Abschließender Sachbericht

Welt der Kinder. Weltwissen und Weltdeutung in Schul- und Kinderbüchern zwischen 1850 und 1918

Leibniz-Einrichtung: Georg-Eckert-Institut. Leibniz-Institut für internationale Schulbuchforschung in Braunschweig
Aktenzeichen: SAW-2014-GEI-2
Projektlaufzeit: Januar 2014 bis Januar 2018
Ansprechpartner: Prof. Dr. Ernesto William De Luca

Final report

CHILDREN AND THEIR WORLD. KNOWLEDGE OF THE WORLD AND ITS INTERPRETATION IN TEXT BOOKS AND CHILDREN'S LITERATURE, 1850-1918

Leibniz-Institute: Georg Eckert Institute for International Textbook Research
Member of the Leibniz Association (GEI)
Reference number: SAW-2014-GEI-2
Project period: Januar 2014 bis Januar 2018
Contact partner: Prof. Dr. Ernesto William De Luca

Inhaltsverzeichnis

Executive summary

Das Projekt „Welt der Kinder“ (WdK) diente der evaluativen Nutzung digitaler Werkzeuge für die Analyse großer Korpora in der historischen Schulbuchforschung und der Gewinnung neuer Erkenntnisse zur Wissensgeschichte des 19. Jahrhunderts durch deren Einsatz. Hierzu arbeitete ein Konsortium aus zwei Leibniz-Instituten (das Georg-Eckert-Institut, Leibniz-Institut für Internationale Schulbuchforschung – GEI und das Deutsches Institut für Internationale Pädagogische Forschung – DIPF) und zwei Universitäten (Technische Universität Darmstadt und Stiftung Universität Hildesheim) interdisziplinär eng zusammen. Das GEI hatte die Projektleitung und arbeitete hermeneutisch für die historische Analyse, während die Kooperationspartner die computer- und informationswissenschaftlichen Aufgaben übernahmen. Ergänzt wurde das Projektteam durch Kooperationspartner im In- und Ausland (Bayerische Staatsbibliothek – BSB, Schweizerisches Institut für Kinder- und Jugendmedien – SIKJM, Universität Zürich, Göttingen Centre For Digital Humanities und die Universitätsbibliothek der Technischen Universität Braunschweig). Ziel war es, die Vermittlung globaler Wissensbestände in deutschen Schulbüchern des ausgehenden 19. Jahrhunderts zu untersuchen und dabei zu fragen, welche Wissensbestände Kindern und Heranwachsenden zwischen 1850 und 1918 zur Konstruktion eines eigenen Weltbildes zur Verfügung standen und wie nach kulturellen Übersetzungen und eigenständigen Entwicklungen in Bezug auf staatlich kontrollierte Wissensbestände in Schulbüchern und stärker marktorientierter Kinder- und Jugendbuchliteratur verlief. In regelmäßigen Treffen wurden die Fragestellungen und Aufgabenverteilungen besprochen und die jeweils nächsten Arbeitsschritte vereinbart. Die Forschungsergebnisse wurden publiziert und auf internationale Tagungen und Workshops diskutiert und verbreitet. Inhaltlich zeigte sich, dass nationale und globale Wissensbestände keine sich ausschließenden Kategorien sind. Ihre Vermittlung und Gewichtung in Schulbüchern hing von mehreren Faktoren ab — insbesondere von der jeweiligen Periode, dem untersuchten Schulfach, der Schulstufe und in geringerem Maße auch dem religiösen Profil der untersuchten Werke. Insgesamt zeigt sich zwar eine Homogenisierung des „nationalen Wissensbestandes“ über die Jahrhundertwende hinweg, doch produzierte der Schulbuchmarkt weiterhin für eine diverse Leserschaft mit regional oder konfessionell unterschiedlicher Schwerpunktsetzung. Schulbücher und Kinder- und Jugendliteratur griffen auch oft auf dieselben Topoi zurück, um Wissen über die außereuropäische Welt für Kinder spannend zu gestalten. Die Übertragbarkeit von digitalen Werkzeugen, die für moderne digitale Korpora entwickelt wurden, wurde für die Anwendung in den Geisteswissenschaften tiefergehend überprüft und getestet. Die im Projekt verfolgten Strategien zeigen den Nutzen und Gewinn digitaler Verfahren zur Analyse großer Korpora sowie das weitere Entwicklungspotential für diese Forschung. Dazu stellt der „Welt der Kinder Explorer“[1] ein wesentliches Ergebnis des Projektes dar. Er demonstriert, wie Texte, die in den Repositorien enthalten sind, exportiert und mit andersartigen Digital-Humanities-Werkzeugen zur weiteren detaillierten Analyse verwendet werden können. Forschende können mit einer Reihe von Texten arbeiten und ihre Suchstrategie konkretisieren, um strukturelle Muster in ihren Quellen aufzudecken, die mit klassisch-hermeneutischen Verfahren nicht erfassbar sind. Digital vorliegende Informationen wurden kombiniert, um spezifische Werkzeuge für die semantische Suche und die statistische Textanalyse zu implementieren, die Forschende dabei unterstützen können, ihre Forschungsfragen besser zu formulieren und den Serendipity-Effekt durch den Einsatz von digitalen Werkzeugen zu unterstützen. Zu diesem Zweck wurden digitalisierte und kuratierte historische Schulbücher des 19. Jahrhunderts auf Seitenebene mit automatisch erkannten Themenmodellen (Topic Models) annotiert und mit zusätzlichen von Hand erhobenen Metadaten versehen. Diese Erweiterungen ermöglichen neben der freien Browsing-Möglichkeit einen komplexen inhalts- und metadatengesteuerten Suchprozess in Schulbüchern.

1 http://wdk.ukp.informatik.tu-darmstadt.de/solr/WdK.dev/browse, zuletzt geprüft am 25. Juli 2018.

Ausgangsfragen und Zielsetzung des Vorhabens

Dem Projekt „Welt der Kinder“ lagen zwei Ausgangsfragen zugrunde, eine historische und eine informations- (und computer-) wissenschaftliche, die in interdisziplinärer Zusammenarbeit von vier Hauptpartnern bearbeitet wurden:

- Georg-Eckert-Institut – Leibniz-Institut für internationale Schulbuchforschung (GEI), Braunschweig
- Deutsches Institut für Internationale Pädagogische Forschung (DIPF), Frankfurt a.M./Berlin
- Institut für Informationswissenschaft und Sprachtechnologie (IWIST), Stiftung Universität Hildesheim, Hildesheim
- Ubiquitous Knowledge Processing (UKP) Lab, Technische Universität Darmstadt, Darmstadt

Unterstützt wurde das Projekt durch fünf Kooperationspartner:

- Institut für Sozialanthropologie und Empirische Kulturwissenschaft (ISEK), Universität Zürich, Zürich
- Bayerische Staatsbibliothek (BSB), München
- Universitätsbibliothek der TU Braunschweig (TUB), Brauschweig
- Göttingen Centre for Digital Humanities (GCDH), Göttingen
- Schweizerisches Institut für Kinder- und Jugendmedien (SIKJM), Zürich.

Mit der historischen Fragestellung wurde untersucht, welches Wissen Kinder im ausgehenden 19. Jahrhundert über die Welt anhand von Schulbüchern der Fächer Realienkunde, Geographie und Geschichte sowie der Kinder- und Jungendliteratur (KJL) der Zeit lernen konnten und wie sich diese neuen Wissensbestände zur Formierung nationalen Wissens im Zuge der Reichsbildung verhielt. Dahinter stand die Beobachtung, dass die zweite Hälfte des 19. Jahrhunderts in Deutschland eine enorme Bildungsexpansion erlebte, bei gleichzeitigem Versuch, das Wissen im Interesse des neuen Nationalstaates zu homogenisieren und damit zum Aufbau einer nationalen Identität beizutragen. Gleichzeitig mussten aber immer größere Wissensbestände nicht nur über Deutschland und seine Nachbarn, sondern über die ganze Welt, in die Schulbücher eingefügt werden, um die neue, sich modernisierende Gesellschaft auf einen globalen Wettbewerb in wirtschaftlicher (und politischer) Hinsicht vorzubereiten. Inwieweit nun dieses neue Weltwissen durch die „nationale Brille“ gefiltert wurden oder ob hier konkurrierende Erzählmuster entstanden, war daher eine weitere Frage im Projekt. Die digitalen Werkzeuge, die zur Analyse eines sonst nicht zu bearbeiteten großen Korpus dienten, sollten so auch genutzt werden, etablierte Forschungsstandpunkte, die an deutlich kleineren Korpora entwickelt wurden, empirisch zu überprüfen.

Für die Computer- und InformationswissenschaftlerInnen des Projektes stand die Fragestellung im Vordergrund, ob die Entwicklung von digitalen Werkzeuge, die für die Analyse von digital born-Texten (Twitter, Facebook, etc.) entwickelt wurden, auch auf historische Textbeispiele übertragen werden können. Damit war die Frage verbunden, welchen Beitrag digitale Werkzeuge, die zur Zeit vor allem in benachbarten Disziplinen wie den Literaturwissenschaften genutzt werden, für die historische Fragestellung des Projektes leisten können. Außerdem wollten die Projektpartner in enger interdisziplinärer Zusammenarbeit klären, inwieweit man mit diesen Werkzeugen der Digital Humanities große Korpora von Texten aus dem 19. Jahrhundert sinnvoll analysieren kann.

Entwicklung der durchgeführten Arbeiten einschließlich Abweichungen vom ursprünglichen Konzept, wissenschaftliche Fehlschläge, Probleme in der Vorhabenorganisation oder technischen Durchführung

Die interdisziplinäre Projektgruppe hat zahlreiche persönliche Arbeitstreffen organisiert, um die Schwierigkeiten räumlich getrennten Arbeitens auf ein Minimum zu reduzieren. Diese Treffen wurden konkret zum Aufbau von Versuchsanordnungen, genutzt die zur Korpusbereinigung und Weiterentwicklung der Tools dienten. Diese wurden dann evaluiert, neue Arbeitsschritte geplant und besprochen. Um den Aussagegehalt der Topics überprüfen zu können, wurde ein Versuchsaufbau generiert, in dem Topics gebildet wurden, die dann mit bekannten Forschungspositionen abgeglichen wurden. Ein Beispiel waren die in der Forschung gut dokumentierte Intervention Kaiser Wilhelms II. und die Verordnungen zweier Schulkonferenzen in Preußen 1890 und 1900, als deren Ergebnis der Anteil antiker Themen in den Schulbüchern zu Gunsten einer stärkeren Beschäftigung mit den nationalen Kriegen des 19. Jahrhunderts zurückgedrängt wurde. Da gymnasiale Geschichtsschulbücher einen großen Bestandteil des Korpus bilden und in diesen im Kaiserreich die klassische Antike einen großen Raum einnahm, war diese Fragestellung gut geeignet, zur Verbesserung des Topic Modelings beizutragen (zur Frage des Topic Modelings als Schwerpunkt des Projektes siehe unten).
Topic Modeling wurde als Methode bevorzugt, da sich mit ihr schnelle Einblicke in das große Korpus (766.260 Seiten aus 3.803 Schulbüchern aus den oben genannten Fächern) gewinnen sowie vorherrschende Themen aufzuspüren ließen. Gleichzeitig wurde eine webbasierte Suchplattform „Welt der Kinder Explorer“ aufgebaut, die über die Projektlaufzeit hinaus der Öffentlichkeit zu Verfügung steht. Diese Plattform ermöglicht unabhängig von der jeweiligen Fragestellung sowohl eine schnelle Durchsuchung und Sortierung der Funde als auch einen Rückgriff auf die im Projekt gewonnen Erkenntnisse, wie etwa die Topics.
Zusätzlich können die aus der Datenbereinigung gewonnen Erkenntnisse (Daten sowie Metadaten) dazu genutzt werden, um die dem Korpus zugrunde liegende Sammlung GEI-Digital[2], eine vom Georg-Eckert-Institut betreute digitale Sammlung historischer Schulbücher bis 1918, zu verbessern.
Mit diesem Versuchsaufbau konnte zum einen die oben vorgestellte Frage zur Veränderung der Epochenzusammensetzung in den Geschichtsschulbüchern relativiert werden; die Antike verlor nicht so sehr an Raum wie bisher angenommen. Als Ergebnis ist zwar ein langsamer Rückgang der Topics, die antike Themen widerspiegeln, zu beobachten, doch scheint dieser nach den im Projekt erreichten Befunden nicht so signifikant, dass man von einem Austausch der Themen reden kann; das Gymnasium blieb ein Hort humanistischer Bildung.
Zum anderen zeigte sich durch die großflächigen Metaanalysen des Korpus die Diversität der Schulbuchproduktion in Deutschland; große Verlage wie Dürr publizierten Schulbücher, die der Reformpädagogik und damit teilweise der Sozialdemokratie nahestanden.
Da bis heute aber keine validen, umfassenden Aussagen über die Nutzung und Verbreitung von Schulbüchern getroffen werden können, kann auch keine Aussage zum Nutzerverhalten – und damit der Verbreitung von Wissensbeständen – gemacht werden.
Weitere Ergebnisse des Projektes betreffen die Auseinandersetzung mit bereits bestehenden Thesen der Geisteswissenschaften. Mit Hilfe der Projektwerkzeuge konnten die Projektmitarbeiter beispielsweise nachweisen, dass neues Wissen schneller in Geographieschulbücher als in Geschichtsschulbücher aufgenommen wurde.

2 http://gei-digital.gei.de, zuletzt geprüft am 25. Juli 2018.

Insbesondere folgende drei Erkenntnisse sind hervorzuheben:

1. Der Versuch, Geographie als vergleichsweise junges Fach sowohl an den Universitäten als auch an den Schulen zu etablieren.
2. Die enge Verzahnung von Schullehrern und geographischen Organisationen im Kaiserreich, denn Lehrer waren lange wichtige Protagonisten in diesen Organisationen.
3. Um das Fach zu etablieren, entwickelte sich eine Reihe von Geographen schon früh zu Propagandisten des deutschen Imperialismus. Dies hängt eng damit zusammen, dass wichtige Akteure dieser Wissensrevolution, wie Ferdinand von Richthofen, nicht nur im außereuropäischen Raum selbst forschend tätig waren, sondern diese Forschungsleistung als direkten Beitrag zum Ausbau des deutschen Kolonialreiches verstanden.

Allen Erkenntnissen lag die Nutzung der im Projekt entwickelten technischen Infrastruktur „Welt der Kinder Explorer“ zugrunde. Für die Forschungsfragen des Projektes, haben die Facettierung (Filterung der Parameter) und die vergleichende Analyse der Worthäufigkeiten eine entscheidende Rolle gespielt. So konnte gezeigt werden, dass in den untersuchten Zeiträumen neue didaktische Konzepte und Vermittlungsansätze in Schulbüchern umgesetzt wurden, obwohl es dafür noch gar keine staatlichen Vorgaben gab. Bezeichnend für diese Entwicklung ist der im Projekt vielfach erbrachte Nachweis, dass die Sprache in Schulbüchern für niedere Schulen deutlich einfacher gehalten war als in denen für höhere Bildungseinrichtungen. Weitere Ergebnisse sind, dass sich Nationalisierung und Globalisierung in ihrer Beschreibung als gegenläufige Prozesse nicht ausschließen. Dies zeigt sich besonders deutlich bei den ineinander verschränkten Erläuterungen von Imperien und Kolonialisierung – obwohl eigentlich historisch voneinander unabhängige, wenn auch parallele Prozesse wurden sie in Schulbüchern häufig miteinander in Verbindung gebracht und spiegeln damit auch das Weltbild der Zeitgenossen in Deutschland. Gleichzeitig wurde das Wissen über die Welt durch die Globalisierung immer größer, immer mehr Fakten und ausführlichere Länderdarstellungen mussten vor allem in die Bücher für höhere Schulen integriert werden. Nationalisierungsprozesse und das Streben nach imperialer Macht geben jedoch das Deutungsschema für das neue Wissen ab 1890 immer ausgeprägter vor; dennoch lassen sich hier zum Beispiel Unterschiede zwischen den Büchern katholischer und protestantischer Verlage finden.

Diese Erkenntnisse wurden außerdem auf der Tagung „Die Welt der Kinder: Weltwissen und Weltdeutung in Schul- und Kinderbüchern des 19. und frühen 20. Jahrhunderts“ diskutiert. Die Veranstaltung wurde von den Projektpartnern an der Universität Zürich organisiert, und diente auch den Vergleich der Schulbuchinhalte mit der zeitgenössischen Kinder- und Jugendliteratur, der in einem reichhaltigen inhaltlichen Austausch resultierte.

Auf der informationswissenschaftlichen Seite erwies sich eine Erhebung zusätzlicher Metadaten als notwendig, um die vorhandenen bibliographische Standards zu ergänzen, und diente so zur Aufbereitung des Korpus und für eine daraus folgende gewünschte komplexe technische Analyse. Über die komplette Projektlaufzeit wurde seitens des GEI Google Refine [3] benutzt, um die umfangreiche (semi-) automatische Metadatenaufbereitung durchzuführen. Die Arbeiten wurden bereits für die Einbindung in den GEI Repositorien vorbereitet.

3 https://code.google.com/archive/p/google-refine/, zuletzt geprüft am 25. Juli 2018.

Bei der Arbeit mit den Volltexten befand sich die Fehlerquote bei der OCR-Erkennung im üblichen Bereich. Momentan existieren keine Verfahren, die eine 100% OCR-Erkennung gewährleisten. Die Anwendung moderner Tools auf historische Texte musste mit folgenden Herausforderungen umgehen:

- Eine höhere Fehlerquote lag vor allem an dem hohen Anteil an Texten in Fraktur-Schrift.
- Ebenso bereiteten die historische Orthographie sowie Bedeutungsverschiebungen auf Wortebene sowohl innerhalb des Untersuchungszeitraums als auch im Vergleich zur Gegenwart unerwartete Probleme.
- Die zur Datenbereinigung zur Verfügung stehenden Wörterbücher waren veraltet oder enthielten keine ungewohnten Schreibweisen. Hierfür war eine händische Korrektur anhand historischer Wörterbücher für das Projekt im Projektzeitraum nicht zu leisten.
- Andere mögliche Fehlerquellen, wie wechselndes Druckbild (zum Beispiel Zitate in altgriechischen oder hebräischen Lettern), blieben dagegen statistisch insignifikant.
- Eine unerwartete Fehlerquelle waren die inkonsistenten Auszeichnungen von Kapitelauszeichnungen, Tabellen, Bildern, Bildunterschriften etc. durch den Dienstleister, der die OCR-Volltextumwandlung geleistet hatte.

Hinzu kam die intensive Vorbereitung und Auseinandersetzung mit dem Themenkomplex Topic Modeling und dem „Welt der Kinder Explorer“. Daher konzentrierten sich die Darmstädter Projektmitarbeiter auf eine prototypische Umsetzung einer semantischen Analyse mittels Opinion Mining, sowie mit weiteren Verfahren wie z.B. Annotationsvergabe, Word Clustering, Word Embeddings und Ontologien (semantische Verknüpfungen), die in einem frühen Experimentierstadium blieben. Eine vertiefte Auswertung ist für ein Nachfolgeprojekt geplant.

Im Ergebnis zeigte sich in diesem Zusammenhang, dass heutige semantische Technologien nur bedingt erfolgreich eingesetzt werden konnten, nicht zuletzt aufgrund der bereits beschriebenen Qualität der Texte und der veränderten Sprache.

Allgemeine semantische Annotationsverfahren sind in der Regel schwer anzuwenden. Eine weitergehende händische semantische Auszeichnung war für diesen Korpus nicht geplant, auch wegen der komplexen semantischen Tiefe. Aufgrund der Sender- und Empfängerposition von Schulbuchwissen war ein bipolar (positiv oder negativ) ausgerichtetes sowie bisher ausschließlich für kurze Texte und im Bereich der Wirtschaftsforschung (Amazon, Google, etc.) zunächst getestete Opinion Mining Methoden schwer auf den Korpus von „Welt der Kinder“ zu übertragen.
Unter Opinion Mining wurde im Projekt unter Anlehnung an neuere Arbeiten der Critical Discourse Analysis die automatische Extraktion meinungstragender Aussagen des Textes, die zur Analyse politischer Standpunkte oder rassistischer Vorurteile herangezogen werden können, verstanden.[4] In Hinblick auf Meinungsanalysen wurden Experimente mit Joint Semantic-Topic Models mit eigens angepassten Versionen des deutschsprachigen Sentiment-Lexikons SentiWS durchgeführt.[5] Die Ergebnisse auf Ausschnitten der WdK-Daten lieferten jedoch keine intuitiv nachvollziehbaren Meinungsanalysen.

4 Für eine Anwendung der Critical Discourse Analysis auf einen verwandten Korpus mit ähnlichen Methoden siehe Rash, Felicity: German Images of the Self and the Other: Nationalist, Colonialist and Anti-Semitic Discourse 1871-1918. Basingstoke: Palgrave Macmillan, 2012.
5 Lin, Chenghua; He, Yulan He: Joint Sentiment/Topic Model for Sentiment Analysis, In: Proceedings of the 18th ACM Conference on Information and Knowledge Management, 375–84. New York, NY, USA: ACM, 2009.

Hinzu kommt, dass historische Texte viele Sichtweisen enthalten und in Aufbau und Inhalt sich von den oben genannten Vergleichstexten unterscheiden. Die im Projekt getesteten Annotationswerkzeuge Pundit[6] und WebAnno[7] boten vielversprechende Möglichkeiten einer Analyse des Korpus. Allerdings war deren Nutzbarkeit für historische Korpora stark limitiert. Deren Anwendung war entweder sehr zeitintensiv (Pundit) oder die automatische Korrektur (wie bei Webanno) lieferte aufgrund der fehlenden Lexika und der nicht vergleichbaren Sprache in den vorhandenen Korpora keine zufriedenstellenden Ergebnisse.
Daher entschied das Projektteam sich für die Kollokationsanalysen als mögliches Ersatzszenario für das Opinion Mining. Allerdings ergab sich während der Projektarbeit, dass die fehlende Redundanz in den Texten und weniger die explizite Meinungsäußerung die Ursache für nicht funktionierende Opinion Mining-Verfahren für Schulbuchanalysen war. Die vorhandenen, sich nicht wiederholenden Meinungen waren nicht geeignet, um ein automatisiertes Modell zur Meinungsfindung zu bilden. Dieses Ergebnis war unerwartet, da aus didaktischen Gründen klare, sich wiederholende Aussagen zu erwarten gewesen wären.
Hermeneutische Vorarbeiten waren für eine weitergehende Analyse zwingend notwendig, da für die Interpretation weiterhin ein Verständnis der historischen Sprache nötig war. Unter anderem war es wichtig, Unterschiede im Sprachgebrauch und in der Orthographie sowie eine Kontextualisierung der Begriffe zu erkennen und für eine moderne Korpusarbeit unabweisbar.

Auf der Repräsentation von Texten über Worthäufigkeiten und den enthaltenen semantischen Einheiten konnten weitere automatische Verfahren getestet werden, die das Erkennen von Mustern ermöglichen. Für die historische Schulbuchforschung stecken solche Verfahren noch in den Anfängen, doch gibt es schon vielversprechende Untersuchungen zu aktuellen Schulbüchern sowie zu Parteiprogrammen, die diese nach den in ihnen vertretenen politischen Einstellungen sortieren. [8] Das Clustering [9] (Gruppieren) von Dokumenten ist dagegen ergebnisoffen und erkennt Gruppen ähnlicher Dokumente, ohne dass Klassen vorgegeben werden müssen. Beim Topic Modeling[10] (Themenmodellierung) wird versucht, thematische Zusammenhänge in Dokumentensammlungen wie GEI-Digital aufzudecken. Dazu wird errechnet, welche Worte häufig zusammen auftreten (Wortlisten, die ein Topic ergeben) und wie relevant jedes Topic für ein Dokument ist.

```
0      0.05117 general armee heer mann truppe schlacht feind franzose festung preußen
1      0.03174 gott tempel mensch gtter priester zeus held erde opfer himmel
2      0.04014 amerika staat afrika europa nordamerika kolonie asien mill insel spanier
3      0.04371 könig sohn kaiser tod vater jahr reich alexander bruder nachfolger
4      0.04955 periode erster abschnitt zweiter karte zeitraum aufl vgl bild dritter
5      0.03801 rom cäsar senat pompejus jahr sulla antonius marius csar provinz
6      0.0602  bauer land geld adel recht bürger gut million staat stand
7      0.04357 mann pferd feind reiter schwert waffe soldat hand ritter heer
8      0.04373 stadt kirche schloß berlin gebäude bau dom tempel denkmal burg
```

Abbildung 1: Liste mit Topics als Ausgabe des Modellierungsprozesses mit den zehn wichtigsten Termen je Topic im „Welt der Kinder“-Projekt

6 http://thepund.it/semantic-web-annotation/ zuletzt geprüft am 25. Juli 2018.
7 https://webanno.github.io/webanno/ zuletzt geprüft am 25. Juli 2018.
8 Slapin, Jonathan B. und Sven-Oliver Proksch. “A Scaling Model for Estimating Time-Series Party Positions from Texts”, in: American Journal of Political Science 52 (2008), 3, 705-722; Slopinski, Andreas und Torsten J. Selck. „Wie lassen sich Wertaussagen in Schulbüchern aufspüren? Ein politikwissenschaftlicher Vorschlag zur quantitativen Schulbuchanalyse am Beispiel des Themenkomplexes der europäischen Integration“, in: JEMMS – Journal of Educational Media, Memory and Society 6 (2014), 1, 124-141.
9 Mitchell, Thomas M. “Machine Learning” (1 ed.). McGraw-Hill, Inc., New York, NY, USA. 1997.
10 Blei, David M. “Topic Modeling and Digital Humanities”, in: Journal of Digital Humanities 2 (1), 2012.

Abbildung 1 zeigt eine Liste mit Topics aus dem „Welt der Kinder“-Projekt. Angezeigt werden die jeweils 10 für das Topic relevantesten Wörter. Ein Vorteil gegenüber dem Dokumenten-Clustering ist, dass Dokumente so gleichzeitig mehreren Themenbereichen zugeordnet werden können und dass durch die Topic-Wörter im Idealfall aussagekräftige Titel für die Themenfelder entstehen. Als ein Ergebnis des Topic Modeling können die Wortlisten für die Topics selbst ausgewertet werden. Weiterhin können die Dokumente einer Sammlung über die Topic-Zuordnung gefiltert und statistische Häufigkeiten von Topics in Ergebnismengen und Untermengen einer Sammlung angegeben werden. Im Projekt „Welt der Kinder“ wird von diesen Möglichkeiten intensiv Gebrauch gemacht. Um selbst damit zu experimentieren, eignet sich ein Topic Modeling Tool, das eine Benutzeroberfläche bereitstellt. So könnte Topic 0 in der Abbildung 1 z. B. als „Dt.- Frz. Krieg 1870/71“, Topic 1 als „Religion im antiken Griechenland“ o. ä. betitelt werden, und man kann für jedes Buch der Sammlung berechnen, wie „präsent“ dieses Thema darin ist. Dabei muss jedoch vorab festgelegt werden, wie viele Themenfelder erwartet werden. Dies und weitere anpassbare Parameter der Modellierung führen dazu, dass die Ergebnisse nicht als objektiv, sondern eher als Betrachtung einer Quellensammlung durch eine speziell für die Fragestellung geschliffene Linse wahrgenommen werden sollten.[11]
Für die Ansätze Joint Semantic-Topic Models (JST) und Dynamic Topic Models (DTM) existierte bis dahin nur eine auf bestimmte Korpusformate und -dimensionierungen spezialisierte Implementation, deren Einsatz auf dem WdK-Korpus erheblichen technischen Aufwand nach sich zog. Für JST wurde dazu ein erheblicher Konvertierungsaufwand und Experimente mit unterschiedlichen Parametereinstellungen unternommen, ohne signifikant bessere Ergebnisse. Die (Neu-)Berechnungen der Topics wurden auf den Servern der TU Darmstadt durchgeführt und standen stets abrufbereit zur Verfügung. Dadurch konnte man verschiedene Topic-Listen und verschieden Subkorpora erstellen, wenn die Daten auf eigene Rechner übertragen wurden.

Topic Modeling erwies sich als erfolgversprechendstes Werkzeug im Projekt. Aber seine Komplexität zeigte, zum Beispiel beim Topic Labelling, die Problematik der Ambiguität. Aufgrund der Zusammenstellung des Korpus sowie seiner statischen Aufbereitung waren die erzeugten Themen hinter den jeweiligen Topics teilweise mehrdeutig. Die Erstellung unterschiedlicher Subkorpora wurde für eine bessere Topic-Analyse umgesetzt, um die Mehrdeutigkeit zu reduzieren. Daher wurden bestimmte Parameter ausgewählt, wie z.B. die Gewichtungen von Topics pro Seite. Ebenso wurden Wörter, die sich nicht unter den 100 wichtigsten Wörtern des Topics befanden, mit Null gewichtet. In einem weiteren Schritt wurden dann auf dieser Seite nur die Topics berücksichtigt, deren Seed Words alle in den Top 100 des Topics vorhanden waren.[12]
Die Herausforderung des Topic Modeling-Verfahrens besteht darin, dass sich semantische Verschiebungen im Zeitverlauf nicht abbilden lassen. Im Projekt hat man sich deshalb entschieden, durch zeitlich (nach Jahrzehnt) getrennte statisch berechnete Topics im Vergleich zu evaluieren, um dieser Problematik auf den Grund zu gehen. Hier war die disziplinäre Interpretation der unterschiedlichen Partner abweichend. Während sich die Historiker gerade für die Bedeutungsverschiebung interessierten, erschienen den Informatikern die relativen Verteilungen innerhalb der Topic-Wortlisten für eine statistische Auswertung nicht ausreichend.

11 Di Maggio, Paul, Manish Nag und David Blei. “Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding”, in: Poetics 41 (6), 2013, 570–606.
12 Schnober, Carsten; Gurevych, Iryna: Combining Topic Models for Corpus Exploration: Applying LDA for Complex Corpus Research Tasks in a Digital Humanities Project, in: Proceedings of the 2015 Workshop on Topic Models: Post-Processing and Applications, TM ’15. New York, NY, USA: ACM, 2015, S. 11–20, hier 17.

Letztlich zeigte sich sehr deutlich, dass aktuelle Texte gut zu bewältigen sind, das Verfahren für historische Texte hingegen nicht ausreichend geeignet ist und daher weitergehender Forschungsbedarf besteht.

Insgesamt handelte es sich so um ein sehr zeit- und arbeitsaufwendiges Verfahren, das enorme Herausforderungen an die möglichen Testläufe in einem zeitlich knapp bemessenen Projekt stellt. Hierzu ist anzumerken, dass interdisziplinäre Arbeit sehr komplex ist und die nicht nur die fachliche Expertise, sondern auch die Fähigkeit der Projektpartner, in einer gemeinsamen bzw. für alle verständlichen Sprachen miteinander zu kommunizieren, für den Erfolg eines solchen Projektes ausschlaggebend ist. Die Expertise aller beteiligten Fachrichtungen ist außerdem sehr wichtig und deshalb bleibt eine kooperative Zusammenarbeit zur Analyse solch großer Korpora sinnvoll. Während der Implementierung der Topic Modeling-Methoden wurde offensichtlich, dass die OCR-Erkennungsfehler ein relevantes Problem bei der automatischen Textverarbeitung darstellen. Deshalb musste daraufhin eine Strategie erarbeitet und Fachwissen aufgebaut werden. Es wurden verschiedene Implementationen von Topic Modeling-Verfahren auf Grundlage von LDA getestet und ein Topic Modeling-Modul für DKPro auf Basis von Mallet entwickelt.[13] Das Modul generiert einerseits Topic Models auf Basis der vorhandenen Textdaten und inferiert andererseits Topic-Verteilungen in einzelnen Dokumenten sowie in frei definierbaren, diachronen und synchronen Teilkorpora. Speziell für die diachrone Analyse wurden erste Experimente mit Dynamic Topic Models mit Ausschnitten des WdK-Korpus durchgeführt.[14]
Es wurde ausführlich mit statistischen Parametern sowie Textfiltern experimentiert. Die daraus resultierenden unterschiedlichen Topic Models wurden gemeinsam mit IWIST und GEI evaluiert.[15] Zentrale Parameter bilden dabei die Anzahl der Topics sowie Wortfilter, die die Modellierung auf Inhaltswörter begrenzen.

Es wurde eine vollständige Pipeline (siehe Abbildung 2) für die Verarbeitung des WdK-Korpus implementiert, bei der der Korpus eingelesen, verarbeitet und gefiltert wird und Topic Models generiert werden. Aus dem Textkorpus erzeugte Topic Models wurden anschließend verwendet, um einzelne Textabschnitte mit Topic-Gewichten zu annotieren. Zu diesem Zweck wurde eine entsprechende Verarbeitungs-Pipeline implementiert. Diese Verarbeitungskette wurde in der Programmiersprache Java umgesetzt, basiert auf dem Framework DKPro[16] und kann nur durch einen Eingriff in den Programmcode angepasst werden, d.h. nicht durch mit der Benutzeroberfläche Arbeitende. Auffällige Fehler können jedoch gesammelt und für den nächsten Vorverarbeitungsdurchlauf berücksichtigt werden. Um einen schnellen Zugriff auf Dokumente zu ermöglichen wurde mit der Suchmaschine Apache Solr[17] ein Index erstellt, der zurzeit 4.039.066 Wörter enthält. Dabei wurden besonders häufig auftretende, jedoch wenig inhaltstragende Wörter vorab entfernt (sog. Stoppwörter).

13 Blei, David M.; Ng, Andrew Y.; Jordan,Michael I.: Latent Dirichlet Allocation.” Journal of Machine Learning Research 3 (March 2003), S. 993–1022; McCallum, Andrew Kachites. MALLET: A Machine Learning for Language Toolkit., 2002;
14 Blei, David M.; Lafferty,John D.: Dynamic Topic Models, in: Proceedings of the 23rd International Conference on Machine Learning, 113–20. ICML ’06. New York, NY, USA: ACM, 2006; doi:10.1145/1143844.1143859.
15 Wallach, Hanna M.; Murray, Iain; Salakhutdinov, Ruslan; Mimno, David: Evaluation Methods for Topic Models, in Proceedings of the 26th Annual International Conference on Machine Learning, 1105–12. Montréal, Québec, Canada: ACM, 2009; Chang, Jonathan; Gerrish, Sean; Wang, Chong; Boyd-Graber, Jordan L.; Blei,David M.: Reading Tea Leaves: How Humans Interpret Topic Models, in: Advances in Neural Information Processing Systems 22, 288–96. Vancouver, British Columbia, Canada, 2009;
16 https://dkpro.github.io/, zuletzt geprüft am 25. Juli 2018.
17 https://lucene.apache.org/solr/, zuletzt geprüft am 25. Juli 2018.

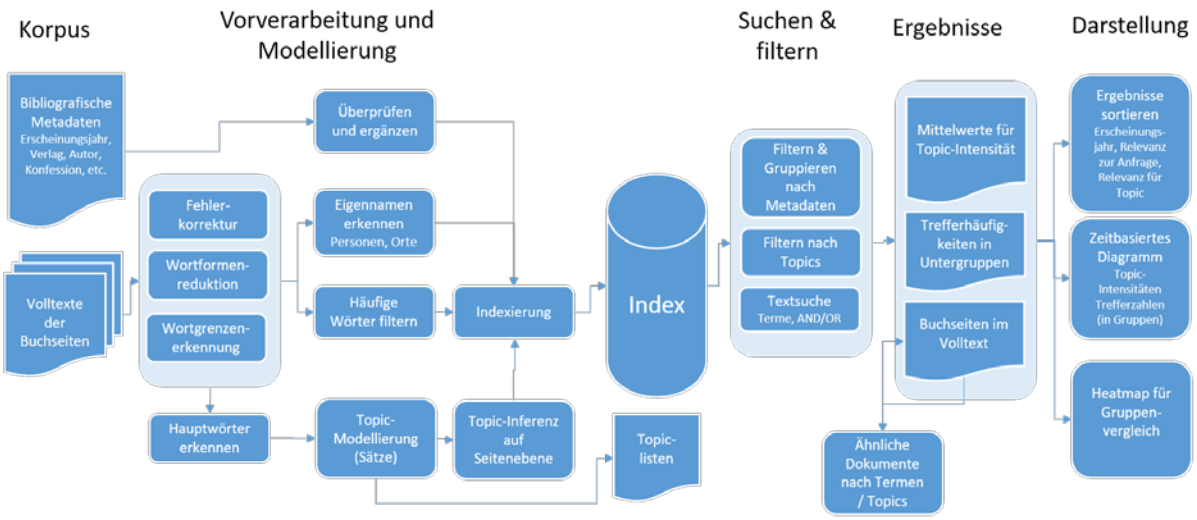

Abbildung 2: Schematische Darstellung der Datenstrukturen und Benutzeroberfläche für das Projekt "Welt der Kinder"

Es gibt zahlreiche Topic Modeling-Varianten und -Implementationen. Die Sondierung der für „Welt der Kinder“ sowohl technisch als auch inhaltlich geeigneten Lösung gestaltete sich aufwändiger als geplant. Die einschlägige Fachliteratur beschreibt einige Fallstudien auch aus dem Bereich der Digital Humanities.[18] Die Optimierung der Parameter für das Topic Modeling lässt sich dennoch nicht automatisieren und ist daher aufwändig. Bei der Evaluierung eines Topic Models ergibt sich zusätzliche Komplexität aus der Tatsache, dass sowohl die Qualität der Topic-Repräsentationen eines Modells als auch der daraus resultierenden Topic-Gewicht-Annotationen für einzelne Dokumente und Teilkorpora zu evaluieren waren. Diese beiden Parameter hängen nicht direkt miteinander zusammen. Ein optimales Topic Model liefert sowohl intuitiv nachvollziehbare Topic-Definitionen als auch aussagekräftige Topic-Gewicht-Annotationen.

Der Aspekt der Kinder- und Jugendliteratur wurde auf Konferenzen und in Publikationen diskutiert. Schwierigkeiten im Vergleich beider Textgattungen (Schulbuch und KJL) mit digitalen Werkzeugen ergaben sich daraus, dass die ursprünglich vorliegenden Digitalisate von zu geringer Qualität waren. Hier unternahm der Projektpartner Bayerische Staatsbibliothek eine aufwendige Nachdigitalisierung.
Ebenso stellte der Kooperationspartner Universitätsbibliothek Braunschweig auf eigene Kosten zum Ende der Projektlaufzeit die im Projekt nachbearbeiteten Digitalisate der Öffentlichkeit zur Verfügung. Grundlage dieses digitalisierten Bestandes bildete die an der Universitätsbibliothek der TU Braunschweig vorhanden Sammlung historisch wertvoller Kinder- und Jugendbücher (Sammlung Hobrecker [19]). Ein tiefgehender Vergleich der beiden Korpora mit einer Analyse anhand digitaler Werkzeuge ist für ein Nachfolgeprojekt vorgesehen.[20]

18 Templeton, Clay; Brown, Travis; Battacharyya, Sayan; Boyd-Graber, Jordan: Mining the Dispatch under Supervision: Using Casualty Counts to Guide Topics from the Richmond Daily Dispatch Cor, in: Chicago Colloquium on Digital Humanities and Computer Science. Chicago, Illinois, USA, 2011; http://www.umiacs.umd.edu/~jbg/docs/slda_civil_war.pdf.
19 Zur Sammlung siehe https://ub.tu-braunschweig.de/recherche/kinderbuch.php, zuletzt geprüft am 25. Juli 2018. Um mehr über die Reichweite dieser Bücher zu erfahren wurde ein Gutachten in Auftrag gegeben.
20 Die Bücher sind unter https://publikationsserver.tu-braunschweig.de/servlets/solr/select?q=category.top%3ADDC\%3A398+AND+state%3Apublished einsehbar. Zuletzt geprüft am 25. Juli 2018.

Darstellung der erreichten Ergebnisse und Diskussion im Hinblick auf den relevanten Forschungsstand, mögliche Anwendungsperspektiven und denkbare Folgevorhaben

Das Hauptziel des Projektes war die Überprüfung der Nutzung von digitalen Werkzeugen im Zusammenhang mit etablierten Forschungsmethoden in den Geschichtswissenschaften. Hierbei ergaben sich neuartige Perspektiven für die Forschungsarbeit und Modifikationen von etablierten Geschichtsinterpretationen.
Folgende Ansätze für die Visualisierung von Topic Models wurden mit den WdK-Daten getestet beziehungsweise sondiert: DiTop-View, Serendip, TMVE, Topic Browser. Forschende können nur in kleinen Datenmengen klare Aussagen erkennen, die selten neue Erkenntnisse bringen. Die Visualisierung von großen Datenbeständen hingegen ist unübersichtlich und unbrauchbar, da zu viele Informationen zeitgleich angezeigt werden.
Es wurde der „Welt der Kinder Explorer“ entwickelt, mit dem der Korpus durchsucht werden kann. Dieses Tool ist für die historische Schulbuchforschung vom großen Nutzen, da Forschende mit statistischen Verfahren geschichtswissenschaftliche Fragestellungen experimentell überprüfen und mit hermeneutischen Methoden abgleichen können. Dies kann als innovative und interaktive Ergänzung zu den etablierten geisteswissenschaftlichen Forschungsmethoden dienen.
Eine andere Innovation ist die Nutzung der variablen Zusammenstellungsmöglichkeiten von Unterkorpora und die Exportfunktion der Originaltexte, sowie auch deren normalisierten Versionen. Diese Exportfunktion erlaubt es individuell, über Topics, Facetten oder Anfragen generierte und so spezifisch an den Untersuchungsgegenstand angepasste Korpora mit anderen Web-Tools (Antconc, voyant-tools, Context, Weblicht etc.) weiter zu ver- und bearbeiten. Das System stellt dynamische Möglichkeiten zur Verfügung, die durch seine vom System erzeugten Indizes weitere Anpassungs- und Einbindungsoptionen schafft.
Es ist geplant, dass das Werkzeug so erweiterbar wird, dass nachdigitalisierte Bücher oder andere Korpora integriert werden können. Für die Recherchen steht momentan nur der kreierte Korpus zur Verfügung. Dieser kann mit anderen Platt- und Darstellungsformen verknüpft werden, da Suchergebnisse exportierbar sind. Die enthaltenen Metadaten können dementsprechend auch nachbearbeitet und verfeinert werden. Außerdem bietet es eine Visualisierung statistischer Trends sowie eine intuitiv bedienbare Nutzeroberfläche.

Am Ende der Projektlaufzeit hat die Projektgruppe Experimente mit Ontologien durchgeführt und eine Anbindung an die Linked Open Data Cloud getestet, um Semantische Technologien und Sentiment Analyse Methoden zu erforschen. Die Ergebnisse sind als Ontologie-Verknüpfungen zum Index des Projektes ergänzt worden und dementsprechend in der Nutzeroberfläche verfügbar. Die Testläufe offenbarten die Chancen für historischen Erkenntnisgewinn, welche eine solche Verknüpfung mit sich bringen können. So wurde mit der Integration der *deutschen-biographie.de* experimentiert, um Hintergrundinformationen über die in den Schulbüchern erwähnten Personen über ein automatisiertes Verfahren für weitere Annotationen zur Verfügung zu stellen.

Ein weiteres erreichtes Ziel der Korpusbearbeitung war, immer wieder neue Abschnitte als Grundlage für die Topicberechnung zu unterteilen; erste Ansätze hierzu wurden im Projekt erprobt. Sie werden nach Ablauf der Förderung in Braunschweig fortgeführt.

Ebenso konnten Synergien mit verschiedenen Plattformen des Georg-Eckert-Instituts hergestellt werden, unter anderem mit den vom niedersächsischen MWK geförderten Projekt „International TextbookCat“ und dem vom BMBF geförderten Projekt „Semantische Konzepte in Schulbüchern“. Das entstandene System wurde vom Georg-Eckert-Institut in die eigene

digitale Infrastruktur integriert und weiterentwickelt, so dass es der Forschung längerfristig zur Verfügung gestellt werden kann. Außerdem ist eine Abstrahierung der Integration der Datenbestände geplant, so dass das System Korpus-unabhängig sein wird. Hier wird die Möglichkeit bestehen, eigene Korpora über standardisierte Schnittstellen einzubinden sowie Forschungsergebnisse von anderen WissenschaftlerInnen weiter zu nutzen und zu analysieren.
Zur historischen Forschung über die Entwicklung von globalisierten Wissensbeständen für die Jugendlichen des Kaiserreichs trug das Projekt durch die empirische Überprüfung und Erweiterung etablierter Forschungsergebnisse bei. Neue Erkenntnisse wurden in zahlreichen Publikationen veröffentlicht und werden in einer Dissertation darstellt, die 2019 an der TU Braunschweig eingereicht wird.
Die Projektpartner der Universität Hildesheim betreuten vor allem die Evaluationen der einzelnen Projektschritte. Ein wichtiger Aspekt lag in der Vermittlung zwischen den verschiedenen Disziplinen, um eine gemeinsame Sprache für das Projekt zu entwickeln, damit ein Verständnis über Disziplingrenzen hinweg entwickelt werden konnte, um die einzelne Arbeitsschritte zu koordinieren. Dieser begleitende Prozess und vor allem die verschiedenen Visualisierungsmöglichkeiten der Projektergebnisse wurden publiziert.
Um die Verwendbarkeit der entwickelten Methoden für die Geisteswissenschaften gewährleisten zu können, wurden über die Projektlaufzeit mehrere Evaluierungen durchgeführt. Eine große abschließende Evaluierung des Prototyps (siehe Abbildung 3) wurde im Rahmen eines Tutorials am Georg-Eckert-Institut ausgeführt. Teilgenommen haben neun ForscherInnen aus dem Institut, ein externer Doktorand, der an einem eng mit dem Projekt verknüpften Thema arbeitet, sowie zwei VertreterInnen der Bibliothek und der Informationsabteilung. Eine detaillierte Auswertung ist dem Sachbericht der Stiftung Universität Hildesheim zu entnehmen.

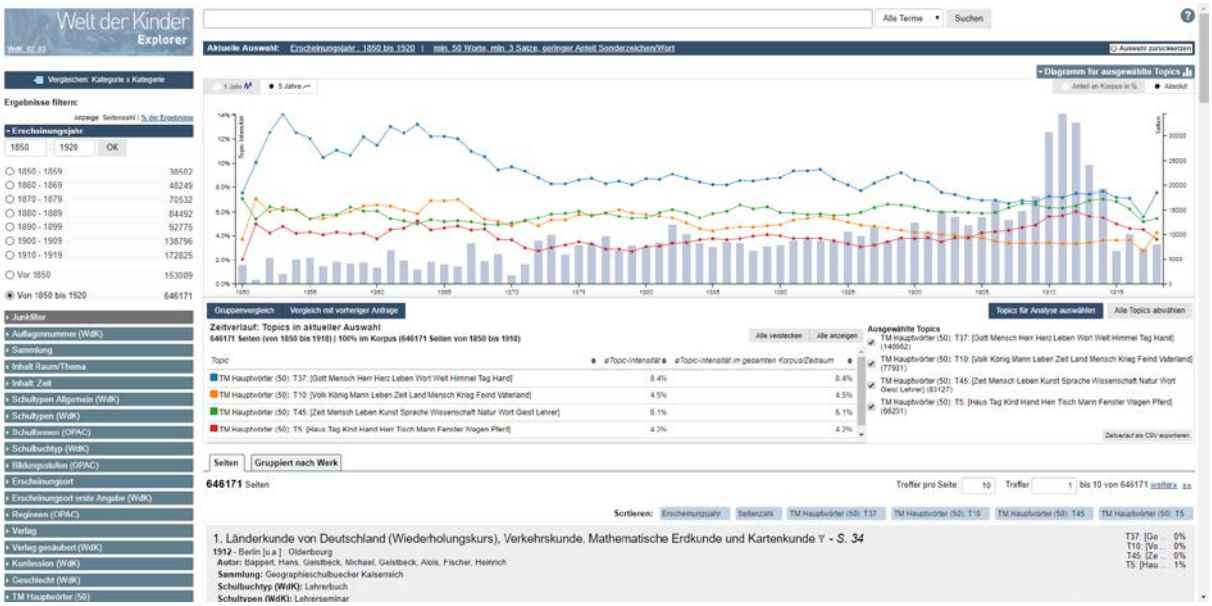

Abbildung 3: Screenshot des Digital-Humanities-Werkzeugs “Welt der Kinder”-Explorer

Die meisten Dienste des GEI erlauben einen Zugriff auf die Daten über standardisierte Schnittstellen (wie z.B. International TextbookCat und GEI-Digital). Je nach Schwerpunkt unterstützen sie die Forschung bei der Recherche und Zusammenstellung von Quellen, bei deren Analyse und Auswertung, deren Publikation und Speicherung.

Aus diesem Grund wurde der „Welt der Kinder Explorer“ in die digitale Infrastruktur des GEI integriert und nimmt damit einen festen Platz innerhalb der Forschungsinfrastrukturen des GEI ein. Kooperationen mit der Bibliothek für Bildungsgeschichtliche Forschung des DIPFs (Prof. Dr. Sabine Reh) und mit der Universität Osnabrück (Prof. Dr. phil. Christian Dawidowski) sind bereits entstanden. Infolgedessen wird derzeit die Einbindung anderer Korpora auf Testsystemen erprobt.

Präsentationen auf Konferenzen zeigten, dass die Gesamtarchitektur nicht nur in der historischen Bildungsmedienforschung auf Interesse stößt. Mit entsprechender Erweiterung kann das Werkzeug so modifiziert werden, dass es auf generelle Korpora digitalisierter Bestände anwendbar wäre.

Stellungnahme, ob Ergebnisse des Vorhabens wirtschaftlich verwertbar sind und ob eine solche Verwertung erfolgt oder zu erwarten ist; Angaben zu möglichen Patenten oder Industriekooperationen

Eine wirtschaftliche Verwertung ist nicht gegeben, da sowohl Software wie auch Publikationen, soweit wie möglich, als OpenSource gemäß den Richtlinien der Leibniz-Gemeinschaft der Öffentlichkeit zur Verfügung gestellt werden.

Die im Projekt entwickelten technischen Infrastruktur „Welt der Kinder Explorer“ steht für andere Institute und Universitäten nachnutzbar über GitHub [21] zur Verfügung, einem Onlinedienst, der Software-Entwicklungsprojekte auf seinen Servern bereitstellt (Filehosting).

Es ergaben sich keine möglichen Patente oder Industriekooperationen.

21 https://github.com/UKPLab/corpus-explorer.git, zuletzt geprüft am 25. Juli 2018.

Angabe der Beiträge von möglichen Kooperationspartnern im In- und Ausland, die zu den Ergebnissen des Vorhabens beigetragen haben

Die Aufgaben verteilten sich in Umfang und Gewichtung für das Projekt gemäß dem Anteil im Gesamtprojekt.

Während die Bayerische Staatsbibliothek, das Schweizerisches Institut für Kinder- und Jugendmedien und die Universitätsbibliothek Braunschweig vor allem bei der Bereitstellung digitalisierter Kinder- und Jugendliteratur sowie der Erweiterung des Bestandes in GEI-Digital wichtige infrastrukturelle Unterstützung leisteten, unterstütze der Lehrstuhl für Populäre Literaturen und Medien, Kinder- und Jugendmedien an der Universität Zürich durch seine Leiterin Professor Dr. Ingrid Tomkowiak das Projekt besonders mit seiner Expertise zur Kinder- und Jugendliteratur. Diese Erfahrung floss auch in die Organisation einer gemeinsamen Tagung an der Universität Zürich vom 4. bis 6. Februar 2016 ein.

Ebenfalls diente die Zusammenarbeit mit Professor Dr. Gerhard Lauer vom Göttingen Centre For Digital Humanities (GCDH) der Erweiterung der literaturwissenschaftlichen Expertise im Projekt. Sein Teilprojekt widmete sich der Erprobung neuer Wege der automatisierten Textanalyse.
Neben der bereits etablierten Möglichkeit zur Untersuchung des Korpus mithilfe von Topic Modeling[22] sollte erkundet werden, in wieweit Ansätze wie Sentiment-Analyse oder auch Argumentanalyse geeignet sind, um historische Muster der Bewertung und Argumentation automatisiert auffinden zu können. Dazu organisierte Professor Lauer am 24. und 25. November 2016 am GCD einen Workshop zur Argumentanalyse, geleitet von Christian Stab und Carsten Schnober (UKP Lab, TU Darmstadt). Der Workshop hatte einen explorativen Charakter. Im Zentrum standen die unterschiedlichen Klassifikations- und Identifikationsverfahren zur automatisierten Erkennung von Argumenten und Argumentationsstrukturen. Dieser Workshop zeigte besonders deutlich die enge Kooperation auf informatorischer und informationswissenschaftlicher Ebene. Der Mitorganisator des Workshops, Carsten Schnober, war sowohl im Projektteam von Frau Professor Dr. Iryna Gurevych am UKP als auch am DIFP tätig. Er widmete sich in seiner Dissertation den für das Projekt wichtigen Algorithmen hinter dem Topic Modeling (zusammen mit Dr. Richard Eckert de Castilho) und überwachte die Topic-Model-Generierung.

Ebenso baute er zusammen mit Ben Heuwing von der Universität Hildesheim die Solr-Plattform auf. Hierfür arbeiteten die Projektmitglieder aus Darmstadt der Implementation an einer Schnittstelle zum Einlesen der „Welt-der-Kinder“-Textdaten (der Bestand aus GEI-Digital) sowie der zugehörigen Metadaten für die weitere Verarbeitung. Dabei wurden einzelne Seiten als Dokumente interpretiert und mit den auf Buchebene vorhandenen Metadaten annotiert. Segmentierungsinformationen zu Kapitelgrenzen, Umschlagseiten, Inhaltsverzeichnissen, Abbildungen und Tabellen wurden aus den Metadaten extrahiert und zu den Annotationen auf Seitenebene hinzugefügt. Die Metadaten lagen zunächst nicht für alle Einträge vollständig vor und wurden durch das GEI nachträglich manuell schrittweise ergänzt. Um die neue Daten den zugehörigen Büchern zuzuordnen und mit den bereits vorhandenen Daten zu kombinieren, wurde ein entsprechender Reader implementiert. Durch (semi-)automatische Korrekturen wurde versucht, möglichst viele OCR-Fehler automatisch zu erkennen und zu korrigieren.

22 http://wdk.ukp.informatik.tu-darmstadt.de/solr/WdK.dev/browse, zuletzt geprüft am 25. Juli 2018.

Ebenso erstellten die Darmstädter Projektpartner einen Prototyp eines „Corpus Explorer zur Darstellung und Nutzbarmachung sämtlicher Verarbeitungsprozesse“, den sog. „Welt der Kinder Explorer“. Mit ihm wurden auch die für die Erforschung historischer Fragestellung automatisch erhobenen statistischen Erkenntnisse visualisiert sowie die Nutzer bei der Suche nach relevanten Textstellen auf Grundlage der automatischen Modellierungen unterstützt. Mit Solritas3 gibt es eine grundlegende Web-basierte Benutzeroberfläche für einen SOLR-Suchindex. Diese wurde in Zusammenarbeit mit dem IWIST an die Welt der Kinder-Daten und -anforderungen angepasst.[23]

Qualifikationsarbeiten, die im Zusammenhang mit dem Vorhaben entstanden sind oder entstehen

GEI Braunschweig:
Dissertation: Maik Fiedler: Weltwissen in deutschen Schulbüchern des 19. Jahrhunderts (betreut von Simone Lässig), In Arbeit.

Universität Hildesheim:
Dissertation: Ben Heuwing: Usability-Ergebnisse als Wissensressource in Organisationen
Masterarbeit: Anastasia Christoforidis (betreut von Ben Heuwing und Thomas Mandl): Visualisierung von Topic-Korrelationen, Universität Hildesheim, 2016
Bachelorarbeit: David Wodausch (betreut von Ben Heuwing und Thomas Mandl): Kollokationsanalyse (2016)
Projekt-Arbeit: David Wodausch (supervised by Ben Heuwing, Thomas Mandl) Kollokationsanalyse, Universität Hildesheim, 2016

TU Darmstadt:
Dissertation: Carsten Schnober [Natural Language Processing] (betreut von Iryna Gurevych). In Arbeit.
Bachelorarbeiten: Ute Winchenbach (betreut von Carsten Schnober, Erik-Lân Do Dinh, Iryna Gurevych): Opinion Mining on Historical German Textbooks (2015).
Masterarbeit: Deborah Buth (betreut von Carsten Schnober, Erik-Lân Do Dinh, Iryna Gurevych): Corpus-Based OCR Post-Processing of German Historical Texts (2015)

23 http://wdk.gei.de/, zuletzt geprüft am 25. Juli 2018.

Publikationsliste

Sammelbände/Themenhefte (peer reviewed):

- Lässig, Simone; Weiß, Andreas (Hrsg.): “Children, Knowledge, and the World in Germany: Reconsidering Cultural and Media Transformations in an Age of Nationalization and Globalization”. Berghahn Books Oxford/New York [beim Verlag, erscheint 2019]
- Weiß, Andreas: Themenheft „Weltwissen und der außereuropäische Raum: Geographieschulbücher und Kinderbücher des 19. Jahrhunderts im internationalen Vergleich“. JEMMS. Journal of Educational Media, Memory, and Society 10 (2018), 1.

Aufsätze/Artikel:

im Erscheinen:

- Heuwing, Ben; Weiß, Andreas: Suche und Analyse in großen Textsammlungen: Neue Werkzeuge und Benutzeroberfläche für die Schulbuchforschung, in: Eckert.Expertise [Erscheint 2018]
- Nieländer, Maret; Weiß, Andreas: Schönere Daten – Nachnutzung und Aufbereitung für die Verwendung in Digital-Humanities-Projekten, in: Eckert.Expertise [Erscheint 2018].
- Lässig, Simone/Weiß, Andreas: Children, Knowledge, and the World in Germany. Reconsidering Cultural and Media Transformations in an Age of Nationalization and Globalization (Introduction); dies. (Hrsg.) in: “Children, Knowledge and the World in Germany”. Berghahn Books Oxford/New York
- Weiß, Andreas: The World at War in German textbooks: Wars as a national projected area of global expansion?, in: Lässig, Simone; Weiß, Andreas (Hrsg.): “World of Children”. Berghahn Books

Erschienen:

- Weiß, Andreas: Introduction: World Knowledge and Non-European Space Nineteenth-Century Geography Textbooks and Children’s Books, in: Weiß, Andreas (Hrsg.): „Weltwissen und der außereuropäische Raum: Geographieschulbücher und Kinderbücher des 19. Jahrhunderts im internationalen Vergleich“. JEMMS. Journal of Educational Media, Memory, and Society 10 (2018), 1, S. 1-9.
- Weiß, Andreas: Reading East Asia in Schools of the Wilhelmine Empire, in: Weiß, Andreas (Hrsg.): „Weltwissen und der außereuropäische Raum: Geographieschulbücher und Kinderbücher des 19. Jahrhunderts im internationalen Vergleich“. JEMMS. Journal of Educational Media, Memory, and Society 10 (2018), 1, S. 10-27.
- Wodausch, David und Maik Fiedler, Ben Heuwing, Thomas Mandl: „Hinterlistig – schelmisch – treulos – Sentiment Analyse in Texten des 19. Jahrhunderts: Eine exemplarische Analyse für Länder und Ethnien“, in: DHd 2018: Kritik der digitalen Vernunft. Konferenzabstracts, Georg Vogeler (Hg.), Köln: Universität zu Köln, 2018, 223-226.
- Fiedler, Maik: Wissensgeschichte aus dem Schulbuch – “Mixed Analysis” oder Diskurs 2.0?, in: Rath, Imke (Hrsg.): Methoden und Theorien der Bildungsmedien- und Bildungsforschung: Ein Werkstattbericht von Nachwuchswissenschaftlerinnen und -wissenschaftlern des Georg-Eckert-Instituts (Eckert. Dossiers 14 (2017)), S. 44-64.
- Christoforidis, Anastasia, Heuwing, Ben, Mandl, Thomas: Visualising Topics in Document Collections: An analysis of the interpretation processes of historians. Everything changes, everything stays the same? Understanding Information Spaces – Proceedings of the 15th International Symposium of Information Science, Schriften zur Informationswissenschaft. Berlin 2017.

- Fechner, Martin; Weiß, Andreas: Einsatz von Topic Modeling in den Geschichtswissenschaften: Wissensbestände des 19. Jahrhunderts, in: Zeitschrift für digitale Geschichtswissenschaft (2017).
- Weiß, Andreas: Der Kolonialkrieg im deutschen Schulbuch des Kaiserreiches, in: Portal Militärgeschichte, 21. November 2016.
- Heuwing, Ben; Mandl, Thomas; Womser-Hacker, Christa: "Combining contextual interviews and participative design to define requirements for text analysis of historical media", in: ISIC: The Information Behaviour Conference. Zadar, 2016.
- dies.: "Methods for User-Centered Design and Evaluation of Text Analysis Tools in a Digital History Project", in: 2016 Annual Meeting of the Association for Information Science and Technology. Kopenhagen, 2016.
- Heuwing, Ben, Projekt Welt der Kinder – Überblick über die informationswissenschaftliche Bedarfsanalyse, HiER 2015 – Proceedings des 9. Hildesheimer Evaluierungs- und Retrievalworkshop. Hildesheim : Universitätsverlag, 2015.
- Heuwing, Ben and Christa Womser-Hacker: Zwischen Beobachtung und Partizipation – nutzerzentrierte Methoden für eine Bedarfsanalyse in der digitalen Geschichtswissenschaft, in: Information – Wissenschaft & Praxis, Bd. 66 (2015) Nr. 5-6, S. 335–344.
- Schnober, Carsten, and Iryna Gurevych. "Combining Topic Models for Corpus Exploration: Applying LDA for Complex Corpus Research Tasks in a Digital Humanities Project", in: Proceedings of the 2015 Workshop on Topic Models: Post-Processing and Applications, TM '15. New York, NY, USA: ACM, 2015, S. 11–20.
- Strötgen, Robert: New information infrastructures for textbook research at the Georg Eckert Institute, in: History of Education & Children's Literature 9 (2014), 1, S. 149-162.

weitere Publikationen:
- Weiß, Andreas; Otto, Marcus: Themenheft „Affekkontrolle", Body Politics: Zeitschrift für Körpergeschichte 5 (2017), 8.
- mit Marcus Otto: Einleitung: Affekte und Affektkontrolle in der Moderne, in: Otto, Marcus; Weiß, Andreas (Hrsg.): Affektkontrolle. Body Politics: Zeitschrift für Körpergeschichte 5 (2017), 8, S. 5-30.
- ASEAN, in: Droit, Emmanuel; Hansen, Jan; Reichherzer, Frank (Hrsg.): Den Kalten Krieg vermessen: Über Reichweite und Alternativen einer binären Ordnungsvorstellung. Berlin: de Gruyter Oldenbourg, 2018, S. 33-44.
- Fiedler, Maik; Weiß, Andreas: Tagungsbericht Von Daten zu Erkenntnissen: Digitale Geisteswissenschaften als Mittler zwischen Information und Interpretation. DHd-Jahrestagung 2015, 23.02.2015 - 27.02.2015, Graz, in: H-Soz-Kult, 06.07.2015; 015, 23.02.2015 – 27.02.2015 Graz, in: H-Soz-Kult, 06.07.2015.
- Weiß, Andreas: Rezension zu: Faure, Romain: Netzwerke der Kulturdiplomatie. Die internationale Schulbuchrevision in Europa, 1945-1989. Berlin, Boston: De Gruyter Oldenbourg, 2015, in: Historische Zeitschrift 304 (2017), 1, S. 295-296.
- Weiß, Andreas: Rezension zu: Erlin, Matt; Tatlock, Lynne (Hrsg.): Distant Readings. Topologies of German Culture in the Long Nineteenth Century. Rochester, NY 2014 , in: H-Soz-Kult, 04.01.2016.

Liste möglicher Pressemitteilungen und Medienberichte (Auswahl)

2017:

Abschlusskonferenz DIGIMET Quellen und Methoden der Geschichtswissenschaft im digitalen Zeitalter - Neue Zugänge für eine etablierte Disziplin? - DIGIMET 2017; zusammen mit dem DHI Washington und dem Verbandes der Historiker und Historikerinnen Deutschlands (VHD).

Beitrag Deutschlandradio
https://www.deutschlandfunkkultur.de/geschichtswissenschaft-im-digitalen-zeitalter-vielleicht.976.de.html?dram:article_id=396905

Videodokumentation der Gerda Henkel Stiftung
Panel 1
https://lisa.gerda-henkel-stiftung.de/digimet_2017_digitale_quellen_digitale_werkzeuge_und_die_notwendigkeit_der_erweiterung_der_historischen_quellenkritik?nav_id=7352

Panel 2
https://lisa.gerda-henkel-stiftung.de/digimet_2017_digital_born_sources_als_herausforderung_fuer_die_zeitgeschichte?nav_id=7328

Panel 3
https://lisa.gerda-henkel-stiftung.de/digimet_2017_lehre_und_ausbildung_der_geschichtswissenschaft_im_digitalen_zeitalter?nav_id=7348

Panel 4
https://lisa.gerda-henkel-stiftung.de/digimet_2017_digitale_infrastrukturen_finanzierung_und_rechtliche_bedingungen?nav_id=7350

Panel 5
https://lisa.gerda-henkel-stiftung.de/digimet_2017_neue_arbeitsweisen_und_die_herausforderungen_der_interdisziplinaritaet?nav_id=7351

Abschlussdiskussion
https://lisa.gerda-henkel-stiftung.de/digimet_2017_quellen_und_methoden_der_geschichtswissenschaft_im_digitalen_zeitalter_neue_zugaenge_fuer_eine_etablierte_disziplin?nav_id=7364

https://twitter.com/GeorgEckert/status/912275973877456896

2016:

https://trendingdeutschland.com/hashtag/dhd2016
Melanie S. @msiemund, 12 Apr 2016: „RT @DARIAHde: So war's: Tagungsberichte der #DHd2016 ReisestipendiatInnen jetzt online https://t.co/H4FESo9pj5 @CLARIN_D @DARIAHde @DHDInfo“
Berichte ReisestipendiatInnen hier: http://dhd-blog.org/?p=6517;
Vortrag Fiedler/ Weiß/ Heuwing/ Schnober „Automatische Textanalysen in der Geisteswissenschaft – Auswertung, Interpretation und Relevanz“:

http://www.dhd2016.de/abstracts/vortr%C3%A4ge-006.html (retweets von DH @ Uni Bern @DH_unibe; sabine seifert @sabine_seifert; prometheus e.V. @prometheus_eV)

https://twitter.com/hashtag/dhd2016?src=hash
Michael Piotrowski @true_mxp, 9 Mar 2016: „Es geht doch um sehr spezielle Werkzeuge für Wissenschaftler, nicht um kommerzielle Dienste, die Nutzer gewinnen müssen 8-O#dhd2016 #v2b“ (?)

http://digihum.de/2016/03/workshop-wissenschaftsgeschichte-und-digital-humanities-in-forschung-und-lehre-07-04-bis-09-04-2016-in-goettingen/

https://twitter.com/search?q=%22welt%20der%20kinder%22&src=typd
AndreasRottensteiner @AndreasRottenst 11. Juli: "Welt der Kinder im 19. Jh" Schulbücher prägen das spätere Weltbild? http://science.orf.at/stories/2784290/… Ich befürchte das Schlimmste für die Gegenwart! - bezieht sich auf Berichterstattung im ORF: http://science.orf.at/stories/2784290
Sabine Scherz @SabineScherz 11. Juli: Mit #BigData der Weltsicht von Schulkindern auf der Spur. "Welt der Kinder" im 19. Jahrhundert" http://science.orf.at/stories/2784290/ … @GerhardLauer
GHI Washington @GHIWashington 15. Jan.: "Die Welt der Kinder" International Conference at the Universität Zürich from February 4-6, 2016 (http://goo.gl/MNavSM)

2015:
Dhd 2015:
Präsentation Welt der Kinder – Geisteswissenschaftliches Asset Management System der Uni Graz: http://gams.uni-graz.at/o:dhd2015.v.055
https://trendingdeutschland.com/tweep/maikfiedler2.html
https://twitter.com/hashtag/dhd2015
https://twitter.com/justtherealmaik

Max Weber Stiftung @webertweets 26. März 2015: Call for Papers: Die Welt der Kinder. Weltwissen und Weltdeutung in Schul- und Kinderbüchern des 19. und frühen 2... http://owl.li/2WvEUa
H-Soz-Kult @hsozkult 25. März 2015: CFP: Die Welt der Kinder. Weltwissen und Weltdeutung in Schul- und Kinderbüchern des 19. und frühen 20. Jahrhunderts http://dlvr.it/96XySq
Stefan Dumont @stefandumont 28. Feb. 2015 : @FredFirlefanz @textarchiv Mein Tweet bezog sich auf die "Welt der Kinder" http://www.gei.de/forschung/europa/welt-der-kinder.html; https://twitter.com/stefandumont/status/571680355666800640

2014:
Christof Schöch @christof77 8. Dez. 2014: "Welt der Kinder" doing Topic Modeling, Opinion Mining / Sentiment Analysis with German textbooks for children. #nedimah
Christof Schöch @christof77 8. Dez. 2014: Now, Carsten Schnober ab."Welt der Kinder". Data av. here: http://gei-digital.gei.de/viewer/ (page-wise scans + raw OCR full text) #nedimah
jugendhilfeportal.de @_fkp_ 30. Apr. 2014 : Projektstart des GEI: „Die Welt der Kinder“ http://bit.ly/1lzbci7
bildungsklick @bildungsklick 29. Apr. 2014: Projektstart: "Die Welt der Kinder". Weltwissen und Weltdeutung in Schul- und Kinderbüchern zwischen 1850 und 1918 http://bikl.de/R91178