Distant Viewing-Forschung mit digitalisierten Kinderbüchern: Voraussetzungen, Herausforderungen und Ansätze

Wiebke Helm, Thomas Mandl, Sigrun Putjenter, Sebastian Schmideler und David Zellhöfer

Bilderkennung durch KI ist derzeit in aller Munde. Ob im Einsatz großer Player wie Google & Co. oder der medizinischen Diagnostik – die Auswertung visuellen Materials spielt eine immer größere Rolle, auch im Bereich der Digital Humanities-Forschung. Nach jahrelanger Fokussierung auf den Text hat diese nun ihr Untersuchungsfeld auf das Bild erweitert. Doch wie sehr dieses Medium beim Digitalisieren von bibliothekarischen Altbeständen bisher vernachlässigt wurde, zeigt der folgende Anwendungsfall, der Desiderate bei der Retrodigitalisierung wie auch Grenzen bestehender Bibliothekskategorien und technischer Systeme in Bezug auf die nachträgliche Inklusion von fachwissenschaftlichen Metadaten und Machine Learning-Prozessen identifiziert. Aus diesen Erkenntnissen werden Aufgaben und Lösungsmöglichkeiten für die zukünftige Forschung abgeleitet.

Image recognition by AI is currently on everyone's lips. The evaluation of visual material plays an increasingly significant role - in everyone's use of large players such as Google & Co. or medical diagnostics as well as in the field of digital humanities research. After years of focusing on texts, it has now extended its studies to the image. To what degree this medium has been neglected in the digitisation of library collections is shown by the following application case. Desiderata in retro-digitisation as well as limits of existing library categories and technical systems will be identified with regard to the inclusion of scientific metadata and machine learning processes at a later point in time. From these findings tasks and options for future research are being derived.

Distant Viewing als Chance für Bibliothek und Forschung

Bibliotheken und Archive haben in den vergangenen Jahren umfangreiche Digitalisierungen ihrer Altbestände durchgeführt und auf diese Weise einen Beitrag zur Sicherung des kulturellen Erbes geleistet. Neben der Bestandserhaltung aus konservatorischen Gesichtspunkten sollte es im Interesse der Beteiligten sein, die digitalisierten Kulturgüter einer breiten Nutzung zuzuführen. Ein Umstand, den auch das neue europäische Urheberrecht für die Anwendbarkeit von Text- und Data-Mining (TDM) insbesondere für Forschungseinrichtungen betont und damit wissenschaftliche Bibliotheken mit Anforderungen konfrontiert, die über das bisher übliche Maß hinausgehen. Denn neben der Online-Bereitstellung der Bestände sollte das Datenmaterial einen Mehrwert für künftige Forschungen bereithalten und die sich wandelnden Forschungsgewohnheiten der Benutzer*innen einbeziehen, da der Stellenwert der Digital Humanities (DH)-Forschung weiter zunehmen wird. Gemeinsam müssen Möglichkeiten erarbeitet werden, digitalisierte Daten für TDM-Verfahren aufzubereiten und niedrigschwellig zur Verfügung zu stellen.

Im vorliegenden Beitrag werden diese Anforderungen exemplarisch an einem Distant Viewing-Forschungsprojekt aus den Digitalen Geisteswissenschaften aufgezeigt, das Illustrationen in historischen Kinder- und Jugendsachbüchern einer automatischen Analyse unterzieht und hierbei u. a. auf die retrodigitalisierten Bestände der Staatsbibliothek zu Berlin (SBB) zurückgreift. Aus der Forschungspraxis berichtend sollen die besonderen Herausforderungen im Umgang mit dem vorhandenen (Bild-)Datenmaterial einerseits und der nachträglichen Integration sowohl von fachwissenschaftlich erhobenen als auch maschinell erzeugten Metadaten in bestehende bibliothekarische Systeme andererseits dargelegt und diskutiert werden. Dabei wird deutlich, dass der Umfang bislang digitalisierter Sammlungen historischer Kinder- und Jugendbücher derzeit noch weiten Raum für Verbesserungen bietet – auch im Hinblick auf zukünftige Nutzungen im Bereich der Digital Humanities.

Das Projekt: Historische Kinder- und Jugendsachbuchillustrationen aus der DH-Perspektive

Für das 19. Jahrhundert sind grundlegende Veränderungen in der Gestaltung von Sachbüchern für Kinder und Jugendliche beobachtbar. Dies ist nicht zuletzt auf modernere Reproduktionsverfahren wie beispielsweise die Lithographie zurückzuführen, die eine Herstellung von Bildern in weitaus größerer Auflagenzahl erlaubt. Dadurch kommt es zur massenhaften Verbreitung von Illustrationen, die das Wissen der Zeit in Bilder transformieren. Darüber hinaus sind Veränderungen in der Bildästhetik und ein an bestimmte Adressaten gebundenes Bildprogramm feststellbar. Diese Prozesse mit Methoden der DH zu untersuchen und sichtbar zu machen, ist Aufgabe des interdisziplinären

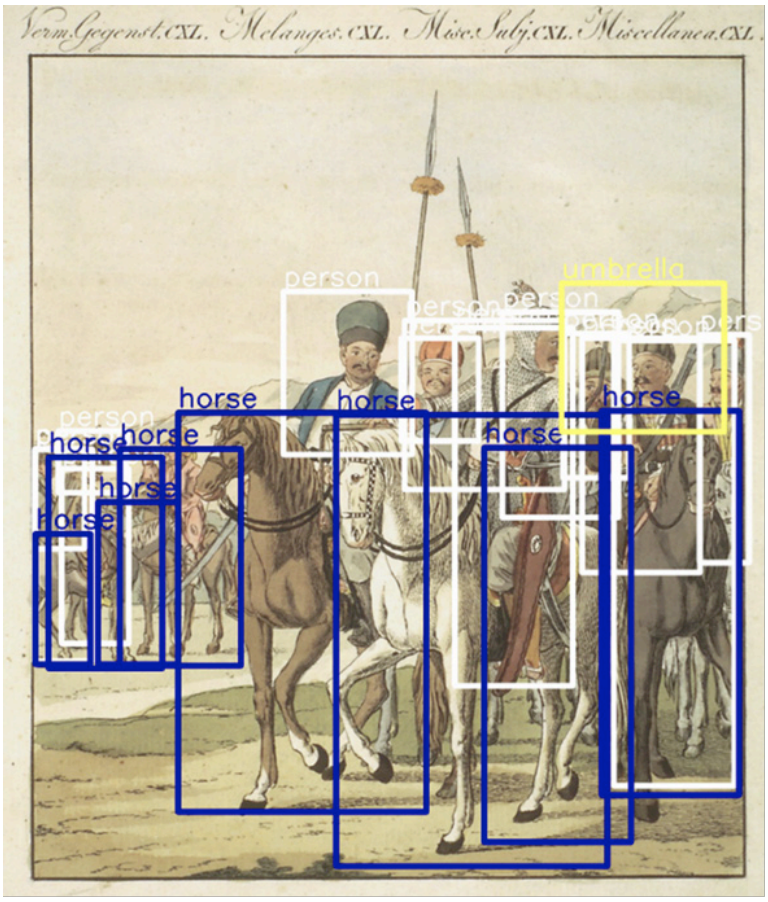

Abbildung 1: Korrekte und fehlerhafte Objektzuweisungen mit YOLO. Zugrunde liegt Bildtafel X des siebenten Bandes von Friedrich Justin Bertuchs „Bilderbuch für Kinder" in der Auflage von 1810. Quelle: BBF

Forschungsprojektes „Entwicklung der Bildikonographie in Wissen vermittelnder Kinder- und Jugendliteratur und Schullehrbüchern des 19. Jahrhunderts. Ein Distant Viewing Ansatz", das seit 2017 von der Fritz Thyssen Stiftung gefördert und in einer Kooperation von Kinder- und Jugendliteraturwissenschaftlern der Universität Leipzig und Informationswissenschaftlern der Universität Hildesheim durchgeführt wird.

Die Datenbasis für die automatische Bildverarbeitung bildet ein Korpus von 600 illustrierten Kinder-und Jugendsachbüchern, die zwischen 1800 und 1900 innerhalb des deutschen Sprachraums verlegt wurden und die bereits als Retrodigitalisate frei zur Verfügung stehen. Sie wurden aus den digitalen Beständen der SBB (*WegehauptDigital*) und der Universitätsbibliothek Braunschweig (*Sammlung Hobrecker*) anhand der Gattungssystematik des „Handbuchs zur Kinder- und Jugendliteratur"[1] zusammengestellt. Sie werden manuell mit forschungsrelevanten Metadaten angereichert, die weit über die Sacherschließungsdaten der Bibliothekskataloge hinausgehen. Deren Sicherung und Verfügbarkeit für andere Wissenschaftler*Innen soll prospektiv u. a. durch die Datenintegration in die Kataloge der SBB erfolgen, weshalb die Erfassung soweit möglich standardisiert stattfindet. Dabei zeigen sich jedoch verschiedene Herausforderungen, die als Desiderate besondere Chancen künftiger Forschungsvorhaben bilden, wie im Folgenden gezeigt werden soll. Hinzu kommt umfangreiches Trainingsmaterial, das u. a. aus der Bilddatenbank *Pictura Paedagogica Online*[2] der Bibliothek für Bildungsgeschichtliche Forschung Berlin generiert wird.

Derzeit wird das Bildmaterial der Volltexte von der Hildesheimer Projektgruppe hinsichtlich der angewandten Reproduktionstechnik und abgebildeten Objekte klassifiziert, es wird nach Ähnlichkeiten durchleuchtet, um Wiederverwendungen sichtbar zu machen, und es wird stilistisch bestimmt, um Aussagen zur Ikonographie von Sachillustrationen für Kinder und Jugendliche des 19. Jahrhunderts treffen zu können. Zum Einsatz kommen dabei Deep Learning-Ansätze wie SDD (Single Shot MultiBox Detector)[3] und YOLO (You Only Look Once)[4], die beispielsweise eine Erkennung von Objekten wie Personen und Tiere in Abbildungen durch einstufige Detektion erlauben. Diese Modelle sind jedoch auf realistische Fotografien trainiert und optimiert, weshalb die Erkennungsrate bei der Anwendung auf historische Buchillustrationen nicht zuletzt aufgrund des Domain-Shift sowie von Objekten in ungewöhnlicher Darstellungsperspektive, heute unüblich gewordenen Gegenständen oder von Anthropomorphisierungen zoologischer oder botanischer Objekte fehlerhaft ist und somit nachtrainiert werden muss (vgl. Abb. 1).

Distant Viewing

Im Bereich der DH-Forschung steht längst nicht mehr nur die ausschließliche Verarbeitung von Texten im Vordergrund. Doch noch immer ist die Zahl der Studien, die eine massenhafte Analyse von Abbildungen vornehmen, vergleichsweise überschaubar. Distant-Viewing-Projekte widmen sich u. a. der Bearbeitung kulturwissenschaftlicher Fragestellungen und nut-

1 Brunken, Otto/ Hurrelmann, Bettina/ Pech, Klaus-Ulrich (Hrsg.): Handbuch zur Kinder- und Jugendliteratur. Von 1800 bis 1850, Stuttgart; Weimar 1998, Sp. 111f.

2 http://opac.bbf.dipf.de/virtuellesbildarchiv/index.html [31.1.2019].

3 Liu, Wei/ Anguelov, Dragomir/ Erhan, Dumitru/ Szegedy, Christian/ Reed, Scott/ Fu, Cheng-Yang/ Berg, Alexander C.: SSD: Single shot multibox detector, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, S. 21-37; https://arxiv.org/abs/1512.02325 [4.2.2019]

4 Redmon, Joseph/ Divvala, Santosh/ Girshick, Ross/ Farhadi, Ali: You only look once: Unified, real-time object detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, S. 779-788; https://arxiv.org/abs/1506.02640 [31.1.2019].

zen dafür Ansätze aus der automatischen Bildverarbeitung.[5] Diese kann für das letzte Jahrzehnt auf erhebliche Fortschritte zurückblicken. So ist es beispielsweise durch Deep Learning-Verfahren[6] gelungen, positive oder negative Stimmungen in einem Bild zu klassifizieren.[7] Durch die automatische Analyse können aber auch Traditionslinien verfolgt, stilistische Ähnlichkeiten ausgewertet und Muster der Wiederverwendung von Abbildungen gefunden werden, wie ein anderes Untersuchungsbeispiel zeigt.[8] Schließlich können mithilfe von automatisierten Verfahren sehr viel mehr Abbildungen und Bildtafeln untersucht und verglichen werden, als es dem menschlichen Betrachter allein möglich ist.

Zur Digitalisierung und bibliothekarischen Erschließung historischer Kinder- und Jugendbücher

Innerhalb des deutschsprachigen Raums existieren zum Teil umfangreiche Kollektionen historischer Kinder- und Jugendbücher. Doch wurden diese bisher lediglich in geringem Umfang, vornehmlich aus Gründen der Bestandserhaltung, digitalisiert. Für Fachwissenschaftler*innen und interessierte Nutzer*innen bleibt der Zugang zu diesen oftmals fragilen und teils unikalen Materialien dadurch immer noch schwierig, eine umfängliche Erforschung (z. B. Big Data-Analyse) mit Methoden der DH ist damit so gut wie ausgeschlossen.

In einem frühen von der Deutschen Forschungsgemeinschaft (DFG) geförderten Projekt der Universität Oldenburg wurden 392 farbig illustrierte Kinderbücher speziell des 19. Jahrhunderts aus den Beständen der Universitätsbibliotheken Oldenburg und Braunschweig digitalisiert. Mittlerweile stellt die Technische Universität Braunschweig ca. 850 Titel[9] ihrer Kinderbuchsammlung, und damit einen Bruchteil ihres eigentlichen Bestandes an Kinder- und Jugendbüchern[10], digital zur Verfügung. Unter diesen Digitalisaten befinden sich 652 retrodigitalisierte Bände der umfangreichen *Sammlung Hobrecker*[11], die hauptsächlich aus der zweiten Hälfte des 19. sowie aus dem frühen 20. Jahrhundert stammen. Im Gegensatz dazu kann die SBB mittlerweile 2.764 digitalisierte Werke der Kinder- und Jugendliteratur anbieten, darunter 1.620 Bände, die im Rahmen des BKM Förderprojektes *WegehauptDigital*[12] in die Digitalen Sammlungen der Staatsbibliothek[13] eingespeist wurden.

Schließlich sei an dieser Stelle noch beispielhaft auf den *Schatzbehalter* der Arbeitsstelle für Kinder- und Jugendmedienforschung der Universität zu Köln (ALEKI)[14] hingewiesen, der sich inzwischen vom reinen Online-Angebot mit einer überschaubaren Anzahl digitalisierter Bände aus dem eigenen Bestand zu einer Plattform für die Metasuche nach einschlägiger digitaler Literatur entwickelt hat. 4.630 Titel lassen sich derzeit über die Oberfläche des *Schatzbehalters* recherchieren, wobei die Standorte der physischen Quellen europaweit verteilt liegen. Obwohl diese Menge auf den ersten Blick beeindruckend wirkt, muss der Vollständigkeit halber angemerkt werden, dass sie zu einem geringfügigen Teil auch Werke der Sekundärliteratur enthält. Vor allem aber wird im direkten Vergleich mit der Fachbibliographie „Alte deutsche Kinderbücher“ von Heinz Wegehaupt deutlich, wie sehr die Digitalisierung in diesem Bereich noch immer im Argen liegt. Während Wegehaupt in den vier Bänden seines maßgeblichen Nachschlagewerks[15]

5 Beispielhaft sei hier nur auf die filmwissenschaftlichen Studien von Taylor Arnold und Lauren Tilton verwiesen. https://distantviewing.org/ [30. Januar 2019].

6 Vgl. hierzu Geißler, Stefan: „Deep Learning lernen“, in: b.i.t.online 20 (2017), S. 383-390.

7 Campos, Víctor/ Jou, Brendan/ Giró-i-Nieto, Xavier: From Pixels to Sentiment: Fine-tuning CNNs for Visual Sentiment Prediction. 2007 http://arXiv:1604.03489v2 [10. Januar 2019].

8 Saleh, Babak/ Elgammal, Ahmed: Large-scale Classification of Fine-Art Paintings: Learning The Right Metric on The Right Feature. 5. Mai 2015. http://arxiv.org/abs/1505.00855 [30. Januar 2019].

9 https://publikationsserver.tu-braunschweig.de/servlets/solr/select?q=%2BallMeta%3AKinderbuchsammlung+%2BallMeta%3Ader+%2BallMeta%3AUniversit%C3%A4tsbibliothek+%2BallMeta%3ABraunschweig+%2Bstate%3A%22published%22+%2Bcategory.top%3A%22mir_genres%5C%3Abook%22+%2BobjectType%3A%22mods%22&fl=*&sort=mods.dateIssued+desc&rows=20&version=4.5&mask=content%2Fsearch%2Fcomplex.xed [30. Januar 2019].

10 Dieser Bestand umfasst für den Publikationszeitraum 1565 bis heute mehr als 28.000 Bände.

11 https://publikationsserver.tu-braunschweig.de/servlets/solr/select?q=%2BallMeta%3Asammlung+%2BallMeta%3Ahobrecker+%2Bstate%3A%22published%22+%2Bcategory.top%3A%22mir_genres%5C%3Abook%22+%2BobjectType%3A%22mods%22&fl=*&sort=mods.dateIssued+desc&rows=20&version=4.5&mask=content%2Fsearch%2Fcomplex.xed [8. Januar 2019]. Zu ergänzen ist, dass der von Peter Düsterdieck verantwortete Bestandskatalog mehr als 8.600 Bände der ehemaligen Sammlung Hobrecker für die Braunschweiger Universitätsbibliothek verzeichnet. Düsterdieck, Peter: Die Sammlung Hobrecker der Universitätsbibliothek Braunschweig. 2 Bde., München u. a. 1985.

12 https://staatsbibliothek-berlin.de/die-staatsbibliothek/abteilungen/kinder-und-jugendbuecher/recherche-und-ressourcen/wegehauptdigital/ [5. Januar 2019].

13 https://digital.staatsbibliothek-berlin.de/ [5. Januar 2019].

14 http://schatzbehalter.aleki.uni-koeln.de/portal/home.html?l=de [5. Januar 2019].

15 Wegehaupt, Heinz: Alte deutsche Kinderbücher. 1507-1850, Berlin 1979.
Wegehaupt, Heinz: Alte deutsche Kinderbücher. 1851-1900, Berlin 1985.
Wegehaupt, Heinz: Alte deutsche Kinderbücher. 1524-1900, Stuttgart 2000.
Wegehaupt, Heinz: Alte deutsche Kinderbücher. 1521-1900, Stuttgart 2003.

für den Erscheinungszeitraum 1507 bis 1900 insgesamt über 13.150 Titel deutscher Kinder- und Jugendliteratur nachwies, verzeichnet der *Schatzbehalter* für denselben Zeitraum nur 3.175 Titel, davon ca. 3.115, die der deutschen Kinder- und Jugendliteratur zuzurechnen sind. Demnach sind gut 10.000 Titel, d. h. über Dreiviertel der von Heinz Wegehaupt nachgewiesenen Gesamtmenge, noch nicht in digitalisierter Form für die Forschenden verfügbar gemacht worden. Hinzu tritt das Problem der Repräsentativität des retrodigitalisierten Materials für Gattungen und Genres typischer Entwicklungen der historischen Kinder- und Jugendliteratur. Die Bestände erweisen sich im Unterschied zu *WegehauptDigital* als besonders heterogen. Sie verfolgen nicht den Anspruch der homogenen Bildung eines repräsentativen Korpus der kinder- und jugendliterarischen Produktion. Somit ist der fachwissenschaftliche Gebrauchswert für historische, systematische und theoretische Perspektiven der kulturwissenschaftlichen historischen Kinder- und Jugendbuchforschung der retrodigitalisierten Kollektion durchaus optimierungsbedürftig.

Bei näherer Betrachtung der Rechercheergebnisse im *Schatzbehalter* zeigt sich auch die Crux der Digitalisate unterschiedlicher Herkunft. Sowohl die Qualität der Scans[16] als auch die Qualität und der Umfang der verfügbaren Metadaten sind äußerst ungleich, obwohl zumindest für Deutschland Standards für die Digitalisierung und Erschließung existieren. Beispielhaft zu nennen wären hier die DFG-Praxisregeln „Digitalisierung“[17]. Doch selbst innerhalb einer Bibliothek können Qualität und Quantität der Metadaten variieren. So erfasst die SBB für Titel, die im Verzeichnis Deutscher Drucke des 18. Jahrhunderts (VD 18) nachgewiesen werden, bibliographische Zitate entsprechend der Regeln der Arbeitsgemeinschaft Alte Drucke beim GBV und SWB[18]. Im Rahmen der zeitlich stark befristeten Durchführung des Digitalisierungsprojekts *WegehauptDigital* war eine solche zusätzliche Erschließung nicht möglich. Hierbei wird ein Kernproblem der Digitalisierung deutlich: die Fokussierung auf die inhaltstragende Medieneinheit mit dem Blick auf die Bereitstellung – ergo der bibliothekarischen Kernaufgabe. Dieser gegenüber steht der Bereich der DH, der neue Fragestellungen z. B. mittels Methoden der Computerlinguistik zu beantworten sucht bzw. durch diese neue Informationen generiert, die im Sinne der Tiefenerschließung bzw. Kontextualisierung auch für Bibliotheken sowie deren Nutzer*innen von Interesse sind. Aus Sicht der DH ist demnach die Digitalisierung nur ein erster Schritt der Zugänglichmachung. Wahrer Mehrwert entsteht jedoch erst, wenn zusätzliche standardisierte, maschinenlesbare oder semantisch annotierte Informationen zu den Daten, wie Volltexte oder Named Entities, bereitgehalten werden. Für das erwähnte DH-Projekt hieße das: Automatisch durch Objekterkennungssysteme erzeugte Metadaten liefern passgenaue Treffer bei der Bildsuche im Bestand retrodigitalisierter Kinder- und Jugendbücher. Für die Klassifikation und die Ähnlichkeitssuche von Buchillustrationen ist ferner zu überlegen, inwieweit Repräsentationen aus Standard-CNN-Modellen[19] für Abbildungen zur Nachnutzung angeboten werden können, um die Pixel-Ebene zu umgehen.

WegehauptDigital

Im Dezember 2012 beschloss das Referat des Beauftragten der Bundesregierung für Kultur und Medien (BKM), das Projekt „WegehauptDigital: Digitalisierung deutscher Kinderbücher 1507-1900“ im Rahmen der sogenannten Digitalisierungsoffensive zu fördern. Dadurch war die Bearbeitung und Digitalisierung historischer Kinder- und Jugendliteratur in größerem Umfang als bisher möglich. Dafür musste jedoch eine Selektion des gut 27.600 Titel umfassenden Monographien-Altbestandes (1507–1913) vorgenommen werden. Diese erfolgte nach dem Kriterium der Gattungszugehörigkeit. So konnte zumindest ein Großteil der zwischen 1679 und 1913 verlegten Sachbücher für Kinder und Jugendliche von einem externen Dienstleister digitalisiert und anschließend bibliothekarisch bearbeitet werden. Im Fokus stand der traditionelle, bibliothekarische Anwendungsfall, die Medieneinheiten zu erschließen, um diese besser auffindbar und damit nutzbar zu machen. Deshalb wurden sämtliche Titel nach der Dewey Dezimal Klassifikation (DDC) erfasst, um aufgrund der globalen Verbreitung des Klassifikationssystems einen sprachunabhängigen, sachlichen Zugang zum Konvolut der digitalisierten Kinder- und Jugendsachbücher zu ermöglichen. Individuell wurde dann außerdem festgelegt, welche Bestandteile eines Bandes das zugehörige Digitalisat ausweisen soll. Neben Inhaltsverzeichnis und Register wurden weitere Strukturdaten wie u. a. Frontispize, separate Bildtafeln und Kartenbeilagen aufgeführt. Diese können nun über die

16 Häufig liegen die Digitalisate schon in der Web-Ansicht nicht in maximaler Auflösung vor. Eine JPG-Komprimierung von 80% wäre jedoch für Forschungszwecke ideal. Darüber hinaus sind eine konforme Auflösung sowie einheitliche Formate wichtig.

17 http://www.dfg.de/formulare/12_151/12_151_de.pdf [5. Januar 2019].

18 https://verbundwiki.gbv.de/display/GAD/Arbeitsgemeinschaft+Alte+Drucke+beim+GBV+und+SWB [5. Januar 2019].

19 Convolutional Neuronal Networks (CNN) sind eine oft erfolgreich eingesetzte Architektur aus dem Bereich der Deep Learning-Verfahren.

Navigationsleiste in den Digitalisierten Sammlungen der SBB angesteuert und aufgerufen werden. Mit der Einstellung vorsegmentierter Erscheinungszeiträume und der Sortierung nach Sachthemen entsprechend der DDC-Sachgruppen wurde eine nutzerfreundliche Rechercheoberfläche auf den Webseiten der Kinder- und Jugendbuchabteilung eingerichtet, die eine spezifische Suche nach Digitalisaten erlaubt.

Ergänzung der traditionellen, bibliographischen Erschließungsdaten

Im Rahmen des vorgestellten Projekts werden von fachwissenschaftlicher Seite Daten generiert, welche die bestehenden Formal- sowie auch die Sacherschließungsdaten systematisch ergänzen und dadurch neue forschungsgetriebene Anwendungsfälle ermöglichen, jedoch auch Bibliotheken vor neue Herausforderungen stellen. Sofern nicht bereits vorhanden, werden im genannten Projekt manuell Angaben zur Entstehungsmethode der Abbildungen (Drucktechnik), zu den an diesen beteiligten Personen (z. B. Zeichner, Stecher, Lithographen), ferner die Gattungszugehörigkeit sowie die intendierte Zielgruppe (Lesealter und Geschlecht) definiert. Darüber hinaus ist es ein Anliegen, die Digitalisate zu referenzieren und den fachwissenschaftlichen Nutzer*innen damit einen Recherchemehrwert anzubieten. Verwiesen werden soll auf einschlägige Fachbibliographien wie die schon erwähnte von Wegehaupt und das von Theodor Brüggemann begründete mehrbändige „Handbuch zur Kinder- und Jugendliteratur“[20]. Diese Daten in den Datensätzen der Formalerschließung adäquat unterzubringen, ist teilweise schwierig, ihre nachträgliche Inklusion in die bereits vorhandenen METS/MODS-Datensätze der Digitalisate nach dem aktuellen Stand der Technik nur schwer realisierbar, sofern man potenziellen späteren Datenverlust im komplexen Zusammenspiel der am Digitalisierungs-Workflow beteiligten Systeme ausschließen möchte (siehe unten). Vergleichbare Herausforderungen ergeben sich bei der maschinellen Anreicherung von Metadaten-Inhalten bzw. der Erschließung von Bildinhalten, deren prinzipielle Machbarkeit am Beispiel der Daten der SBB demonstriert werden konnte.[21]

Als weitgehend trivial kann die Implementierung der Informationen über die angewandte Drucktechnik der Illustrationen sowie bibliographischer Zitate angesehen werden. In beiden Fällen bietet das Kategorienschema des GBV eine entsprechende Pica3-Kategorie: 2277 für bibliographische Zitate im Bereich alter Drucke bzw. 4061 für die Illustrationsangabe oder sonstige physische und technische Angaben. Jedoch ist mit Blick auf die Kategorie 2277 anzumerken, dass die „Liste der fachlichen Nachschlagewerke der GND“ mit einer Ausnahme bisher keine Norm für die Zitation von Referenzwerken aus dem Bereich der Kinder- und Jugendliteraturforschung vorsieht. Diesbezügliche Abweichungen in der Datenbasis bestehender Sammlungen – sofern Quellenangaben überhaupt angeführt werden – sind damit vorprogrammiert.

Schwieriger wird es hinsichtlich der Unterbringung des künstlerischen Personals, das an der Entstehung des Werks beteiligt war. Sofern es sich bei dem Werk nicht um einen Einblattdruck handelt, gilt dieser Personenkreis lediglich als Beiträger und wird nach den Regeln der Resource Description and Access (RDA) in der den alten Drucken vorbehaltenen Kategorie 3050 mit der Beziehungskennzeichnung *BeiträgerIn k.* sowie einem entsprechenden MARC-Code ($4ctb) bedacht. Spezifischere Angaben über die Funktion dieser Personen müssten in einer Fußnote (4201) festgehalten werden. Da in diesem Bereich, außer dem Ablageort, wenig standardisiert ist, wird bereits an diesem Beispiel deutlich, dass eine maschinelle Auswertung, um Fragestellungen aus dem Bereich der DH zu beantworten, äußerst komplex würde. Eine vergleichbare negative Auswirkung auf den traditionellen, bibliothekarischen Anwendungsfall lässt sich hier jedoch nicht erkennen.

Wünschenswert im Rahmen einer automatisierten Umsetzung, aber aktuell schwierig, wäre in diesem Fall zudem das Einsetzen und die automatische Verknüpfung über die Identifikationsnummer des entsprechenden GND-Normdatensatzes, da hierzu robuste Named Entity Recognition (NER) zum Einsatz kommen müsste, wenn die Entität nicht eindeutig über einen Stringvergleich auflösbar ist. Durch diesen Anreicherungsschritt, in Kombination mit einem kontrollierten Vokabular oder einer entsprechenden Ontologie, um Rollen standardisiert zu beschreiben, entstünde ein wesentlicher Mehrwert außerhalb des bibliothekarischen Anwendungsfalls für Forscher*innen im Bereich der DH. Aktuell erforscht die SBB entsprechende Techniken im Rahmen des vom Bundesministerium für Bildung und Forschung geförderten QURATOR-Verbundprojekts[22].

20 Handbuch zur Kinder- und Jugendliteratur. Begr. v. Theodor Brüggemann, 6 Bde., Stuttgart, Weimar 1982-2008.

21 Siehe Zellhöfer, David: Exploring Large Digital Libraries by Multimodal Criteria, in: Norbert Fuhr/ László Kovács/ Thomas Risse/ Wolfgang Nejdl (Hrsg.): Research and Advanced Technology for Digital Libraries – 20th International Conference on Theory and Practice of Digital Libraries, Cham 2016, S. 307-319.

22 https://3pc.de/forschung/qurator/ [5. Januar 2019].

Auch hinsichtlich der Definition der Zielgruppe eines Werks stößt die Projektgruppe an die Grenzen des Pica3-Kategorienschemas. Zwar ist hierfür die Kategorie (1133) vorgesehen, jedoch erweist sich die Liste des möglichen Vokabulars als eng gesteckt und abgeschlossen: Jugend, Kind, Lehrer, Leseanfänger, Schüler, Sehbehinderter, Vorschulkind. Die Termini beziehen sich auf unsere zeitgenössische Literatur, historische Kategorien unterschiedlicher Entwicklungsstufen vom Kleinkind („für die lieben Kleinen") bis zum jungen Erwachsenen („für die reifere Jugend") sind ebenso wenig vorgesehen wie eine Differenzierung zwischen den Geschlechtern („für brave Mädchen" / „für wißbegierige Knaben"). Die Standardschlüssel der Hilfstafel 1 der DDC (-081-088 Verschiedene Personengruppen) könnten hier hilfreich sein.

Das nachträgliche, automatisierte Anhängen einer Schlüsselzahl ist allerdings problematisch, da es immer wieder Fälle gibt, in denen entweder keine oder keine zweite Schlüsselzahl angehängt werden kann oder es für die Synthese der Notation besondere Anweisungen gibt. Insofern haben wir uns für eine Eintragung der Schlüsselzahl in die Kategorie der für Nutzer*innen sichtbaren und über den OPAC recherchierbaren Fußnoten (4201) entschieden, wohl wissend, dass es sich dabei um eine Notlösung handelt, die eine konsistente Datenanalyse ausschließt, die aber zum jetzigen Zeitpunkt immerhin die Sicherung der gewonnenen Erkenntnisse in standardisierter Form ermöglicht.

Ingest von externen Metadaten in Workflowsysteme

Lässt man die Problematiken der bibliothekarischen Erschließung beiseite und widmet sich den praktischen Problemen der Anreicherung von DH-spezifischen oder fachwissenschaftlichen (Meta-)Daten in den typischen Digitalisierungsworkflowsystemen der Bibliotheken, wie Kitodo oder Goobi, wird eine weitere Herausforderung deutlich. Die nachträgliche Einspielung der oben beispielhaft genannten Daten in die entsprechenden METS/MODS-Datensätze der Digitalisate wäre zwar grundsätzlich denkbar, jedoch würden sämtliche Informationen gelöscht, sobald die XML-Daten des zugehörigen Kitodo-Satzes geändert und neu als METS/MODS-Datensatz exportiert werden, da das zentrale Workflowsystem keine Kenntnis von solchen Daten außerhalb des tradierten Digitalisierungsworkflows hat. In diesem Punkt müsste also entweder Kitodo in seiner Funktionsweise geändert oder aber der gesamte Vorgang jedes einzelnen, retrospektiv anzureichernden Datensatzes noch einmal zurückgenommen, bearbeitet und anschließend erneut hochgeladen werden. Offen bleibt auch der (technische) Ingest von Daten, die außerhalb der Bibliothek erzeugt wurden, da die gängigen Workflowsysteme hier noch wenig Unterstützung bieten. Bibliotheken riskieren folglich die Konsistenz der von ihnen bereitgestellten Daten, wenn die bestehenden Workflowsysteme mit Eigenentwicklungen ergänzt werden, welche die eigentlich wünschenswerten Daten aus dem Forschungsumfeld maschinell erweitern. Ein Desiderat ist demnach ein standardisierter Ingest von „Fremddaten", die nicht primär dem bibliothekarischen Basisanwendungsfall dienen. Denkbar wäre hier die Verknüpfung mittels Verweisen innerhalb von METS/MODS-Dateien wie im Fall von Volltexten im ALTO-Format (siehe unten). Hierbei muss jedoch kritisch geprüft werden, welche Standards für Metadaten unterschiedlicher Provenienz angelegt werden sollten, um wirklich einen Mehrwert im Bereich DH zu schaffen.

Auch im Hinblick auf die Informationen, die durch die automatisierte Bilderkennung erzielt werden, spielt dieses Problem eine erhebliche Rolle. Die auf diese Weise gewonnenen Schlagworte müssten nachträglich den bereits intellektuell ausdifferenzierten Strukturdaten hinzugefügt werden. Immerhin gäbe es dafür im Bereich der „Structural Map" einen korrekten Ablageort. Noch problematischer gestaltet sich die Unterbringung von Informationen zu Bildern, die sich innerhalb einer Textseite befinden. Diese sind nämlich nicht jeweils als Strukturdatum ausgezeichnet, es gibt in einem solchen Fall also keinen Anker im METS/MODS-Format, mit dem bibliographische Informationen verknüpft werden könnten. Ein möglicher Ansatzpunkt hierfür sind ALTO-Dateien, welche die SBB für Werke, die durch eine Optical Character Recognition (OCR) prozessiert wurden, anbietet. Kritisch aus Sicht der DH ist hierbei jedoch, dass standardisierte Pfade verlassen werden müssten (auch wenn die alldurchdringenden XML-Formate des Bibliothekssektors gerade dies gestatten) und das Handling verschiedener Metadatenformate benötigte Werkzeugketten unnötig komplex gestaltet. Zwar bietet der ALTO-Volltext den Illustration-Tag, jedoch ist dieser nicht in jedem Fall korrekt zugeordnet, da er aus der Texterkennung resultiert. Somit muss die Identifikation von Abbildungen auf Buchseiten vorerst weiterhin technisch gelöst werden, wofür sich CNNs bewährt haben. Hierzu liefert ein Benchmarking mit historischen Dokumenten erste Hinweise.[23]

23 Mehri, Marouna/ Heroux, Pierre/ Gomez-Krämer, Petra/ Mullot, Rémy: „Texture feature bench-marking and evaluation for historical document image analysis", in: International Journal on Document Analysis and Recognition 20 (2017) 1, S. 325–364.

*Abbildung 2: Die Autor*innen dieses Beitrages v.l.n.r.: Thomas Mandl, Sigrun Putjenter, Wiebke Helm, David Zellhöfer und Sebastian Schmideler.*

Chancen und Potenziale künftiger Forschung

Die vorangegangenen Ausführungen haben verschiedene Problemfelder benannt, mit denen sich Forschungsprojekte aus dem Bereich der DH, insbesondere Projekte mit Distant Viewing-Ansatz, konfrontiert sehen, wenn sie ihre Studien auf Daten wissenschaftlicher Bibliotheken aufbauen. Umgekehrt wurden Desiderate und Herausforderungen aufgezeigt, die bibliothekarische Einrichtungen im Hinblick auf die Datenpflege ihrer digitalen Sammlungsbestände managen müssen. Es wurden von bibliothekarischer und informationstechnischer Seite Überlegungen und Lösungsvorschläge für den Import externer Forschungsdaten in bestehende Formal- und Sacherschließungsdaten ebenso diskutiert wie die Integration externer Metadaten in bestehende Workflowsysteme. Anhand der beschriebenen Anwendungsfälle konnte deutlich gemacht werden, dass es weiterer gemeinsamer Anstrengungen bedarf, um Distant Viewing-Forschung mit automatisch erzeugten (bibliothekarischen) Daten zu erleichtern. Der intensive und offene Austausch von Erfahrungen beider Seiten – der Forschung und der Bibliotheken – kann hierbei nur förderlich sein. ❙

Dieser Beitrag entstand im Rahmen des genannten, von der Fritz Thyssen Stiftung geförderten interdisziplinären Forschungsprojekts, der wir für ihre Unterstützung danken.

Wiebke Helm
Wiss. Mitarbeiterin im vorgestellten Kooperationsprojekt
Institut für Pädagogik und Didaktik im Elementar- und Primarbereich
Universität Leipzig
Marschnerstraße 31 H 3, 04109 Leipzig
wiebke.helm@uni-leipzig.de

apl. Prof. Dr. Thomas Mandl
Leiter der Projektgruppe „Entwicklung der Bildikonographie in Wissen vermittelnder Kinder- und Jugendliteratur und Schullehrbüchern des 19. Jahrhunderts. Ein Distant Viewing Ansatz“
Institut für Informationswissenschaft u. Sprachtechnologie
Stiftung Universität Hildesheim
Lübecker Straße 3, 31141 Hildesheim
mandl@uni-hildesheim.de

Sigrun Putjenter
Stv. Leiterin der Kinder- und Jugendbuchabteilung
Staatsbibliothek zu Berlin – Preußischer Kulturbesitz
Westhafenstr. 1, 13353 Berlin
sigrun.putjenter@sbb.spk-berlin.de

Dr. Sebastian Schmideler
Leiter der Projektteilgruppe zum historischen Kinder- und Jugendsachbuch, Wiss. Mitarbeiter im Fachbereich Grundschuldidaktik Deutsch, Institut für Pädagogik und Didaktik im Elementar- und Primarbereich
Universität Leipzig
Marschnerstraße 31 H 3, 04109 Leipzig
sebastian.schmideler@uni-leipzig.de

Dr.-Ing. David Zellhöfer
Stv. Leiter der Abteilung Informations- und Datenmanagement
Staatsbibliothek zu Berlin – Preußischer Kulturbesitz
Potsdamer Str. 33, 10785 Berlin
david.zellhoefer@sbb.spk-berlin.de