht gegeneinander ausspielen. Kernkompetenzen wie die Erschließung sollten nicht leichtfertig aufgegeben, sondern vielmehr ihre Anwendungsbereiche erweitert werden. Der Wert der bibliothekarischen Daten, die längst auch im Semantic Web angekommen seien, werde selbst von Bibliothekar/inn/en noch immer unterschätzt.

3 Vgl. Heidrun Wiesenmüller, Leonhard Maylein und Magnus Pfeffer, „Mehr aus der Schlagwortnormdatei herausholen: Implementierung einer geographischen Facette in den Online-Katalogen der UB Heidelberg und der UB Mannheim," *B.I.T. online* 14, Nr. 3 (2011): 245–252, http://nbn-resolving.de/urn/resolver.pl?urn=urn:nbn:de:bsz:16-opus-125556.

4 Vgl. Magnus Pfeffer und Heidrun Wiesenmüller, „Resource Discovery Systeme," in *Handbuch Informationskompetenz*, hrsg. Wilfried Sühl-Strohmenger, 2. Auflage (Berlin, Boston: De Gruyter Saur, 2016), 105–114, http://doi.org/10.1515/9783110403367-012, hier 113.

5 Vgl. Jens Lazarus, „Das machen wir selbst: Der Aufbau eines eigenen Artikelindex als Alternative zu proprietären Angeboten" (Vortrag auf dem 6. Bibliothekskongress in Leipzig am 14.03.2016), http://nbn-resolving.de/urn/resolver.pl?urn:nbn:de:0290-opus4-23591.

Als Erfolgsgeschichte benannte die Referentin die Ausbreitung der Gemeinsamen Normdatei (GND) über den bibliothekarischen Bereich hinaus. So gibt es z.B. in der Wikipedia knapp 400.000 Links zur GND. Ein Beispiel für die konkrete Anwendung sind die Personenseiten in der Deutschen Digitalen Bibliothek. Diese basieren auf dem Datendienst „Entity Facts" der Deutschen Nationalbibliothek und bieten sowohl Informationen aus der GND als auch Verlinkungen zu verschiedenen Systemen, welche mit GND-Nummern arbeiten.

Abschließend setzte sich die Referentin mit der Automatisierung auseinander. Nach ihrer Einschätzung ist derzeit die maschinell unterstützte Erschließung der Königsweg, der das Beste aus beiden Welten vereinigt: die Qualität der intellektuellen Erschließung und den möglichst rationellen Einsatz der Ressource Mensch. Die Frage sei jedoch, ob dies eine dauerhafte Lösung ist oder nur ein Zwischenschritt auf dem Weg zur Vollautomatisierung. Die Kombination von „Big Data" mit künstlicher Intelligenz und maschinellem Lernen habe in jüngster Vergangenheit dazu geführt, dass Maschinen Nachrichtenmeldungen schreiben, an der Börse handeln und Krebs diagnostizieren können. Anders als früher seien nicht nur Routinetätigkeiten von Automatisierung betroffen, sondern auch viele Aufgaben, für die traditionell hochqualifiziertes Personal eingesetzt wurde.[6] Die Bayerische Staatsbibliothek teste derzeit ein auf künstlicher Intelligenz beruhendes Recherchewerkzeug (Yewno).[7]

Jedoch gebe es auch manches, was gegen das Szenario einer bevorstehenden Vollautomatisierung spreche. So seien die Ergebnisse maschineller Übersetzungswerkzeuge noch immer ernüchternd; auch die Qualität automatischer Erschließungsverfahren sei weiterhin begrenzt. Für generell nicht automatisierbar hält die Referentin überdies die Weiterentwicklung der Erschließungssysteme. Auch würden stets intellektuell erschlossene Vergleichskorpora als Trainingsmaterial für maschinelle Verfahren benötigt. Schließlich gebe es auch in der Wirtschaft immer mehr Stimmen, die den Trend zur Vollautomatisierung grundsätzlich in Frage stellen. Wichtig sei es, so das Fazit der Referentin, die richtige Balance zu finden: Wo sonst gar keine Erschließung geleistet werden könne, sei eine rein maschinelle Erschließung natürlich sinnvoll. Die intellektuelle Erschließung bleibe jedoch zumindest im Kernbereich weiter wichtig. Sie müsse möglichst rationell erfolgen, wozu auch eine maschinelle Unterstützung gehöre.[8]

6 Vgl. z.B. Martin Ford, *Rise of the Robots: Technology and the Threat of a Jobless Future* (New York: Basic Books, 2015), sowie Carl Benedikt Frey und Michael A. Osborne, *The Future of Employment: How Susceptible are Jobs to Computerisation* (Oxford: Oxford Martin Programme on Technology and Employment, 2013), zuletzt geprüft am 14.08.2017, http://www.oxfordmartin.ox.ac.uk/publications/view/1314.

7 Zu Yewno vgl. Berthold Gillitzer, „Vom Recherchesystem zum inferentiellen Service – ein Paradigmenwechsel? Yewno, ein semantischer Discovery Service im Pilotversuch an der Bayerischen Staatsbibliothek," *ZfBB* 64, Nr. 2 (2017): 71–78, http://doi.org/10.3196/186429501664227. Für eine kritische Auseinandersetzung mit Yewno vgl. Heidrun Wiesenmüller, „Eindrücke vom Bibliothekartag in Frankfurt (Teil 2)," *Basiswissen RDA* (Blog), 30.06.2017, zuletzt geprüft am 14.08.2017, https://www.basiswissen-rda.de/bibliothekartag2017-teil-2/.

8 Vgl. dazu auch Gerhard Stumpf, „'Kerngeschäft' Sacherschließung in neuer Sicht: Was gezielte intellektuelle Arbeit und maschinelle Verfahren gemeinsam bewirken können" (Vortrag auf der VDB-Fortbildungsveranstaltung „Wandel als Konstante: neue Aufgaben und Herausforderungen für sozialwissenschaftliche Bibliotheken" am 22./23. Januar 2015 in Berlin), http://nbn-resolving.de/urn/resolver.pl?urn:nbn:de:bvb:384-opus4-30027.

Der Digitale Assistent

Ein konkretes Beispiel für ein Werkzeug, das die intellektuelle Erschließung mit maschinellen Methoden in hervorragender Weise unterstützt, ist der „Digitale Assistent“ (DA), dem ein großer Teil des Workshops gewidmet war. Der DA ist das Ergebnis einer gelungenen Zusammenarbeit zwischen Bibliotheken – zunächst der ZB Zürich, später auch der UB Stuttgart – und der Schweizer Firma Eurospider Information Technology AG, welche 1995 als „Spin-off“ der ETH Zürich gegründet wurde.

Für den Einsatz in Baden-Württemberg wurde der DA erheblich überarbeitet und erweitert und wird in dieser Form als „DA-2“ bezeichnet. Imma Hinrichs (UB Stuttgart) erläuterte und demonstrierte die Funktionalitäten des Systems.

Abb. 1: Imma Hinrichs erläutert den Digitalen Assistenten. Foto: UB Stuttgart/Frank Wiatrowski

Der DA-2 unterstützt die verbale Erschließung zum einen durch eine komfortable webbasierte Oberfläche und das „Einsammeln“ von Erschließungen, die bereits in anderen Katalogen vorhanden sind (Fremddaten) und die im DA-2 als Vorschläge angezeigt werden. Dabei werden auch Schlagwörter aus anderen Schlagwortsystemen als der GND berücksichtigt und über eine approximative Übersetzung (überwiegend mittels Konkordanztabellen) in GND-Schlagwörter übertragen. Außerdem bietet der DA-2 eine Ähnlichkeitssuche, bei der die Metadaten eines zu erschließenden Titels über einen großen, separat hinterlegten Index mit den Metadaten bereits erschlossener Titel abgeglichen werden. Die Erschließung von bis zu vierzig „ähnlichsten“ Titeln kann angesehen und nachgenutzt werden. Unter der gleichen Oberfläche kann man aber auch selbst GND-Schlagwörter und Formangaben suchen, Schlagwortnormsätze ansehen und sich dazu die Hierarchien sowie verwandte oder über sonstige Beziehungen verknüpfte Schlagwörter anzeigen lassen. Für die Anzeige großer, mehrstufiger Hierarchien kann man über einen Link zur frei zugänglichen „WebGND“ von Eurospider wechseln.

Im DA-2 kann voreingestellt werden, dass immer Schlagwortfolgen vergeben werden sollen; es ist aber auch möglich, ohne die Bildung von Folgen zu verschlagworten.

Auch im laufenden Betrieb wurde der DA-2 ständig verbessert, z.B. durch die Einbindung von Inhaltsverzeichnissen, Klappentexten u. ä. oder beim Wechsel von RAK zu RDA im Bereich der Formschlagwörter bzw. Formangaben. In ihrer vorletzten Folie verwies die Referentin auf Publikationen zum DA-2, insbesondere auf einen Aufsatz in o-bib, in dessen Anhang (S. 173-183) beispielhaft das Vorgehen bei der Verschlagwortung mit dem DA-2 dokumentiert wurde.[9] Die Teilnehmenden am Workshop bekamen den DA-2 jedoch live gezeigt und konnten erleben, wie leicht die Bedienung ist und wie unkompliziert eine Verschlagwortung von der Hand gehen kann.

Armin Kühn vom Bibliotheksservice-Zentrum Baden-Württemberg (BSZ) erläuterte die Rolle des BSZ innerhalb des Gesamtworkflows. Für jede Bibliothek aus dem Südwestdeutschen Bibliotheksverbund (SWB), die den DA-2 nutzen möchte, wird zunächst einmalig eine Grundlieferung ihres Bestands in den DA-2 eingespielt (nach den Vorgaben der Bibliothek). Im laufenden Betrieb liefert das BSZ jede Nacht für alle Teilnehmerbibliotheken (derzeit UB Stuttgart, WLB Stuttgart, UB Tübingen und BLB Karlsruhe) eine Liste mit den Identnummern der neuen Titeldatensätze an Eurospider. Am nächsten Tag stehen diese für die Bearbeitung im DA-2 zur Verfügung. Die im DA-2 vergebenen Schlagwörter werden wiederum an das BSZ zurückgeliefert und in der folgenden Nacht in die Verbunddatenbank eingespielt. Dazu müssen sie aus dem Lieferformat MARC 21 in das Pica-Format des SWB konvertiert werden. Je nach Einstellung im DA-2 kommen die Schlagwörter entweder in die für Schlagwortfolgen vorgesehenen Felder oder in diejenigen für Einzelschlagwörter.

Ein großer Vorteil des Verfahrens ist, dass die Verschlagwortung ausschließlich innerhalb des DA erfolgt; es muss also nicht zwischen zwei Systemen hin und hergewechselt werden. Auch müssen sich die Erschließenden nicht mit den Pica-Feldern und dem Handling des Katalogisierungs-Clients (WinIBW) auseinandersetzen, sondern können die sehr komfortable Oberfläche des DA-2 nutzen. Ebensowenig müssen sie sich darum kümmern, ob ein RAK- oder ein RDA-Datensatz vorliegt – denn innerhalb des DA-2 werden die Formaspekte in beiden Fällen gleich behandelt und als Teil der Schlagwortfolge erfasst. Erst beim Einspielen in den SWB wird ein Unterschied zwischen den beiden Typen gemacht. Handelt es sich noch um einen RAK-Datensatz, so bleiben die Formaspekte als Bestandteil der Schlagwortfolge erhalten (Formschlagwörter). Liegt hingegen ein RDA-Datensatz vor, so werden die Formaspekte aus den Schlagwortfolgen herausgelöst und stattdessen in die speziell dafür vorgesehenen Felder geschrieben.

Lukas Fischer (Eurospider) präsentierte Ergebnisse einer Masterarbeit, die Ursula Jud-Reichlen für ihren Abschluss im Masterstudiengang Bibliotheks- und Informationswissenschaften an der Universität Zürich (2017) angefertigt hat und zu der der Referent die Auswertungen durchgeführt hatte. Ursula Jud-Reichlen untersuchte in ihrer Arbeit die Qualität von GND-Erschließungsdaten, die der

9 Imma Hinrichs, Gérard Milmeister, Peter Schäuble und Helge Steenweg, „Computerunterstützte Sacherschließung mit dem Digitalen Assistenten (DA-2)," *o-bib* 3, Nr. 4 (2016): 156–185, https://doi.org/10.5282/o-bib/2016H4S156-185.

DA-2 als Vorschläge bereitstellt und die aus SWB-eigenen und fremden GND-Erschließungen und approximativen GND-Übersetzungen aus fremden Schlagwortsystemen bestehen. Sie bewertete die GND-Erschließungen randomisierter Stichproben deutsch- bzw. englischsprachiger Titel aus den Bereichen Wirtschaft bzw. Recht und stellte einen eigenen Goldstandard her, mit dem Recall und Precision berechnet wurden. Außerdem wurden die approximativen Übersetzungen und die Ähnlichkeitssuche untersucht. Wie Fischer erläuterte, ergab die Untersuchung von Jud-Reichlen, dass die Qualität der Erschließungsdaten, die der DA-2 bereitstellt, im Bereich Sachschlagwörter sowohl in Hinsicht auf Recall als auch auf Precision gut ist. Um die Erschließung mit Formangaben zu verbessern, schlägt die Autorin automatische Verfahren vor, die z.B. das Inhaltsverzeichnis oder auch das Vorwort einer Publikation nach einschlägigen Wörtern bzw. Merkmalen auswerten, aus denen Formangaben abgeleitet werden können. Die Qualität der approximativen Übersetzungen konnte in der Masterarbeit nicht zu allen Stichprobenmengen ausgewertet werden, jedoch stellte Jud-Reichlen fest, dass die Übersetzungen gute GND-Vorschläge als Ergänzung zu originären GND-Fremddaten liefern können. Die Ähnlichkeitssuche erbringe derzeit allerdings zu 40 % der Titel noch keine Vorschläge. Als Fazit sehe Jud-Reichlen den DA-2 als geeignetes Instrument für die kooperative Erschließung und betone, dass Kooperation und Absprachen, aber auch die Weiternutzung von Erschließungsdaten wichtig seien.

Regine Beckmann von der Staatsbibliothek zu Berlin – Preußischer Kulturbesitz (SBB-PK) berichtete von einem Test des DA-2, den sie mit Mitarbeitenden der Staatsbibliothek durchgeführt hat. Die Ausgangslage stelle sich an der SSB-PK ähnlich dar wie an vielen anderen großen Bibliotheken mit Pflichtexemplarzugang und vielen Sonderabteilungen. Die Möglichkeit zur direkten Fremddatennutzung sei gering, die Zahl elektronischer Publikationen nehme zu und die Personalressourcen würden bei gleichzeitigem Wandel des Berufsbildes immer knapper. Bei der Erschließung elektronischer Ressourcen, beim Austausch von Fremddaten und auch bei der Synchronisation der Sacherschließungsinformationen zwischen mehreren Manifestationen eines Werkes seien viele Probleme noch nicht gelöst. Zudem stehe einerseits Inhaltserschließung unter dem Verdacht, zu teuer und überflüssig zu sein, andererseits gebe es auch keine gesicherten Erkenntnisse zu Aufwand und Nutzen von Inhaltserschließung.

Am Test nahmen zwölf Personen aus dem Referat Sacherschließung und den Fachreferaten, die dem Hauptbestand und verschiedenen Sonderabteilungen zugeordnet sind, teil. In einer Auftaktsitzung wurde gefragt, wie momentan gearbeitet werde und wie die Erschließenden gerne arbeiten würden. Derzeit wird in unterschiedlichen Systemen nach Fremddaten recherchiert, es wird händisch aus anderen Erschließungssystemen „übersetzt“ und kopiert, als Hilfsmittel werden Wortlisten, Skripte und Tabellen angelegt. Neben einer benutzerfreundlicheren Oberfläche werden automatisierte Verfahren gewünscht, die die genannten Routinen übernehmen. Im Test wurden die herkömmlichen Arbeitsschritte und die im DA-2 benötigten Arbeitsschritte gezählt und miteinander verglichen. Beim Einsatz des DA-2 wurden deutlich weniger Recherchen in anderen Systemen getätigt, als sie sonst nötig sind. Auch die Zahl der Eingaben bei der eigentlichen Verschlagwortung war beim Einsatz des DA-2 sowohl bei Fremddatennutzung als auch bei Schlagwortsuche und eigener Verschlagwortung deutlich reduziert und die für die Verschlagwortung benötigte Zeit daher sehr viel kürzer. Da weit über die Hälfte der Titel im Test noch nicht verschlagwortet waren, wurde der DA-2 insbesondere

bei moderner englischsprachiger Forschungsliteratur und bei E-Books von den Testpersonen als deutliche Entlastung gesehen. Aufgrund der Einbindung von Rameau im DA-2 verspreche man sich auch im Bereich französischer Forschungsliteratur eine deutliche Entlastung. Der Import der Daten in den GBV-Katalog wird zurzeit von der Verbundzentrale des GBV (VZG) programmiert und muss noch geprüft werden; auch Performanz, Datenlage und Ähnlichkeitssuche werden erst im Echtzeitbetrieb vollständig messbar werden.

Die Mehrheit derer, die am Test teilgenommen haben, möchte gerne mit dem DA-2 arbeiten. Es gebe aber auch noch weitere Wünsche: So möchte man z.B. die Basisklassifikation, andere Klassifikationen und hausinterne Codes mit dem gleichen Werkzeug vergeben, es sollten mehr Schlagwortfolgen (bisher nur aus dem SWB) angezeigt, mehr Quellkataloge eingebunden und weitere Schlagwortübersetzungen angeboten werden. Man möchte einen direkten Zugang von der Anzeige eines Schlagwortnormsatzes zu den damit verknüpften Titeldaten haben und am liebsten den DA-2 in die Katalogisierungsoberfläche WinIBW eingebunden sehen. Eine Kooperation mit dem Konkordanzprojekt Coli-conc der VZG[10] wäre ebenfalls wünschenswert. Da der DA-2 den Erschließenden aber auch im derzeitigen Zustand schon viel Handarbeit abnehme, schaffe er Kapazität für eine qualitätvolle intellektuelle Erschließung. Der Test habe überdies – wie Regine Beckmann berichtete – dazu angeregt, mit einer grundsätzlichen Diskussion über die Inhaltserschließung an der SBB-PK zu beginnen.

Den Vortragsblock zum Digitalen Assistenten beschloss Peter Schäuble, Inhaber von Eurospider, mit einer Betrachtung, welche Erkenntnisse aus den Erfahrungen mit dem DA-2 und aus der Masterarbeit von Ursula Jud-Reichlen gewonnen werden könnten. Schon Ende des 19. Jahrhunderts gab es die Ansicht, Sacherschließung sei unnötig. Hermann Escher (1857-1938), der erste Direktor der ZB Zürich, setzte jedoch damals bereits auf den Schlagwortkatalog, um gerade den nicht-wissenschaftlichen Benutzerkreisen von Bibliotheken den Zugang zu deren Buchbestand zu ebnen. Heute sind Sacherschließungsdaten überall in den Verbünden zu finden und eignen sich sowohl für die Nachnutzung bei der intellektuellen Erschließung in den Bibliotheken als auch für das Trainieren und Evaluieren beim Einsatz automatischer Erschließungsverfahren, die die Informationswissenschaft bereitstellt. Interessant sei, dass mehr als 50 % der erschlossenen Titel nicht mehr als zwei Sachschlagwörter aufwiesen, obwohl viel mehr vergeben werden könnten. Aber sowohl in Hinsicht auf den Erschließungsaufwand als auch in Hinsicht auf die Auffindbarkeit sei eine Anhäufung von Sachschlagwörtern wenig zielführend.

Aufgefallen war dem Referenten, dass Bibliothekswissenschaft und Informationswissenschaft unterschiedliche Kriterien anwenden, wenn sie Erschließungsqualität beschreiben. In der Bibliothekswissenschaft schaut man auf die Indexierungstiefe und die Indexierungsbreite, während die Informationswissenschaft den Fokus auf die Auffindbarkeit legt und daher Ausbeute und Präzision misst, dafür aber auch stets standardisierte Testkollektionen benötigt. Bibliothekswissenschaft und Informationswissenschaft stehen insofern in einem Gegensatz zueinander. Einem Wortschatz wie der GND haftet informationstheoretisch immer ein „curse of dimensionality“ an, da ein Begriff meist

10 Zu Coli-conc vgl. Uma Balakrishnan, „DFG-Projekt Coli-conc: Das Mapping-Tool Cocoda,“ *o-bib* 3, Nr. 1 (2016): 11–16, https://doi.org/10.5282/o-bib/2016H1S11-16.

mehrere Bedeutungsdimensionen hat. Als Beispiel führte der Referent die Begriffe „Bundesrat“ und „Bundesrätin“ an, die die gleiche Eigenschaft aufweisen, wenn man auf die Funktion schaut. Jedoch tritt jeweils eine zweite, auf das Geschlecht bezogene Eigenschaft hinzu, sodass sich beide Begriffe informationstheoretisch nicht auf einen Vektor abbilden lassen. Zieht man jedoch beim Vektor „Bundesrat“ die Eigenschaft „Mann“ ab und addiert die Eigenschaft „Frau“, so gelangt man immerhin in die Nähe des Vektors „Bundesrätin“. Insofern kann die Informationswissenschaft dem Problem des „curse of dimensionality“ durch die Modellierung mit Kontextbegriffen beikommen.

Dennoch sei, wie Schäuble erläuterte, der Weg zu einer computerunterstützten Sacherschließung noch lang. Auf diesem Weg identifizierte er sechs Stationen bzw. „Generationen“: Die erste Generation der Sacherschließung verlief ausschließlich intellektuell und ohne Computer. In der zweiten Generation wurde Pionierarbeit geleistet: Es wurden einfache Merkmale für einen automatischen Vergleich entwickelt. Die dritte Generation war von Systemen mit einfachen Vergleichsmethoden gekennzeichnet. Heute befinde man sich nach verschiedenen Weiterentwicklungen und dem Aufkommen einer Wettbewerbssituation zwar schon in der vierten Generation, jedoch stünden der sinnvolle Einsatz maschinellen Lernens (fünfte Generation) und die automatische Generierung von Merkmalen sowie die Verarbeitung der daraus folgenden Datenfülle mit einer entsprechenden Rechenleistung (sechste Generation) noch aus. Ein Austausch zu Innovationen in Bibliotheken und anderen Gedächtnisinstitutionen sei sinnvoll und wichtig. Dafür soll der von Eurospider neu gegründete Newsletter „www.InnoBib.News“[11] einen Beitrag leisten. Zum Abschluss seines Vortrags versprach Schäuble für die nächte Weiterentwicklung, den DA-3, eine dreifache Tyrolienne (Seilrutsche) aus Fremddaten, approximativen Übersetzungen und Erschließung ähnlicher Titel, dazu ein Sicherheitsnetz aus „Erschließung ohne einsetzbare Vorschläge“ und, nicht zu vergessen, Spaß beim Fahren mit der dreifachen Tyrolienne.

Maschinelle Inhaltserschließung an der ZBW und der DNB

Abgerundet wurde der Workshop durch die Vorstellung der Aktivitäten zur automatischen Sacherschließung an der ZBW – Leibniz-Informationszentrum Wirtschaft und der Deutschen Nationalbibliothek (DNB).[12]

Martin Toepfer berichtete von den Forschungsaktivitäten bzw. laufenden Arbeiten der ZBW im Bereich der maschinellen Indexierung. Die ZBW entwickelt mit eigenen Kräften automatische Sacherschließungssysteme. Ziel ist ein multilingualer, themenbezogener Sucheinstieg im Rechercheportal EconBiz. Unabhängig vom Format und der wachsenden Zahl an Publikationen soll eine lückenlose, einheitliche Erschließung aller Publikationen erfolgen. Neben der Unabhängigkeit von kommerziellen Anbietern sieht die ZBW die Vorteile einer Eigenentwicklung im Auf- und Ausbau von hauseigenen Kompetenzen und der Nachnutzung des sowohl in der Bibliothek als auch bei Forschungspartnern vorhandenen Know-hows.

11 *InnoBib.News*, zuletzt geprüft am 16.08.2017, https://www.eurospider.com/de/innobib.

12 Ein weiterer geplanter Beitrag von Priska Bucher (ZB Zürich) zum Projekt „FREmdDaten-Anreicherung von Sacherschließungsdaten“ (FRED) musste leider entfallen.

Die ZBW versteht sich als forschende Einrichtung und ist infolgedessen auch auf dem Gebiet neuer Indexierungsverfahren und Verfahren zur Thesaurusanreicherung tätig. Für automatische Verfahren werden verschiedene Ansätze kombiniert: Der Crosskonkordanz-Ansatz (Autoren-Keywords – STW-Begriffe[13]), ein dokumentorientierter Ansatz (mit Kontextauswertung) und ein assoziativer selbstlernender Ansatz. Dabei war eine Erkenntnis, dass einfache Ansätze mit weniger Parametern oftmals bessere Ergebnisse liefern als komplexe Verfahren und die Kombination verschiedener Verfahren.[14] Toepfer kündigte weitere Ergebnisse für die „Joint Conference on Digital Libraries (JCDL)" 2017 an.[15]

Sehr viel umfassender und weitergehender sind die Aktivitäten der DNB, über die Elisabeth Mödden berichtete. Bereits seit 2012 werden automatisiert DDC-Sachgruppen für Netzpublikationen und Zeitschriftenartikel vergeben – mittlerweile wurden auf diesem Weg über 900.000 Dokumente klassifiziert. 2015 startete die automatische Vergabe von medizinischen DDC-Kurznotationen für den gleichen Sammlungsausschnitt (mittlerweile über 130.000 erschlossene Dokumente). Der Hauptteil des Vortrags befasste sich mit der seit 2014 praktizierten automatischen Vergabe von GND-Schlagwörtern mithilfe einer Software der Firma Averbis. Das Verfahren wird derzeit in mehreren Bereichen eingesetzt: einerseits bei online vorliegenden wissenschaftlichen Monografien (u. a. Hochschulschriften, Monografien von wissenschaftlichen Verlagen), Publikationen von Book-on-Demand-Verlagen und Aufsätzen (u. a. von Springer), andererseits bei gedruckten Monografien der Reihen B und H.

Abb. 2: Vortrag von Elisabeth Mödden. Foto: UB Stuttgart/Frank Wiatrowski

13 STW steht für den Standard-Thesaurus Wirtschaft, zuletzt geprüft am 19.08.2017, http://zbw.eu/stw/.

14 Vgl. Martin Toepfer und Andreas Oskar Kempf, „Automatische Indexierung auf Basis von Titeln und Autoren-Keywords: Ein Werkstattbericht," 027.7 *Zeitschrift für Bibliothekskultur* 4, Nr. 2 (2016): 84–97, https://doi.org/10.12685/027.7-4-2-156.

15 Vgl. Martin Toepfer and Christin Seifert, „Descriptor-Invariant Fusion Architectures for Automatic Subject Indexing: Analysis and Empirical Results on Short Texts", Vortrag auf der ACM/IEEE Joint Conference on Digital Libraries (JCDL) 2017 am 20.06.2017 in Toronto, ON (Canada) (IEEE, 2017), publiziert auf IEEE Xplore am 27.07.2017, https://doi.org/10.1109/JCDL.2017.7991557.

Bei den vollständig elektronisch vorliegenden Dokumenten basiert die Schlagwortzuordnung nicht auf dem Gesamttext, sondern nur auf den ersten 80.000 Zeichen. Bei den Print-Publikationen werden die eingescannten Inhaltsverzeichnisse ausgewertet. Die bibliografischen Daten und die elektronischen Textausschnitte werden zunächst computerlinguistisch verarbeitet. Dabei werden beispielsweise Artikel und Verben erkannt (und ignoriert) und Wörter in Nominalphrasen auf Stammformen zurückgeführt sowie segmentiert. Aus „entzündliche Erkrankungen" wird so beispielsweise „entzuend" und „krank", aus „des Herzmuskels" wird „herz" und „muskel". Im nächsten Schritt erfolgt die Termidentifikation; dafür werden die ermittelten Zeichenketten mit aus der GND generierten Wörterbuchdateien abgeglichen. Ein besonderes Problem stellt die Disambiguierung dar. Beispielsweise matcht „kurs" nicht nur mit einem, sondern mit mehreren GND-Begriffen: mit „Kurs" (im Sinne von „Lehrgang"), „Kurs (Navigation)", „Kurs (Devisen)" (Verweisung zu „Wechselkurs"), „Kurs (Aktie)" (Verweisung zu „Aktienkurs") und „Kurs (Wertpapier)" (Verweisung zu „Wertpapierkurs"). Hier kommen verschiedene Methoden zum Einsatz, u. a. das Prüfen auf Synonyme zu einem der in Frage kommenden Begriffe im ausgewerteten Textausschnitt. Nicht alle ermittelten GND-Schlagwörter werden auch tatsächlich in das Katalogisierungssystem der DNB eingespielt, sondern jeweils nur diejenigen mit der höchsten Konfidenzrate (je nach Konfiguration maximal vier bzw. acht Schlagwörter).

Von besonderem Interesse ist die Evaluation des Verfahrens und die Qualität der auf diesem Weg ermittelten Schlagwörter. Wie Mödden erläuterte, werden regelmäßig Stichproben unterschiedlicher Art durchgeführt. Dazu gehört eine Einordnung der vergebenen Schlagwörter in die Kategorien „sehr nützlich", „nützlich", „wenig nützlich" oder „falsch". Aufgetretene Fehler werden analysiert und nach Fehlertypen charakterisiert; teilweise können durch Änderungen in der GND (z.B. Neuansetzung eines fehlenden Schlagworts) oder in den Wörterbuch-Routinen Verbesserungen erreicht werden. In der Präsentation wurde beispielhaft ein Screenshot eines aktuellen Analyseprotokolls gezeigt (Folie 21): Eine Online-Dissertation mit dem Titel „Die Ordnung der Dinge durch die Malerei. Jan van Kessels Münchner Erdteile-Zyklus" etwa bekam das Sachschlagwort „Kessel" zugeteilt und ein Werk mit dem Titel „Die Tennis Bibel" das Schlagwort „Bibel". Je nach Fachgebiet ergeben sich bessere oder schlechtere Ergebnisse: So funktioniere – so die Referentin – die maschinelle Schlagwortvergabe in der Medizin besonders gut und in der Informatik besonders schlecht.

Im Ausblick berichtete die Referentin, dass die Verfahren künftig auch auf andere Objekttypen ausgeweitet sowie für englischsprachige Dokumente eingesetzt werden sollen (über eine Konkordanz zwischen GND und LCSH). Außerdem werde an einem Vorschlagstool zur Pflege der GND gearbeitet; das System soll aktiv Hinweise auf möglicherweise in der GND fehlende Konzepte geben. Künftig sollen außerdem die maschinell generierten Schlagwörter auch in den Datendiensten der DNB ausgeliefert werden. Wie sehr die Inhaltserschließung der DNB dadurch ihren Charakter verändern wird, war zum Zeitpunkt des Workshops freilich noch nicht abzusehen.

Exkurs: Das neue DNB-Konzept zur Inhaltserschließung

Denn erst kurz nach dem Workshop wurde das neue Konzept der DNB für die Inhaltserschließung veröffentlicht.[16] Darin wird als Ziel formuliert, nur noch dann etwas intellektuell inhaltlich zu erschließen, „wenn maschinelle Verfahren entweder nicht zur Verfügung stehen, keine ausreichenden Ergebnisse liefern oder intellektuell erstellte Daten für die Weiterentwicklung der maschinellen Verfahren benötigt werden." Begründet wird die neue Strategie mit dem Anwachsen der zu bearbeitenden digitalen Publikationen, dem Auseinanderdriften zwischen der Erschließung von digitalen und Print-Materialien sowie einer neuen Sicht auf Erschließung „als zyklisches Verfahren (...), bei dem Erschließungsdaten immer wieder verändert und aktualisiert werden" können.

Bereits zum 1. September 2017 werden die Reihen B und H umgestellt. Deutschsprachige Publikationen in diesen Reihen erhalten ab diesem Zeitpunkt zusätzlich zu den DDC-Sachgruppen maschinell ermittelte GND-Schlagwörter; diese werden ab Mitte Oktober in den Datendiensten erscheinen. Änderungen gibt es auch in der klassifikatorischen Erschließung: Bisher wurden für die Reihen B und H vollständige DDC-Notationen vergeben; diese werden durch „DDC-Kurznotationen" ersetzt, „die derzeit von der DNB entwickelt werden". Zumindest für digitale Publikationen ist offenbar eine maschinelle Erzeugung dieser Kurznotationen geplant. Voraussichtlich 2018 soll die seit 2006 betriebene Erschließung mit vollständigen DDC-Notationen[17] dann grundsätzlich aufgegeben werden.

Es ist bedauerlich, dass diese Pläne auf dem Workshop noch nicht bekannt waren und folglich von den Teilnehmerinnen und Teilnehmern nicht diskutiert werden konnten. Ein zentraler Punkt ist natürlich die Qualität und Vollständigkeit der maschinell generierten Schlagwörter, die ab Herbst in die Kataloge zahlreicher deutscher Bibliotheken gelangen werden. Eine aktuelle Statistik über die Ergebnisse der Stichprobenprüfungen wurde leider in der Präsentation von Elisabeth Mödden nicht vorgelegt. Eine Studie über maschinell erschlossene digitale Publikationen von 2013 ergab für die Precision (also die Nützlichkeit der ermittelten Schlagwörter) Werte zwischen 0,38 in der Informatik und 0,62 in der Wirtschaft (wobei 1,0 für „sehr nützlich" und 0,0 für „falsch" steht); zumeist lagen die Werte zwischen 0,45 und 0,55.[18] In dieser Studie wurde das Ergebnis der maschinellen Indexierung auch auf Vollständigkeit hin geprüft – also daraufhin, ob alle für das Dokument relevanten Schlagwörter gefunden wurden (dies würde einen Recall von 1,0 bedeuten). Hier ergaben sich für die meisten Sachgruppen Werte zwischen 0,65 und 0,75. Überträgt man die Ergebnisse auf ein Dokument mit

16 „Grundzüge und erste Schritte der künftigen inhaltlichen Erschließung von Publikationen in der Deutschen Nationalbibliothek," DNB, zuletzt geprüft am 14.08.2017, http://www.dnb.de/DE/Erwerbung/Inhaltserschliessung/grundzuegeInhaltserschliessungMai2017.html. Vgl. auch „Änderung der Inhaltserschließung in den Metadaten der Deutschen Nationalbibliografie ab 1. September 2017," DNB, zuletzt geprüft am 14.08.2017, http://www.dnb.de/DE/Erwerbung/Inhaltserschliessung/aenderungInhaltserschliessungSeptember2017.html. Mittlerweile wurde von der DNB noch ein weiteres Papier zu diesem Thema veröffentlicht: Ulrike Junger und Ute Schwens, *Die inhaltliche Erschließung des schriftlichen kulturellen Erbes auf dem Weg in die Zukunft*, zuletzt geprüft am 19.08.2017, http://www.dnb.de/SharedDocs/Downloads/DE/DNB/inhaltserschliessung/automatischeInhaltserschliessung.pdf.

17 Zur Einführung der DDC bei der DNB und den damit verbundenen Zielen vgl. Magda Heiner-Freiling, „RSWK und DDC: Sacherschließung auf zwei Beinen," *Dialog mit Bibliotheken* 17, Nr. 3 (2005): 4–13, zuletzt geprüft am 14.08.2017, http://www.ddc-deutsch.de/Subsites/ddcdeutsch/SharedDocs/Downloads/DE/publikationen/heinerFreiling2005rswkUndDDC.pdf.

18 Vgl. Sandro Uhlmann, „Automatische Beschlagwortung von deutschsprachigen Netzpublikationen mit dem Vokabular der Gemeinsamen Normdatei (GND)," *Dialog mit Bibliotheken* 24, Nr. 2 (2013): 26–36, zuletzt geprüft am 14.08.2017, http://d-nb.info/1048376788/34, hier 32f.

vier maschinell generierten Schlagwörtern, so bedeutet dies, dass nur zwei davon als sehr nützlich oder nützlich einzuordnen wären und ein bis zwei wichtige Schlagwörter fehlen würden.

Es wäre interessant gewesen zu erfahren, ob es seitdem zu nennenswerten Verbesserungen bei der Qualität gekommen ist und wie es sich auswirkt, wenn der Software nur ein eingescanntes Inhaltsverzeichnis zur Verfügung steht. Dies könnte, wie Stichproben zeigen, zu erheblich schlechteren Ergebnisse führen.[19] Mittlerweile hat sich auch Klaus Ceynowa, der Generaldirektor der Bayerischen Staatsbibliothek, in einem Beitrag in der Frankfurter Allgemeinen Zeitung zu dem neuen Konzept der DNB zu Wort gemeldet und damit eine intensive Diskussion ausgelöst.[20]

Aber auch die Aufgabe der DDC-Tiefenerschließung wirft erhebliche Fragen auf. Denn bisher ergab sich in den Katalogen wissenschaftlicher Bibliotheken eine relativ hohe Abdeckung mit DDC-Erschließung: Bei internationaler Literatur (insbesondere aus dem englischen Sprachraum) werden vollständige DDC-Notationen mit den Fremddaten geliefert, für die deutschsprachige Literatur wurden sie von der DNB erstellt. Das Zusammenspiel ermöglicht nicht nur eine sprachübergreifende Recherche, sondern bietet auch große Chancen z.B. für Konkordanzen und Linked-Data-Projekte. Diese Einheitlichkeit wird nun – so ist zu befürchten – wieder verloren gehen.

Vor dem Hintergrund der neuen Strategie der DNB gewinnt der Stuttgarter Workshop zur computerunterstützten Inhaltserschließung an Bedeutung über den Tag hinaus. Denn der dort schwerpunktmäßig thematisierte Digitale Assistent steht sozusagen beispielhaft für ein Gegenmodell zur Positionierung der DNB. Anstatt vollständig auf maschinelle Methoden zu setzen, steht das Alternativmodell auf zwei Säulen: Es kombiniert eine umfassende Nachnutzung vorhandener Erschließungsdaten – also die klassische bibliothekarische Kooperation – mit maschinell erzeugten Vorschlägen. Die Erschließung bleibt dabei in der Hand eines menschlichen Indexierers bzw. einer menschlichen Indexiererin. Man darf gespannt sein, welche Rolle diese beiden Modelle in den nächsten Jahren bei der Weiterentwicklung der Erschließungsverfahren spielen werden.

Zitierfähiger Link (DOI): https://doi.org/10.5282/o-bib/2017H3S94-105

19 Vgl. Heidrun Wiesenmüller, „Das neue Sacherschließungskonzept der DNB in der FAZ," *Basiswissen RDA* (Blog), 02.08.2017, zuletzt geprüft am 14.08.2017, https://www.basiswissen-rda.de/neues-sacherschliessungskonzept-faz/, dort Kommentare #7 und #24.

20 Klaus Ceynowa, „In Frankfurt lesen jetzt zuerst Maschinen," *Frankfurter Allgemeine Zeitung* vom 31.07.2017, zuletzt geprüft am 14.08.2017, http://www.faz.net/-gqz-909kq. Vgl. auch Wiesenmüller, „Das neue Sacherschließungskonzept der DNB in der FAZ".