ANNE PURSCHWITZ

Netzwerke des Wissens - Thematische und personelle Relationen innerhalb der halleschen Zeitungen und Zeitschriften der Aufklärungsepoche (1688-1818)

Journal of Historical Network Research x (2018) 109-142.

Keywords

Topic-Modelling, Journals, Enlightenment, Discourse, communicative networks

Abstract

Given that the amount of available digital sources is constantly growing, studies on the acquisition and transmission of knowledge in the public sphere tend to be methodologically selective in their analysis of historical discourses. These studies, as a rule, tend to offer a very partial image from an intellectual history perspective of the production, representation, reception, and evaluation of knowledge. A more productive approach would be to analyse the ways in which pre-modern knowledge production functioned based on the broadest possible sample of sources. The project will seek, with a computer-based methodology, to consistently compile and systematically analyse all 356 journals and periodicals published in Halle between 1688 and 1815, as well as to gather information on all the actors participating in the process (authors, editors, printers et cetera). The periodical press of the Enlightenment is particularly relevant for this type of approach. Newspapers and periodicals allowed for a broader circulation of the ideas of the educated and for faster reactions to news, controversies, and publications, as they became a widely accepted form of interpersonal

C²DH LUXEMBOURG CENTRE FOR CONTEMPORARY AND DIGITAL HISTORY

UNIVERSITÉ DU LUXEMBOURG

communication. They thus filled a public and institutional gap and provided an open and adaptable medium for various public discourses without thematic limitations. We qualitatively go beyond standard discourse analysis by using a wealth of text and network data digitally available to reconstruct the breadth and depth of all the types of discourses of the Enlightenment, the combined quantitative methods for text and network analysis will allow us to trace the impact of the many involved parties right into the structure of the discourses.

1 Einleitung*

Anhand einer Beispielstudie möchte unser Projekt: 'Gesellschaftliche Wissensproduktion in der Aufklärung – Text- und netzwerkanalytische Diskursrekonstruktion. Die halleschen Zeitungen und Zeitschriften (1688-1815)' dazu beitragen, methodische Grundlagen zu entwickeln, um Diskursthemen zu identifizieren, Diskursverläufe miteinander zu vergleichen und, wenn möglich, für die Aufklärungsepoche Diskursmodelle zu entwickeln, die zum einen die Verortung einzelner Kontroversen in einem Gesamtspektrum ermöglichen und zum anderen eine Antwort auf die Frage bieten können, welche Themen im Betrachtungszeitraum von Relevanz waren. Mit einer solch umfassenden Herangehensweise und durch die Heranziehung einer Vielzahl von Dokumenten sollen den Fragen nach der gesellschaftlichen Reichweite und den thematischen Schwerpunkten der Aufklärung neue Impulse gegeben werden. Der Fokus des Projektes liegt in der Entwicklung einer Methode, die in der Lage sein wird mediale Quellenbestände in Korrelation mit strukturellen Rahmenbedingungen auf eine bestimmte Frage und deren Entwicklung hin zu untersuchen. In einem zweiten Schritt sollen unter Einbeziehung von Metadaten thematische, personelle und institutionelle Netzwerke rekonstruiert, visualisiert und für eine inhaltliche Interpretation aufbereitet werden. Durch die Etablierung von Modellen und Kennzahlen für Diskurse wird es dann möglich sein, unterschiedliche Diskurse miteinander in Beziehung zu setzen und zu vergleichen. Zu diesem Zweck gilt es, innerhalb der für den Zeitraum 1688–1815 nahezu vollständig erschlossenen periodischen Presselandschaft der Stadt Halle diskursive Ordnungsmuster zu identifizieren, aber auch die entsprechenden Rahmenbedingungen zu

---

* **Acknowledgments**: This article is based on a paper read at the *12th Historical Network Research Workshop* „Historische Netzwerkforschung. Kommunikation in Netzwerken – Netzwerke der Kommunikation. Thematische, methodische und theoretische Perspektiven historischer Netzwerkanalyse zwischen Geschichts- und Kommunikationswissenschaft" convened by Matthias Bixler and Erik Koenen and held from 20-21 April 2018 at the University of Bremen.
Corresponding Author: Dr Anne Purschwitz, Chair of Economic and Social History / IZEA, Martin Luther University Halle-Wittenberg, anne.purschwitz@izea.uni-halle.de

analysieren, die die lokale Diskurslandschaft prägten. Das bedeutet, dass neben Texten auch die Ebene der personellen Akteure, der Institutionen und Medien entsprechend zu berücksichtigen ist. Erst auf diese integrative Weise lässt sich Wissensgeschichte im Sinne Sarasins als Gesellschaftsgeschichte definieren.[1]

Im Folgenden sollen die Möglichkeiten einer Kombination von Topic-Modeling und Netzwerkanalyse an einem ausgewählten Beispiel überprüft und der Mehrwert einer solchen methodischen Koppelung gegenüber einer singulären Anwendung des jeweiligen Verfahrens dargestellt werden.

2 Zielsetzung

Ziel des Projektes ist es, für ein Zentrum der Aufklärung in Deutschland Entwicklungsfaktoren, Dynamik und Funktionsweise von Wissenssystemen zu ergründen. Unser Projekt möchte die Frage beantworten, wie, wann und (möglichst auch) warum während der Aufklärungsepoche ein bestimmtes Wissen in Form einer öffentlichen Thematisierung gesammelt, verbreitet und genutzt beziehungsweise ebenso unterdrückt, vernachlässigt oder ignoriert wurde. Im Vordergrund steht zunächst die Aufgabe, inhaltliche Beziehungen zwischen Texten herauszuarbeiten, ohne sie in ihrer Gänze manuell lesen und bearbeiten zu müssen, um mithilfe dieser Bezüge thematische Verknüpfungen und daran anschließend Diskurse zu rekonstruieren.

Während der Aufklärung übernahm insbesondere die Presse stetig umfassendere Funktionen bei der Wissensvermittlung, was zur Generierung eines Mediums führte, das entsprechend schnell auf immer häufiger zu registrierende Neuerungen, Kontroversen und Publikationen reagieren und sich zugleich zu einer weithin akzeptierten Form der interpersonellen Kommunikation entwickeln konnte.

Die damit einhergehende Expansion und Veränderung der öffentlichen Kommunikation und des öffentlichen Raumes führten zugleich zu einer stärkeren Fragmentierung von Inhalten, Akteuren und Medien. Und dennoch war gerade die periodische Presse als Gruppenphänomen in der Lage die Kommunikationspartizipation zu erweitern; sie verband individuelle Kommunikationsstränge und galt zugleich als lokaler Spiegel der gelehrten Welt.[2] Ihre spezifischen Charakteristika ermöglichten jedem hinreichend Lesefähigen, am Wissensdiskurs teilzunehmen – sie bildeten eine, wenn nicht die entscheidende Arena für öffentliche Diskurse, aus diesem Grund stehen sie im Zentrum unseres Projekts.

---

1 Sarasin „Wissensgeschichte“, S. 171.

2 Gierl, „Kompilation“, S. 72.

Um ein möglichst umfassendes und dichtes Bild der Wissenskultur im Raum Halle während der Aufklärung zu erhalten, bietet es sich an, einen diskursgeschichtlichen Ansatz zu wählen, da nur so das gesamte Spektrum von Wissen und den damit verknüpften Akteuren Berücksichtigung erfahren kann. Anknüpfend an Jäger wird unter Diskurs ein Fluss von Wissen durch die Zeit verstanden,[3] wobei in der kritischen Diskursanalyse auch Ansätze der Interdiskursanalyse [4] Berücksichtigung erfahren. Die Anwendbarkeit der Diskursanalyse auf ein sehr umfangreiches Quellenkorpus wie periodisch erscheinende Presseerzeugnisse es bilden, stellt zugleich die größte methodische Schwierigkeit dar. Doch nur durch eine solche Betrachtung kann ein Blick auf die Medialität und Prozessualität von Wissen und auf Journale als intermediäre Vermittler von Wissen eröffnet werden.

Für die konkrete Herangehensweise des Projekts bedeutet das zunächst zu untersuchen, welche Themen in welchen Journalen, wann, mit welchen inhaltlichen Präferenzen und in welcher Intensität diskutiert wurden. Die mithilfe eines Topic-Modelling generierten Themen können jedoch erst durch zusätzliche Verfahren mit ergänzenden Informationen in Beziehung gesetzt werden, so zum Beispiel den an den Themen beteiligten Akteuren, Journalen, strukturellen Rahmenbedingungen et cetera. In dieser Hinsicht eröffnet gerade die Rekonstruktion von Netzwerken in Korrelation mit Themen und weiterführenden (Meta)Daten nahezu unbegrenzte Möglichkeiten. Da sie beispielsweise durch die Verbindung mit Informationen zu Publikum, fachlicher Verortung (der Akteure und/oder Journale), Mitarbeit von Kupferstechern, Verlegern, Druckern bis zum Preis und Verbreitungsgebiet der Journale die modellierten Themen in ihrer Ausprägung näher bestimmen, ihre Genese interpretativ unterstützen und einen detaillierten Vergleich zwischen unterschiedlichen Themen ermöglichen kann. Erst durch eine solche Verknüpfung wird ersichtlich (und möglichst auch visualisiert) werden, welche Themen wann, in welchen Medien, unter Einbeziehung welcher Akteure und vor dem Hintergrund welcher Rahmenbedingungen, im Vordergrund standen. Daraus können dann unterschiedliche Aspekte zu Fragen der Ähnlichkeit von Diskursen (wie beispielsweise Thema, Zeitpunkt, Intensität, Frames und Diskurskonstellationen), [5] der Vernetzung von Diskursen[6] oder der Definition des öffentlichen Raumes als entscheidender Voraussetzung für Wissenszirkulation[7] abgeleitet werden.

---

3 Vgl. Jäger; Jäger, Deutungskämpfe, S. 13-37.
4 Vgl. Link, „Diskurs".
5 Kleinen-von Königslöw, Arenen-Integration, S. 51-58.
6 Benkler et al., „Social Mobilization".
7 Dahlgren, „The Internet" und Papacharissi, „A privat sphere".

3 Datengrundlage

Die Grundlage für dieses Vorhaben bilden die 360 zwischen 1688 und 1815 in Halle erschienenen Zeitungen und Zeitschriften und die damit in Beziehung stehenden persönlichen Relationen der an ihnen, auf unterschiedlichen Ebenen, beteiligten Akteure. Aktuell sind 290 Zeitschriften vollständig durch ihre Inhaltsverzeichnisse erfasst und die für die Erschließung entwickelte Datenbank weist mehr als 30.000 Beitragsverzeichnungen auf; sie umfasst rund 1.800 Personeneinträge, von denen etwa 900 eindeutig identifizierbar sind.[8] Dieses grundlegende Korpus wird beständig ergänzt und soll letzten Endes alle Beiträge der halleschen Zeitschriften und (wenn auch mit methodischen Schwierigkeiten verbunden) Zeitungen enthalten. Die Artikel sind auf unterschiedlichen Ebenen und unter Nutzung von Mehrfachzuweisungen mit Personen verknüpft. Die Kombinationen zwischen Person und Zeitschrift können durchaus vielfältig sein. Berücksichtigung erfahren momentan Verleger-, Herausgeber-, Autoren- und Rezensentenbeziehungen. Die dafür erforderlichen Informationen stammen aus den Zeitschriften selbst, wurden zudem aber durch prosopographische Recherchen ergänzt. Die Fülle von Verknüpfungen zwischen Zeitschriften und Personen offenbart bereits, dass gerade die periodische Presse der Aufklärungsepoche als ein engmaschiges Netzwerk verstanden werden muss, in dem nicht nur einzelne Journale, meist durch personelle Querverbindungen, miteinander in Bezug standen. Vielmehr zeitigen persönliche Kontakte und Verflechtungen für die Wissensvermittlung und -generierung der Aufklärungsepoche eine große Relevanz und bewirken durchaus Wechselwirkungen mit den – beispielsweise in der periodischen Presse auffindbaren – Kontroversen. Durch die Rekonstruktion von multimodalen Netzwerken können, neben den 'klassischen' persönlichen und inhaltlichen Verknüpfungen, zudem beispielsweise die Adressierung von Texten sowie die Kommunikationsrichtung Berücksichtigung erfahren, gleichzeitig aber auch die Journale selbst als Akteure des Wissens in den Blick genommen werden.

Eine grundlegende Schwierigkeit beinhalten Journalbeiträge, die nur mit Kürzeln, Initialen oder Chiffren versehen sind. In diesen Fällen wird zunächst versucht, die Autoren zu identifizieren. Dafür wurde am Institut für Informatik der Martin-Luther-Universität Halle-Wittenberg ein Tool entwickelt, das in der Lage ist, automatische Abfragen an eine eigens dafür erstellte 'Pseudonymdatenbank' zu richten. Die darin enthaltenen

---

8 'Eindeutig identifizierbar' bedeutet in diesem Zusammenhang, dass eine Person mit Vornamen und Nachnamen bezeichnet werden kann, nicht zwingend beinhaltet sind an dieser Stelle Lebensdaten und eine Rekonstruktion ihrer Existenz, zumal zahlreiche Personen kaum Spuren hinterlassen haben.

Informationen entstammen zum Teil unterschiedlichen Pseudonymlexika,[9] aber ebenso den häufig am Ende eines Jahrgangs in den Periodika selbst verzeichneten Autoreninformationen und werden fortlaufend ergänzt.[10] Das Verfahren selbst sucht durch einen Abgleich zwischen den unterschiedlichen Datenbanken nicht bloß nach vollständigen Übereinstimmungen, sondern berücksichtigt auch Ähnlichkeiten und Wahrscheinlichkeiten, so zum Beispiel bei variierenden Namensschreibweisen. Ebenso können die teilweise in den Datenbanken verfügbaren Lebensdaten für eine genauere Zuordnung mit einbezogen werden. Nur im Idealfall, das heißt bei einer genauen Übereinstimmung von Nachnamen, mindestens zwei Vornamen und einer Kongruenz der Lebensdaten mit dem Zeitpunkt der Publikation wird ein Autor eindeutig identifiziert (ca. 50 Personen). In der Mehrzahl der Fälle kann hingegen ein Personenkreis bestimmt werden, der dann durch die (nicht-automatisierte) Bestimmung thematischer Schwerpunkte beziehungsweise eine Fachzuordnung der Autoren weiter eingegrenzt werden muss (bisher bei ca. 200 Personen).

Ein weiteres Problem der Periodika der Aufklärungsepoche sind die zahlreichen anonym erschienenen Beiträge. In diesen Fällen kann die Frage nach persönlichen Bezügen nicht beantwortet werden. Die Suche nach Beziehungen zwischen den Texten erhält hier einen deutlich höheren Stellenwert, denn auch die veröffentlichten Artikel, Rezensionen, Anmerkungen und Abhandlungen bilden ein Netzwerk; dies nicht länger nur auf Grundlage persönlicher Relationen zwischen Verlegern, Herausgebern oder Autoren, sondern vielmehr nehmen die Texte aufeinander Bezug oder thematisieren ähnliche Gegenstände in anderen Zusammenhängen und Medien. Diese Bezüge können mithilfe der Topic-Modellierung aufgefunden werden.

4 Topic-Modeling

Unter dem Schlagwort distant reading werden seit einigen Jahren neue Möglichkeiten diskutiert, um in der Menge der (zunehmend auch digital verfügbaren) Texte durch die Anwendung algorithmenbasierter Verfahren intertextuelle Zusammenhänge zu erkennen. Hierzu gehören beispielsweise Text Mining, Clustering oder Topic-Modeling.[11]

---

9 Bisher bearbeitet wurden zunächst: Weller, Lexicon pseudonymorum; Eymer, Pseudonymen-Lexikon; Holzmann, Pseudonymen-Lexikon und in Auszügen Parthey, Mitarbeiter.

10 Unveröffentlichte Bachelorarbeit: Mladenova, Automatisierung.

11 Vgl. dazu: Crane, „What Do You Do“.

In Hinblick auf die automatische Sprachverarbeitung müssen dabei zwei Ebenen voneinander unterschieden werden, zum einen die sprachlich-stilistische und zum anderen die inhaltlich-thematische Analyse. Im Folgenden liegt der Fokus auf den Möglichkeiten einer inhaltlich-thematischen Strukturierung eines Textkorpus. Das dafür nutzbare Topic-Modeling bezeichnet eine quantitative Methode der Textanalyse, die auf ein Korpus von Dokumenten angewendet werden kann und dabei die Wortverteilung innerhalb dieser Textsammlung statistisch ermittelt. Ziel ist das Aufdecken 'verborgener' semantischer Strukturen, ohne dafür explizit über semantisches Wissen verfügen zu müssen.[12] Topic-Modeling folgt dabei der Hypothese, dass Wörter, die in demselben Kontext vorkommen, dazu tendieren, eine ähnliche Bedeutung zu besitzen. Dadurch könnten grundlegend wiederkehrende Themen, Motive und unter Hinzuziehung weiterer Daten auch Diskurse identifiziert werden. Die Topics (Themen) helfen somit Komplexität zu strukturieren.[13]

Zunächst erfolgt die Modellierung eines Standard-LDA-Topic-Models.[14] Dieses bietet im Unterschied zu Clustering-Verfahren[15] den Vorteil, dass die Wörter in einem Dokument einzeln den Topics zugeordnet werden können. Topic-Modelle betrachten ein Dokument somit als eine Mischung von Topics. Eine Textanalyse mithilfe von Topic-Modellen kann als Knowledge-Discovery-Prozess beschrieben werden und folgt den Schritten 1. Datenauswahl, 2. Datenvorverarbeitung und Datentransformation, 3. Datenanalyse und 4. Evaluation des Modells.[16] Die Topics selbst basieren auf der Verteilung von Wahrscheinlichkeiten von Wörtern, die jeweiligen Dokumente hingegen auf der Verteilung von Wahrscheinlichkeiten von Topics. Vorgegeben sind somit die in den Dokumenten enthaltenen Wörter, während die 'versteckten' Topics in einem iterativen Prozess aufgedeckt werden. Dieses rein mathematische Verfahren funktioniert sprachunabhängig und kann durch die Herstellung von inhaltlichen Zusammenhängen zwischen Wörtern und Dokumenten den Bedeutungsgehalt eines Textes erschließen. Um Topic-Modeling möglichst zielführend einsetzen zu können, wird zunächst ein (möglichst geschlossenes beziehungsweise eingrenzbares) Korpus benötigt. Im Unterschied zu zahlreichen anderen informatischen Analyseverfahren bedarf das Topic-Modeling jedoch keiner Metadaten, die die Topics inhaltlich vorab definieren.

---

12 Vgl. dazu: Lim et al., „Nonparametic"; Mimno und Blei, „Bayesian Checking"; Blei, „Probabilistic Topic Models"; Blei und Lafferty, „Dynamic Topic Models".

13 Zum Begriff ‚topic' vgl.: Jacobi et al. „Quantitative analysis".

14 Einen guten Überblick dazu bieten: Anthes, „Topic models" und Sievert und Shirley, „LDAvis".

15 Vgl. dazu bspw.: Stier et al. „Election Campaigning".

16 Vgl dazu ausführlich: Papilloud und Hinneburg, Qualitative Textanalyse, S. 28-36.

Das Haupthindernis in Hinblick auf historische, in Fraktur gedruckte Texte besteht in der Tatsache, dass diese nicht maschinenlesbar beziehungsweise erst durch mühsame Verfahren aufbereitet werden müssen, aus diesem Grund können Natural-Language-Prozesse nicht problemlos ansetzen. Digitalisierungsverfahren, die historische Texte in PDF- bzw. Bilddateien umwandeln, bieten hier zunächst keinen Mehrwert, da sie mit automatisierten Verfahren nicht lesbar sind.[17]

In unserem Projekt nutzen wir den, von Alexander Hinneburg an der Martin-Luther Universität Halle-Wittenberg entwickelten TopicExplorer[18]. TopicExplorer nutzt zurzeit die LDA-Implementierung von Mallet[19]. Die genutzte Parametrisierung folgt keinem festen Hyperparameter für LDA, sondern optimiert diese in regelmäßigen Abständen während des Trainings des Topic-Modells. Die Nutzung des TopicExplorers wird aktuell in einem permanenten Wechselspiel, an die Spezifika historischer Dokumente, und die damit verbundenen Schwierigkeiten, angepasst.

Da bisher nur für einen Teil der betreffenden Journale OCR-Volltexte vorliegen, erfolgten die ersten Topic-Modellierungen im Projekt ausschließlich anhand der Artikelüberschriften. Die zu untersuchende Haupteinheit der Zeitungen und Zeitschriften bilden somit die einzeln erfassten, vollständigen bibliographischen Angaben der in ihnen beinhalteten Artikel. In einem zweiten Schritt soll die Funktionalität dieses Vorgehens durch exemplarische Versuche mit Volltexten validiert werden. In den meisten Fällen sind die Titelinformationen jedoch durchaus umfangreich und beinhalten beispielsweise bei Rezensionen auch den Titel und Verfasser des rezensierten Werkes, in anderen Fällen, wie den biographischen Nachrichten, werden die Artikelüberschriften hingegen erst verständlich, wenn sie mit dem jeweiligen Zeitschriftentitel in Bezug gesetzt werden. Um diese methodische Schwierigkeit zunächst zu umgehen, wurde mit einem Datensatz gearbeitet, der aktuell sehr kurze Titel ausschließt. Die daraus resultierende ähnliche Dokumentenlänge kommt der Funktionalität der Topic-Modellierung entgegen. Schwieriger könnte sich die Situation gestalten, wenn die Volltexte

---

17 Die bisher zur Verfügung stehende OCR-Software erfordert aktuell noch zahlreiche Vorverarbeitungen, Korrekturen oder Trainings, sodass sie auf ein so heterogenes Korpus, wie es der Bestand unterschiedlicher Zeitschriften darstellt, kaum angewendet werden kann (vgl. Breuel, „The OCR"; Springmann, „OCR" und Stollwerk, „Machbarkeitsstudien").

18 Vgl. Hinneburg et al, „TopicExplorer"; Hinneburg et al., „Exploring Document Collections" und Röder et al., „Exploring the Space". Ausführlicher zum TopicExplorer und der dahinter stehenden Programmierung siehe: https://github.com/hinneburg/TopicExplorer.

19 Vgl. https://github.com/mimno/Mallet.

der Zeitungen und Zeitschriften in eine Topic-Modellierung einbezogen werden, da der Umfang der Beiträge in den halleschen Journalen große Schwankungen aufweist.[20]

Die Topic-Modellierung nutzt die verfügbaren bibliographischen Angaben (im Folgenden als Dokument bezeichnet) und ordnet möglichst jedes Wort (Token) eines Dokuments Themen zu, wobei die Ausgangsanzahl der Themen variiert werden kann und die Themen selbst Ähnlichkeiten zwischen den einzelnen Dokumenten erklären sollen. Im Resultat erhält man zunächst unabhängige Annahmen über die Daten, die aufgrund von Plausibilitätskriterien in Themengruppen gegliedert werden, wobei auch immer Mehrfachzuweisungen möglich und gewünscht sind.

Relativ aufwendig gestaltet sich die Vorverarbeitung der Dokumente. Satzzeichen und Stoppwörter, Einzelbuchstaben und Zahlen müssen entfernt werden. Mithilfe des TreeTagger[21] werden die Wörter der Dokumente in Wortarten klassifiziert (Part-Of-Speech tagging). Für historische Texte stehen zudem Tools zur Verfügung, die eine Vereinheitlichung und Anpassung der variierenden Schreibweisen an die aktuelle Rechtschreibung ermöglichen.[22] Zwar stellen abweichende Schreibweisen für das Topic-Modeling grundlegend keine Schwierigkeit dar, da nur die Zeichenfolge für die Analyse Relevanz hat; Probleme können jedoch bei der Lemmatisierung entstehen, sodass eine Normalisierung durchaus sinnvoll ist. Anschließend können die Wörter nach Wortarten für die Topic-Analyse sortiert werden. Substantive tragen dabei die meisten inhaltlichen Bedeutungsdetails eines Textes. Für die erste Überblicksanalyse erfolgte aus diesem Grund eine Konzentration auf Substantive, weitere Wortarten sollen in späteren Verfahren hinzugenommen werden. Da sehr häufige und sehr seltene Wörter die Genauigkeit von Topic-Modellen mindern, wurden niedrig-frequente Wörter entfernt und hochfrequente Begriffe einer detaillierten manuellen Selektion unterworfen. Durch die anschließende Lemmatisierung wurde das Vokabular für die Topic-Analyse so genau wie möglich bestimmt. Um die spätere Interpretation der generierten Topics zu erleichtern, erfolgte die Aufnahme von Metadaten zu jedem Dokument. Dazu zählen, Autor, Veröffentlichungszeitpunkt und Journaltitel. Diese Metadaten werden nicht in die Topic-Generierung einbezogen, stehen aber für die Interpretation der Topics in der späteren Visualisierung zur Verfügung, beziehungsweise können sie nach der

---

20 Vorstellbar wäre in diesem Zusammenhang, die innere Struktur langer Beiträge (Zwischenüberschriften) zu übernehmen oder jeweils auf der Ebene der Seiten zu arbeiten.

21 Vgl. Schmid, „Prohabilistic".

22 Beispielsweise das Tool CAB (Cascaded Analysis Broker) Historical Text, siehe: Jurish, „Canonicalizing".

Modellierung durch SQL-Abfragen an die Datenbank variabel abgefragt werden. Die Festlegung auf zunächst 30 Themen erfolgte nach mehreren Test-Modellierungen. Grundlegend bietet der genutzte TopicExplorer aber die komfortable Möglichkeit mehrere Themen zusammenzufassen und wieder zu trennen, so dass in Hinblick auf die Themenanzahl ein vom Bearbeiter später zu nutzender Spielraum besteht.

Als ein Problem erwies sich bei Betrachtung der ersten Ergebnisse die Sprache. Eine nicht geringe Anzahl von Dokumenten ist in lateinischer oder französischer Sprache verfasst. Diese wurden zwar gleichberechtigt in die Analyse einbezogen und entsprechend Themen zugeordnet, doch besteht noch keine Möglichkeit einer automatisierten Übersetzung. Daher wurden beispielsweise zwischen synonymen lateinischen Begriffen und deutschsprachigen Bezeichnungen keine Beziehungen hergestellt. Dies führte dazu, dass sich insbesondere die französischsprachigen Beiträge nahezu ausschließlich in eigenständigen Themen wiederfanden, ohne einen Bezug zu ähnlichen Themen deutschsprachiger Dokumente aufzuweisen. Diese Beobachtung lässt es als zwingend notwendig erscheinen, eine sprachliche Angleichung (Übersetzung) vorzulagern. In unserem Beispiel betrifft dies cirka ein Fünftel der Dokumente.

Eine für unseren Kontext sehr spezifische Schwierigkeit bilden die sowohl innerhalb der Dokumente als auch im Zusammenhang mit den bibliographischen Informationen zu den Zeitungen und Zeitschriften enthaltenen Personennamen und Ortsbezeichnungen. Insbesondere bei Rezensionen und auch in naturwissenschaftlichen Journalen stellen beispielsweise die namentlichen Bezüge zu Autoren von wissenschaftlichen Abhandlungen einen entscheidenden Orientierungspunkt dar, da anhand dieser Informationen bewertet werden kann, auf welche Gelehrte sich andere Publikationen bezogen. Bei der Bereinigung der Wortlisten wurde bewusst auf Initialen verzichtet, da dies in einem frühen Versuch zu Themenlisten geführt hatte, die zum Beispiel nur Einzelbuchstaben (S., P., W. et cetera) enthielten. Indessen erscheint es nicht zielführend, nur mit Nachnamen in die Analyse zu gehen, da in diesem Fall keine Eindeutigkeit erreicht werden kann. Andererseits werden zahlreiche Namen nicht als Namen erkannt, bzw. es erfolgt nicht in allen Fällen eine entsprechende Zuweisung zu Themen. Dies führte zu einem zusätzlichen vorbereitenden Element – der Normierung von Namensschreibweisen. Wobei es notwendig war, die Vornamen als feste Einheiten an die Nachnamen zu binden und auch die Form dieser Koppelung zu normieren.

Etwas schwieriger gestaltet sich die ähnlich gelagerte Fragestellung in Hinblick auf Ortsnamen. Diese finden sich ebenfalls häufig in Dokumenten, weisen aber in vielen Fällen keine analoge Relevanz für die Themengenerierung auf, wie dies bei den Personennamen angenommen

werden muss. Häufig beziehen sich Ortsbezeichnungen, insbesondere in Rezensionszeitschriften, auf den Verlagsort – diese Information ist zwar nicht ohne Interesse, bildet aber nur in Ausnahmefällen eine entscheidende Komponente für die Themenmodellierung. Die dadurch provozierte Häufung führte in ersten Tests zu einer Verzerrung der Themenlisten. So fanden sich in einem Thema auffallend viele Ortsnamen (Berlin, Paris, Petersburg, England, Brüssel, Venedig et cetera), deren Beziehung durch die anderen dem Thema zugeordneten Begriffe nicht ersichtlich wurde. In dieser Hinsicht müssen noch weitere Versuche erfolgen, um zu testen, wie sinnvoll es sein kann, auf Ortsnamen vollständig zu verzichten, bzw. um zu entscheiden, in welchen Fällen diese für die Topic-Modellierung unerlässlich sind. Dies ist beispielsweise bei Reisebeschreibungen, in Verbindung mit der Frage, welche Weltteile als besonders interessant und erforschenswert galten, wie auch in Bezug auf Staatenbeschreibungen offensichtlich.

Die bisher skizzierten Schwierigkeiten, Probleme und weiter zu entwickelnden Anforderungen sollen jedoch keineswegs verdecken, dass bereits beeindruckende Ergebnisse erzielt werden konnten.

4.1 Der TopicExplorer in der praktischen Anwendung[23]

Abbildung 1 gibt einen Eindruck von der Nutzeroberfläche des TopicExplorers;[24] in der unteren Leiste ist ein Teil der (in ihrer Anzahl vorgegebenen) 30 Themen zu erkennen.

Die Topics werden als Wortlisten dargestellt, deren Wörter nach absteigender Wahrscheinlichkeit sortiert sind, sodass die häufigsten Wörter in einem Topic oben stehen. Um das gesamte Topic-Spektrum von links nach rechts ohne inhaltliche Brüche durchsehen zu können, werden die Topics linear nach wortbasierter Ähnlichkeit geordnet. Die berechnete Themenreihenfolge wird als Regenbogenskala dargestellt, wobei die Farbe als visuelle Hash-Funktion dient, da eine Ähnlichkeit der Farbe auf eine Ähnlichkeit zwischen den Themen hinweist. In der oberen Hälfte hingegen sind die, einem Thema zugeordneten, Dokumente aufgelistet.

---

23 Ausführlich zur Handhabung des TopicExplorers: https://blogs.urz.uni-halle.de/hallejournals/kategorie/methoden/topic-modelling/.

24 Der Prototyp des TopicExplorers zu den halleschen Journalen kann unter dem Link: http://topicexplorer.informatik.uni-halle.de:8080/HALJOUR_0_te/# eingesehen werden.

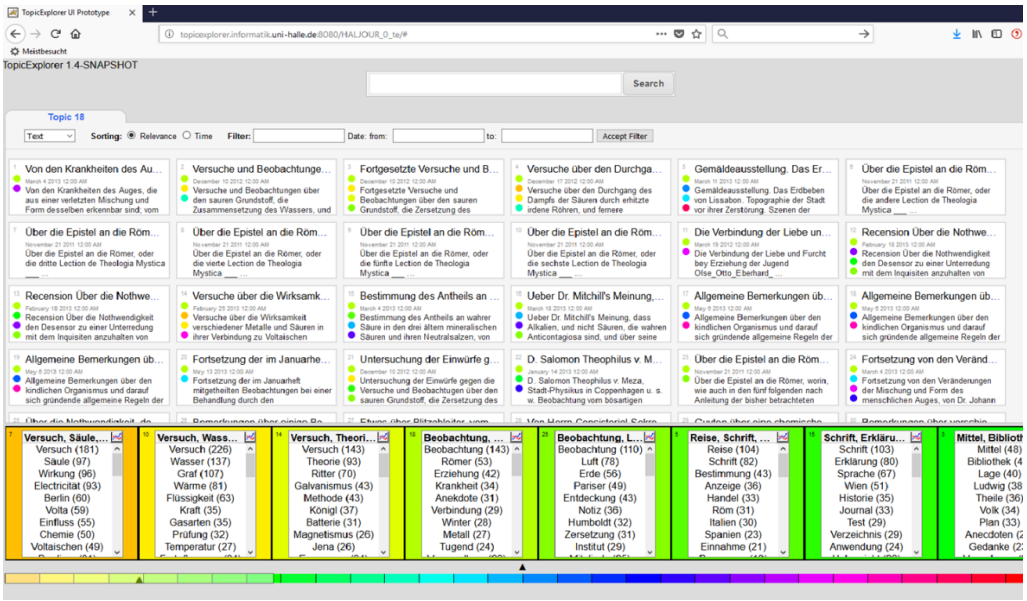

Abbildung 1: Screenshot TopicExplorer zu den halleschen Journalen.

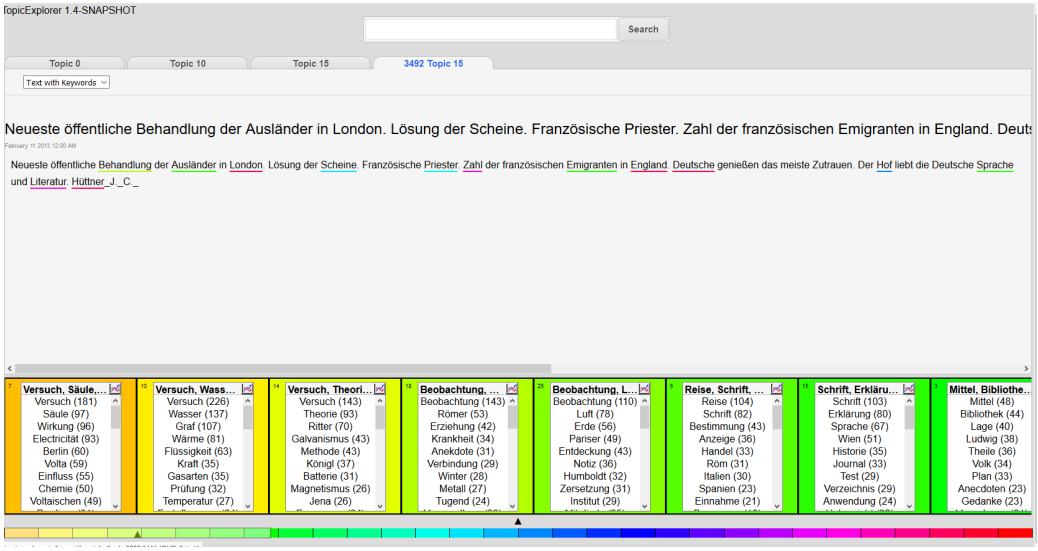

Abbildung 2: Screenshot TopicExplorer zu den halleschen Journalen.

Durch die farbigen Punkte neben den Dokumenten wird die jeweilige Themenzuweisung (je Dokument die vier häufigsten Themen, basierend auf der Zuordnung von Token eines Dokuments zu Themen) verdeutlicht. Die Oberfläche des TopicExplorer ermöglicht es nahe beieinanderliegende Themen zusammenzuführen, sie bietet eine Suchfunktion mit Auto-Vervollständigen, unterschiedliche zeitliche Filterfunktionen und die Option, für jedes einzelne Dokument die Zuordnung von Wörtern zu Themen sichtbar zu machen (Abbildung 2). Im TopicExplorer selbst kann die zeitliche Entwicklung jedes Topic visualisiert werden.

4.2 Ergebnisse des Topic-Modeling

Betont werden muss an dieser Stelle ausdrücklich, dass die generierten Themen zunächst keinerlei Interpretation beinhalten, diese muss erst durch den Bearbeiter, beziehungsweise durch das Hinzuziehen ergänzender Informationen geleistet werden.[25] Einige der generierten Themen weisen bereits auf den ersten Blick eine relativ hohe Verständlichkeit auf. Doch nicht alle abgebildeten Themen offenbaren eine so hohe Klarheit. Um sie erschließen, verorten und interpretieren zu können, werden zwingend zusätzliche Informationen benötigt. Die dafür erforderlichen Metadaten können gleichzeitig dazu dienen, die Struktur und Stabilität der generierten Topics zu überprüfen. Für eine adäquate Darstellung und Einbeziehung dieser Metadaten, erscheint die Netzwerkanalyse als das vielversprechendste Instrument.

Der TopicExplorer speichert alle Daten der Texte und des Topic-Modells in einer relationalen Datenbank. Für weiterführende Analysen und Datenexporte kann auf diese Datenbank, unabhängig vom TopicExplorer, mithilfe von SQL-Abfragen zugegriffen werde.[26] So kann beispielsweise für jedes Thema hinterfragt werden, welche Autoren und/oder Journale zu welchem Zeitpunkt involviert waren. Diese Abfragen stellen gewichtete Beziehungen zwischen Thema und Journal, Thema und Autor und Thema, Autor und Journal her. Dadurch kann gezeigt werden, wie sich Themen in Relation zu/mit Zeitschriften oder Akteuren ausbilden. Des Gleichen besteht die Möglichkeit, die zeitliche Entwicklung aller – oder einzelner Themen – in Relation zu den Journalen zu setzen; damit kann beispielsweise die Frage beantwortet werden, ob bestimmte Zeitschriften Themenkonjunkturen bedingen. Dies erscheint

---

25 Vgl. zur Problematik der Interpretation von Topics: Chang et al., „Reading Tea Leaves"; Sievert und Shirley, „LDAvis"; Kataria et al., „Context senxitive"; Riddell, „Text Analysis"; Tangherlini und Leonard, „Traweling in the Sea".

26 Eine Auswahl gängiger SQL-Statements findet sich unter: https://github.com/hinneburg/TopicExplorer/blob/master/sql-analytics.md.

aktuell insbesondere in Hinblick auf die naturwissenschaftlichen Zeitschriften als sehr naheliegend.

5. Topic-Modeling und Netzwerkanalyse (experimentelles Beispiel)

Da sich unser Projekt noch in einem Frühstadium befindet, sollen im Folgenden insbesondere die Visionen einer Kombination zwischen Topic-Modeling und Netzwerkanalyse an einem exemplarischen Beispiel in ersten Ansätzen aufgezeigt werden. Grundlage dafür bildet das Thema 4. Alle folgenden Darstellungen gehorchen dabei keinem vorangestellten Narrativ, dass durch die voraussetzungsfreie Anwendung einer Topic-Modellierung gerade verhindert werden soll, sondern stellen – wie auch die in Beziehung gesetzten Vergleiche – ausschließlich aus dem Projekt und den genutzten Quellen hervorgegangene Interpretationen dar.

Dem Thema sind knapp 1.000 Dokumente zugeordnet, die keinesfalls alle gleichermaßen in eine Netzwerkanalyse einbezogen werden sollen und können. Vor diesem Hintergrund erfolgt über eine entsprechende SQL-Abfrage, die die Metadaten der Topic-Modellierung berücksichtigt, eine Hierarchisierung der involvierten Journale und Autoren. Dafür wird die Summe aller Dokumente mit einer Tokenzuordnung zu Thema 4 in Relation zu Journalen und Autoren abgefragt. Für die Gesamtmenge dieser Dokumente wird dann die durchschnittliche Anzahl der dem Thema zugeordneten Token berechnet. Diese Zahl dient im Folgenden zugleich für die Gewichtung der Relationen innerhalb der Netzwerkanalyse. In die spätere Untersuchung werden nur durchschnittliche Werte >1 einbezogen. Durch entsprechende SQL-Abfragen können unterschiedliche Listen generiert werden. Zum einen eine nach absteigender Relevanz sortierte Liste zu den an Thema 4 beteiligten Journalen (Tabelle 1) und eine gleichermaßen strukturierte Aufzählung der involvierten Autoren (Tabelle 2). Damit wird zum Ausdruck gebracht, welche Autoren/Journale in welcher Intensität mit dem Thema verknüpft sind.

Durchschnittliche Anzahl Token der summierten Dokumente	Journal_Titel (Kurztitel)
3,67	Magdeburg-halberstädtische Blätter
2,00	Magazin für die Religion
1,70	London und Paris
1,50	Beyträge zur Beförderung der Erkenntnislehre
1,50	Merkwürdige Krankengeschichten
1,33	Allgemeine Geschichte der europäischen Staaten
1,33	Auserlesene und noch nie gedruckte Gedichte
1,33	Compendiöse Bibliothek der schönen Wissenschaften
1,20	Neue Reisebemerkungen
1,19	Magazin für die neue Historie und Geographie
1,16	(Fortgesetzte) Betrachtungen über die Werke Gottes in der Natur
1,13	Journal der Physik
1,13	Der Andächtige
1,13	Magazin der deutschen Critik
1,09	Der Naturforscher
1,09	Deutsche Bibliothek der schönen Wissenschaften
1,06	Allgemeine historische Bibliothek
1,06	Annalen der Physik

Tabelle 1: Thema 4. Durchschnittliche Anzahl Token der summierten Dokumente je Journal über den Gesamtzeitraum (Ausschnitt).

Durchschnittliche Anzahl Token der summierten Dokumente	Name_Autor
2,00	Campbell, H.
2,00	Dohm, Christian Wilhelm
2,00	Gilbert, Ludwig Wilhelm
2,00	Hensler, Philipp
2,00	Monroe, Donald
2,00	O.
2,00	Porteau
2,00	Rambach, Friedrich Eberhard
2,00	Schloiser, Johann Heinrich
1,84	Hüttner, J. C.
1,78	Aubigny gen. Engelbrunner, N.
1,33	Berthollet
1,22	Hostig, K. G.

Tabelle 2: Thema 4. Durchschnittliche Anzahl Token der summierten Dokumente je Autor über den Gesamtzeitraum (Ausschnitt).

Bereits eine solche Darstellung kann, mit Vorwissen über die fachliche Zuordnung von Autoren und Journalen, Interpretationsansätze bieten, so handelt es sich offensichtlich um ein Thema mit dem geographischen Schwerpunkt England, mit dem eine Vielzahl thematisch unterschiedlicher Journale in Beziehung standen. Die Inhalte reichen von Geschichte (*Allgemeine Geschichte der Europäischen Staaten*, *Allgemeine historische Bibliothek*), Unterhaltung (*London und Paris*), Literatur (*Auserlesene und noch nie gedruckte Gedichte*, *Magazin der deutschen Critik*, *Deutsche Bibliothek der schönen Wissenschaften*) über Naturwissenschaften (*Journal der Physik*, *Der Naturforscher*), bis hin zu Theologie (*Magazin für die Religion*, *Betrachtungen über die Werke Gottes in der Natur*). Der Schwerpunkt liegt jedoch auf einer allgemeinen Wissensvermittlung über die aktuellen politischen, wissenschaftlichen und sozialen Verhältnisse in England (*Magdeburg-halberstädtische Blätter*, *Neue Reisebemerkungen*, *Magazin für die neue Historie und Geographie*). Mit einer solchen Überblicksdarstellung kann jedoch, auch unter Hinzuziehung der Metadaten zu Autoren und Journalen, noch nicht beantwortet werden, ob alle Bereiche im zeitlichen Verlauf in gleicher Intensität abgedeckt wurden, beziehungsweise welchen inhaltlichen Verschiebungen das Thema unterlag.

Ausgehend von den Möglichkeiten des TopicExplorers kann die Entwicklung des Themas über den Gesamtzeitraum betrachtet werden (Abbildung 3). Diese Visualisierung bietet zudem die Option, für jeden beliebigen Zeitpunkt die vier jeweils dominanten Token anzeigen zu lassen.

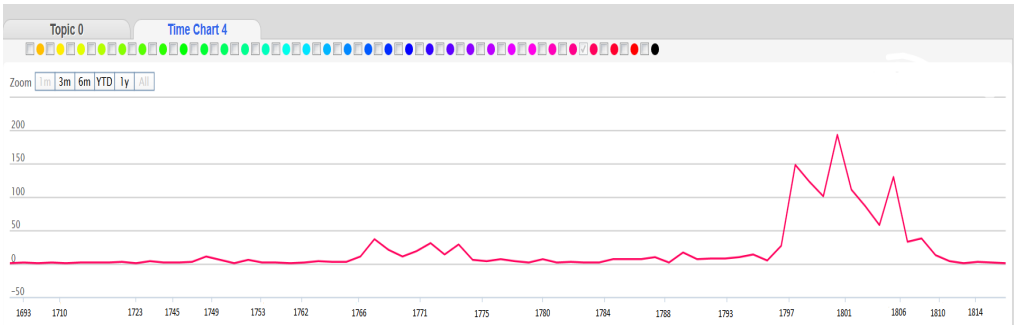

Abbildung 3: Zeitliche Entwicklung von Thema 4. Darstellung im TopicExplorer.

An dieser Stelle kann somit die grundlegende Entwicklung eines Themas, einhergehend mit begrifflichen Verschiebungen und Schwerpunktlegungen reflektiert werden. Erklärungen dafür, warum es beispielsweise zu einer Zunahme der Beiträge kam, beziehungsweise welche Faktoren die erkennbaren Schwankungen verursachen, bietet der TopicExplorer nicht.

Um diese Fragen zu beantworten, werden die Informationen der generierten Listen durch eine Netzwerkanalyse miteinander in Beziehung

gesetzt. An dieser Stelle in die Netzwerkanalyse integriert werden Metadaten zu den beteiligten Journalen (Fachzuordnung, Erscheinungsweise, Erscheinungszeitraum, Herausgeber, Drucker, Verleger et cetera) und Autoren (persönliche Beziehungen, Beruf, Aufenthaltsort, Publikationen, Fachgebiet et cetera). Für die sich anschließende Darstellung liegt der Schwerpunkt auf der fachlichen Zuordnung der Journale und Autoren, um damit inhaltliche und thematische Verschiebungen innerhalb der Entwicklung des Themas aufzeigen zu können. Infolge dieser Ergänzung von Daten entsteht ein multimodales Netzwerk, dessen, für die Wissensrekonstruktuion elementare Akteurstypen: Thema, Autor und Journal darstellen. Die entsprechenden gerichteten Relationen bringen zum Ausdruck, welche Journale bzw. Akteure sich an dem entsprechenden Thema beteiligen und welcher Fachrichtung sie zugeordnet werden können – Mehrfachzuweisungen sind dabei möglich. Die Gewichtung der Relation zwischen Journal/Autor und Thema erfolgt anhand der SQL-Abfragen in der Datenbank des TopicExplorers, die in diesem Fall mit zeitlichen Grenzen versehen sind.

Mithilfe dieser Daten kann quasi das Ego-Netzwerk von Thema 4 im Zeitverlauf rekonstruiert werden, wobei die zeitliche Dimension zum einen an die Erscheinungsdauer der Journale und zum anderen an das Publikationsdatum der zugeordneten Beiträge gebunden wird. Thema 4 stellt das ‚Ego' dar, während Journale und Autoren die ‚Alteri' bilden. Mit dieser Konstellation kann das Netzwerk um Thema 4 zunächst isoliert von anderen Themen in seiner Entwicklung und Ausprägung untersucht werden. In Hinblick auf das behandelte Thema ergibt sich ein Zeitraum zwischen 1689 und 1817, der für eine vergleichende Analyse in vier mathematisch etwa gleich große Abschnitte eingeteilt wird. Anhand einer solchen Strukturierung, die zunächst nicht an herausragenden Ereignissen, sondern vielmehr an einer gleichmäßigen Verteilung über den Gesamtzeitraum orientiert ist, können thematische, mediale und personelle Schwerpunktverschiebungen erforscht werden. Eine generelle Problematik liegt, insbesondere für die ersten Jahrzehnte des Betrachtungszeitraums, in der häufig fehlenden Zuordnung von Beiträgen zu Autoren, da eine namentliche Kennzeichnung von Artikeln erst im Lauf des 18. Jahrhunderts eine zunehmende Verbreitung erfuhr.

Die Netzwerkvisualisierung selbst erfolgt mit Gephi[27] unter Verwendung des Fruchterman Reingold Algorithmus[28]. Für den ersten Abschnitt zwischen 1689 und 1725 finden sich sieben Zeitschriften, die sich in etwa gleicher Intensität an Thema 4 beteiligten (Abbildung 4). Gruppiert man die involvierten Journale nach ihrer inhaltlichen Ausrichtung, so findet sich eine –

---

27 https://gephi.org/ und https://github.com/gephi/gephi. Version 0.9.2.

28 Vgl. Fruchterman und Reingold, „Graph drawing".

dem grundlegenden Schwerpunkt der halleschen Journale dieser Zeit entsprechende – Konzentration auf Rezensionszeitschriften mit theologischer und staatswissenschaftlicher Ausrichtung.

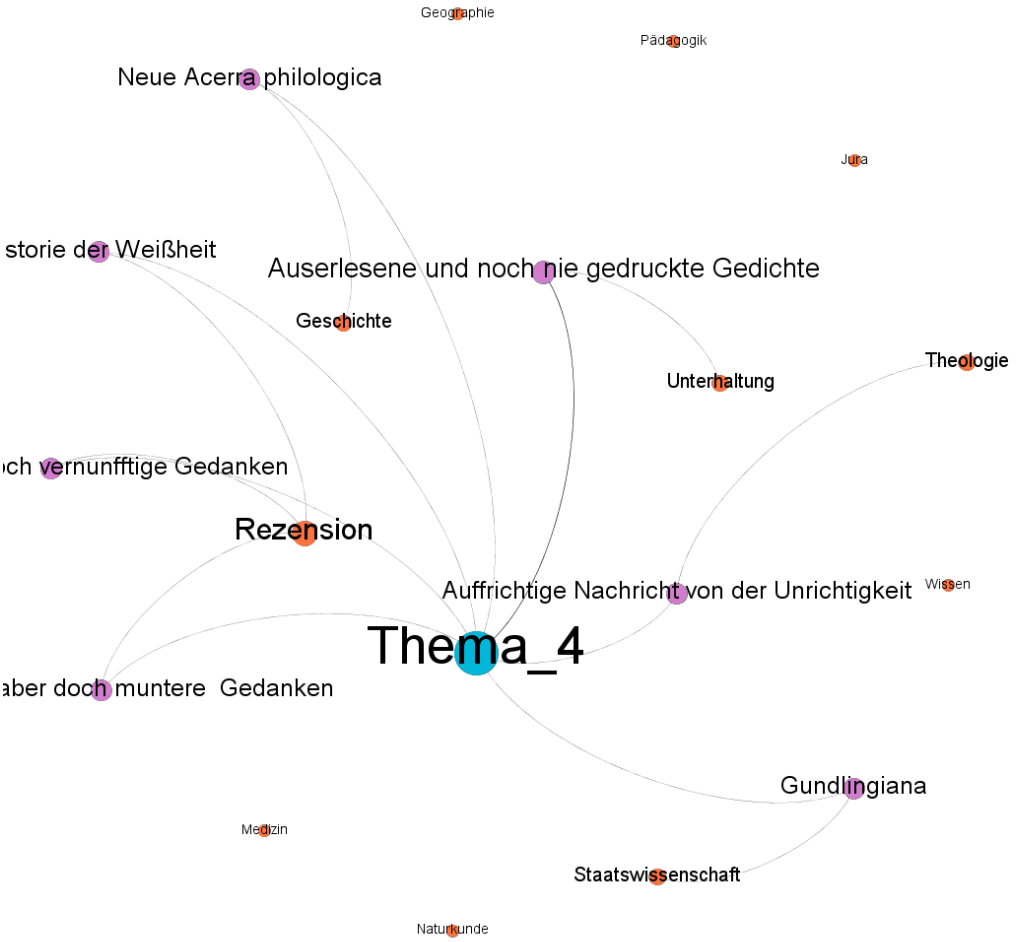

Abbildung 4: Thema 4 Netzwerk zwischen 1689 und 1725. Violet = Journale, rot = Fachzuordnung, Knotengröße und Schriftgröße = Gewichteter Grad.

Keiner der entsprechenden Beiträge ist dabei mit einem Autorennamen gekennzeichnet. Gleichzeitig verläuft das Thema sehr flach, das bedeutet die Diskursintensität des Themas ist nur gering ausgeprägt (Abbildung 3), eine hervorstechende Auseinandersetzung mit England ist somit noch nicht existent, sondern vielmehr Nebenprodukt derjenigen Journale Halles, die es sich zur Aufgabe gemacht haben, eine große Bandbreite von Themen, Gegenständen oder Publikationen zu betrachten. Aus diesem Grund findet sich das Thema, meist relativ isoliert in den zeitgenössisch herausragenden halleschen Journalen. So beispielsweise in den von Christian Thomasius

herausgegebenen Zeitschriften *Freymüthige jedoch vernunfft- und gesetzmäßige Gedanken über allerhand fürnehmliche aber neue Bücher (1689-1690)* und den *Ernsthaffte aber doch munter und vernünfftige Thomaische Gedancken und Erinnerungen über allerhand außerlesene juristische Händel (1720-1721).*[29] Die staatswissenschaftliche Dimension wird vorrangig durch die von Nicolaus Hieronymus Gundling herausgegebenen *Gundlingiana*[30] *(1715-1732)* und die *Neue Acerra philologica*[31] *(1715-1723)* abgedeckt. An dieser Stelle gilt es im Projekt Maße zu entwickeln, die Aussagen darüber ermöglichen, ab welcher ‚Beteiligungsintensität' und ‚Akteursvielfalt' von einem Diskurs gesprochen werden kann. Zudem erforderlich ist die Etablierung unterschiedlicher Diskurstypen (beispielsweise Laien- oder Expertendiskurs).

Erst in der zweiten Periode zwischen 1726 und 1756 lassen sich einzelne Autoren identifizieren. Auffällig ist, dass die grundlegende Themenentwicklung in ihren Schwerpunkten den allgemeinen Tendenzen des halleschen Zeitschriftenmarktes, der zu diesem Zeitpunkt unterhaltenden Schriften einen größer werdenden Raum eröffnete, folgt. Dieser Beobachtung widersprechend sticht die pädagogische Konnotation des Themas heraus, die sich in dieser Intensität nicht im Gesamtspektrum der halleschen Journale wiederfindet (Abbildung 5).

Diese thematische Konzentration wird hauptsächlich von der pädagogischen Zeitschrift *Altes und Neues von Schulsachen (1752-1755)*, herausgegeben von Johann Gottlieb Biederman getragen. An diesem Beispiel kann belegt werden, dass, da die grundlegende Auseinandersetzung mit dem englischen Thema noch sehr gering ausfiel, bereits ein Journal eine wesentliche Verschiebung in der inhaltlichen Ausrichtung bedingen konnte ohne dabei zwingend eine Kontroverse initiieren zu müssen. Im Unterschied dazu beschäftigten sich die literarisch-kulturelle Publikumszeitschrift *Bemühungen zur Beförderung der Critik und des guten Geschmacks (1743-1747)*, herausgegeben von Christlob Mylius und Johann Andreas Cramer und die von Friedrich Eberhard Rambach veröffentlichte theologische Fachzeitschrift *Beyträge zur*

---

29 Ein weiteres Journal Thomasius' das diesen Themenbereich berührt ist die *Historie der Weißheit und Thorheit (1693)*.

30 Vollständiger Titel: *Gundlingiana: darinnen allerhand zur Jurisprudentz, Philosophie, Historie, Critic, Litteratur und übrigen Gelehrsamkeit gehörige Sachen abgehandelt werden*.

31 Vollständiger Titel: *Neue Acerra philologica oder Gründliche Nachrichten aus der Philologie, und denen römischen und griechischen Antiquitaeten: darinnen die schweresten Stellen aller auctorum classicorum der studierenden Jugend zum Besten in einer angenehmen Erzehlung kürtzlich und gründlich erkläret werden*.

Beförderung der Erkänntniß (1754-1759)[32] mit durchaus variierenden Elementen des englischen Themas.

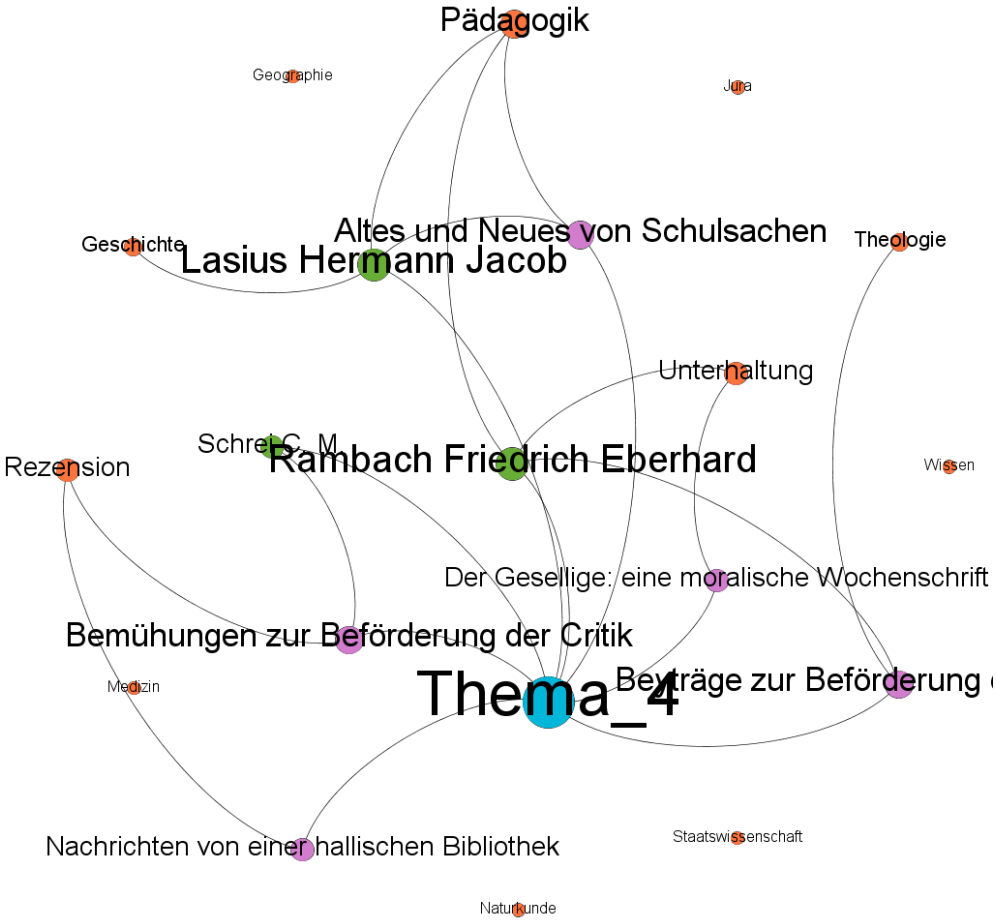

Abbildung 5: Thema 4 Netzwerk zwischen 1726 und 1756. Violet = Journale, rot = Fachzuordnung, grün = Autoren, Knotengröße und Schriftgröße = Gewichteter Grad.

---

32 Vollständiger Titel: *Beyträge zur Beförderung der Erkäntniß der Wahrheit zur Gottseligkeit aus den Schriften ausländischer und bewährter Gottesgelehrten*, als Fortsetzung der in Leipzig und Greifswald 1750 erscheinenden *Samlung auserlesener Abhandlungen ausländischer Gottesgelehrten, zur Unterweisung des Verstandes und Besserung des Herzens*.

Auch für die hervorstechenden Autoren kann konstatiert werden, dass aufgrund der noch eingeschränkten Breite des Themas einzelne Personen schnell zu einer relativen Verzerrung führen können. Sowohl Rambach, als auch Lasius beteiligten sich nur punktuell an einer Darstellung englischer Themen, da jedoch kaum weitere Autoren namentlich greifbar gemacht werden können, erscheinen die betreffenden Personen prägender als sie es tatsächlich waren. Um diesen, einem Mangel an Daten geschuldeten Verzerrungen zu begegnen, muss im Projektverlauf in Hinblick auf die Netzwerkrekonstruktion eine Lösung für den Umgang mit anonym erscheinenden Beiträgen angestrebt werden.

In der dritten Phase zwischen 1757 und 1787 weitete sich, entsprechend der allgemeinen Expansion des halleschen Zeitschriftenmarktes, die Anzahl der involvierten Journale und Autoren aus, wodurch zugleich inhaltliche Interessenverschiebungen bedingt wurden. Geschichte und Geographie rückten, gerade bei der Betrachtung ausländischer Staaten zunehmend in den Mittelpunkt, wie auch Fragen des Rechts. Zugleich offenbart sich, dass personelle Verknüpfungen, auch über Staatsgrenzen hinweg vor allem von Naturwissenschaften und Theologie getragen wurden. In diesem Bereich bestand relativ früh ein intensives Kommunikationsbedürfnis, das die gelehrten Briefwechsel endgültig in den Hintergrund drängte und zu einer personalisierten Wissenschaftspraxis in den naturwissenschaftlichen und theologischen Journalen führte. Eine Vielzahl herausragender Naturwissenschaftler publizierte Forschungsergebnisse in naturwissenschaftlichen und medizinischen Zeitschriften und auch Theologen nutzten die Möglichkeit, durch Publikationen in Journalen ihre Überzeugungen und Überlegungen einer größeren – auch weniger gebildeten – Öffentlichkeit zur Verfügung zu stellen. Parallel gewährleisteten Kompilationen, Zitationen, Rezensionen und Übersetzungen in beiden Bereichen eine relativ schnelle Verbreitung. Die hervorstechenden Autoren sind aufgrund dessen, mehrheitlich in Naturwissenschaften und Theologie verankert (Abbildung 6).

Generell ist der Stellenwert theologischer Betrachtungen und Abhandlungen interessant. Diese wurden zum einen von einflussreichen Autoren getragen, andererseits spiegelt sich hier der in Halle, auch in der publizistischen Praxis, lange Bestand habende theologische Schwerpunkt wider. Dieser führte nahezu im gesamten Betrachtungszeitraum zu einer beeindruckenden Stabilität theologischer Journale und Themen. Einige von ihnen setzten sich dabei offensichtlich auch mit theologischen Fragen in anderen räumlichen Kontexten auseinander. Zu ihnen gehörten beispielsweise das *Magazin für die Religion (1780-1781)* herausgegeben von Johann Salomo

Semler, die *Betrachtungen über die Werke Gottes (1772)*[33] veröffentlicht durch Christoph Christian Sturm oder der ebenfalls von Sturm, gemeinsam mit Jakob Friedrich Feddersen, herausgegebene *Der Andächtige (1773-1774).* Gleichbleibend finden sich Rezensionen in allen fachlichen Bereichen. Diese thematischen Schwerpunkte wurden zwischen 1757 und 1787 von einer Vielzahl von Journalen getragen, wie auch die generelle Zunahme von periodisch erscheinenden Schriften eine Verzerrung des Gesamtbildes, wie im zweiten Zeitabschnitt, zunehmend verhinderte.

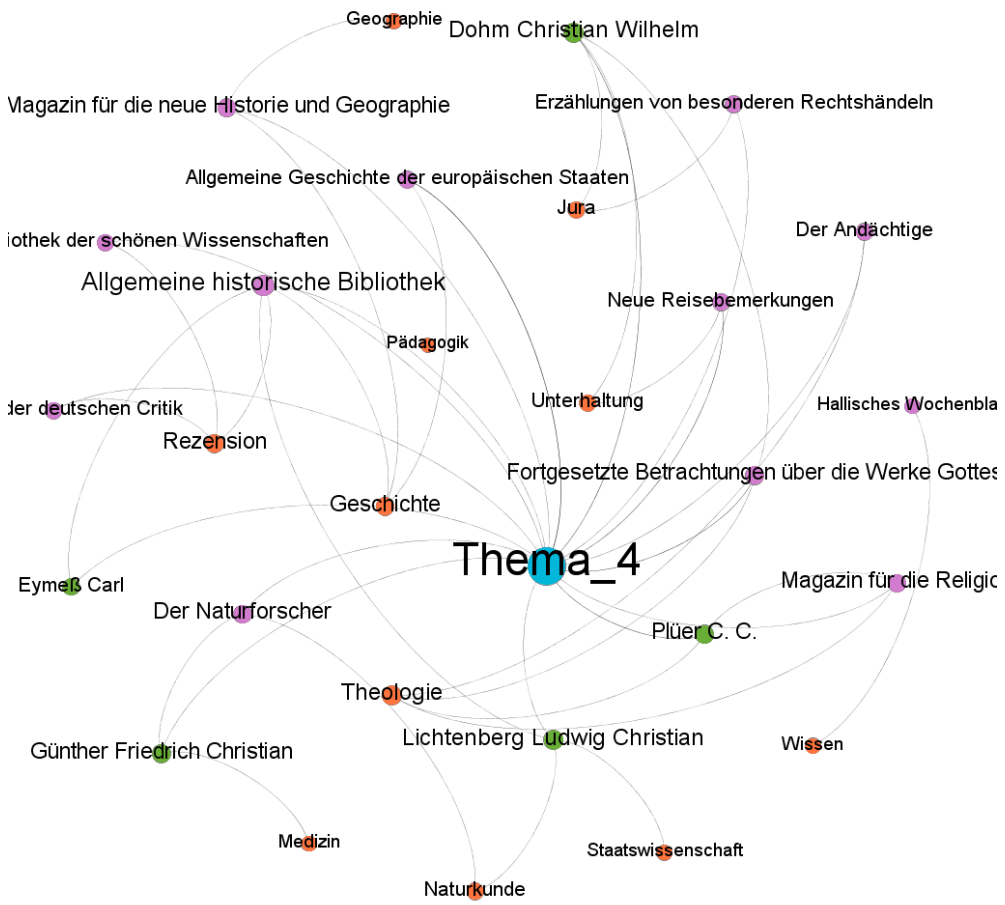

Abbildung 6: Thema 4 Netzwerk zwischen 1757 und 1787. Violet = Journale, rot = Fachzuordnung, grün = Autoren, Knotengröße und Schriftgröße = Gewichteter Grad.

---

33 Vollständiger Titel: *Betrachtungen über die Werke Gottes im Reiche der Natur und der Vorsehung auf alle Tage des Jahres.*

Für die letzte Periode zwischen 1788 und 1818 ist ein neuerlicher Anstieg der Beiträge zu verzeichnen, wobei die inhaltliche Vielfalt gegenüber zwei klaren, wenn auch divergierenden, Interessenschwerpunkten abnahm (Abbildung 7).

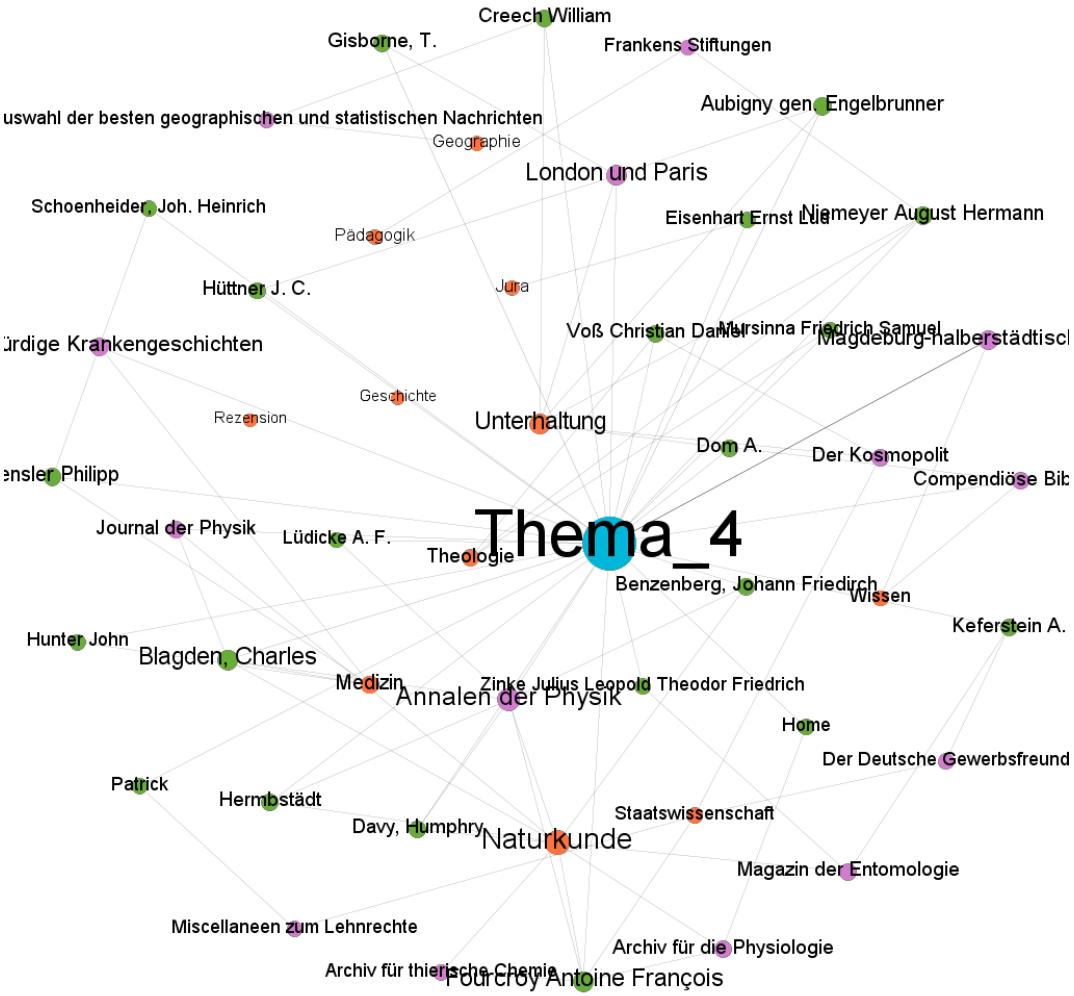

Abbildung 7: Thema 4 Netzwerk zwischen 1788 und 1818. Violet = Journale, rot = Fachzuordnung, grün = Autoren, Knotengröße und Schriftgröße = Gewichteter Grad.

Diese zwei Hauptinteressengebiete bilden die Unterhaltung, in ihrem Schwerpunkt getragen vom Journal *London und Paris*, das über alle interessant erscheinenden Moden, kulturellen Ereignisse, Literatur, bildende Künste et cetera aus den beiden Metropolen Paris und London informierte. Ebenso berichteten kosmopolitisch ausgerichtete Zeitschriften und die allgemeine Tagespresse mit wechselnder Intensität aus und über London. Der diesem Bereich gegenüberliegende Schwerpunkt bestand in naturwissenschaftlichen

Abhandlungen. Qualitativ hochwertige naturwissenschaftliche Journale (insbesondere die *Annalen der Physik*) prägten seit 1790 den halleschen Markt. Es gelang zahlreichen medizinisch-naturwissenschaftlichen Zeitschriften sich in Halle zu etablieren, wobei betont werden muss, dass sich in diesen Fällen der Verbreitungsraum weit über die Grenzen Halles erstreckte. So weisen das *Magazin für Entomologie,* das *Archiv für Physiologie,* das *Journal der Physik* oder die *Merkwürdigen Krankengeschichten* eine intensive inhaltliche und personelle Vernetzung weit über die Grenzen der halleschen Universität auf. Der sich in der zeitlichen Darstellung des TopicExplorers offenbarende massive Anstieg seit 1797 (Abbildung 3) wurde damit zu großen Teilen von einer etwa gleich gewichteten Mischung aus neu begründeten naturwissenschaftlichen und unterhaltenden periodischen Schriften getragen. Im Unterschied dazu können kaum noch prägende Autoren ausgemacht werden. Andererseits war es unter Umständen das Verdienst von Günther, Rambacher, Dohm und anderen ein breites Interesse an England geweckt zu haben, das in den Jahren nach 1800 zwar zu einer zunehmenden Konzentration auf zwei Themenschwerpunkte führte, dabei jedoch von einem beeindruckenden Personenkreis getragen wurde.

Dieses beispielhafte Netzwerk soll im weiteren Projektverlauf um zusätzliche Informationen ergänzt werden, wie auch die zeitliche Einteilung entlang des im TopicExplorers generierten Verlaufs oder herausragenden historischen Ereignissen variierend strukturiert werden wird.

5 Ausblick

Dieser kurze Einblick in die Zielsetzung des Projekts sollte an einem kleinen Ausschnitt beleuchten, worin die Chancen und Möglichkeiten einer Kombination von Topic-Modeling und Netzwerkanalyse liegen. Keiner der beiden Ansätze allein würde eine solche Fülle von Daten berücksichtigen können und zudem flexibel angepasste Interpretationen und Vergleiche ermöglichen. Wendet man die dargestellte Herangehensweise auf die Gesamtheit der generierten 30 Topics und ihrer Repräsentanz in den halleschen Journalen an und nimmt zunächst noch keine weitere Unterscheidung in unterschiedliche zeitliche Phasen vor – kann ein Blick auf das Gesamtnetzwerk zwischen 1689 und 1824 bereits erste Eindrücke vermitteln. Berücksichtigt wurden in der folgenden Darstellung je Topic die zehn relevantesten Journale (Abbildung 8). Die struktuelle Vernetzung zwischen einzelnen Themen kann dabei über Netzwerkdichte und durchschnittlichen Grad betrachtet werden.

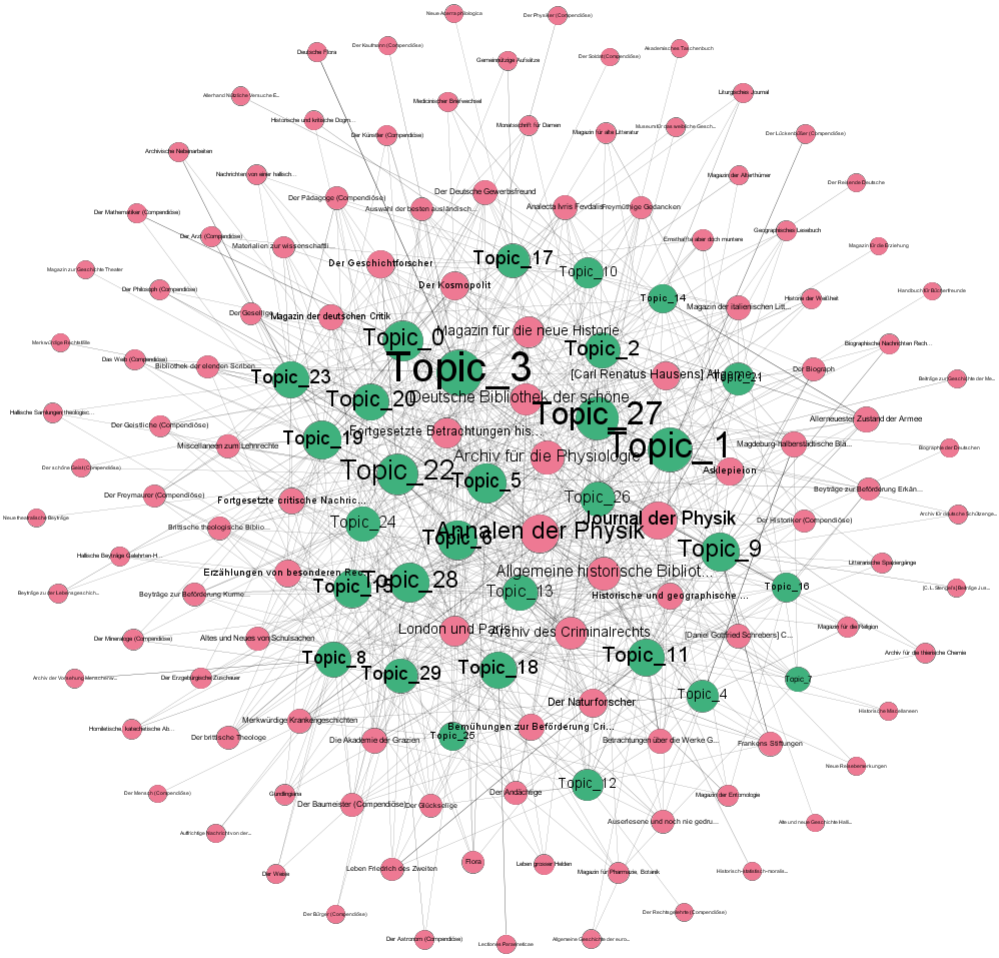

Abbildung 8: Topics mit Journalen (1689-1817). Grün = Topic, pink = Journal, Knotengröße = gewichteter Grad, Schriftgröße = Betweenness Centrality. Netzwerkdurchmesser = 4, mittlerer gewichteter Grad = 70,2.

Als besonders zentrale Themen und damit unter Umständen (Wissens)Diskurse, über den Gesamtzeitraum, erweisen sich Topic 3 mit Rezensionen und Beiträgen zu politischen und staatswissenschaftlichen Schwerpunkten, Topic 1 aktuelle Nachrichten zu Politik und vor allem regionaler Wirtschaft beinhaltend und Topic 27 mit eher beschreibenden Artikeln zu Geschichte und Verfassung unterschiedlicher Länder. Andere Themen, die häufig nur punktuell auftreten, beziehungsweise relativ spät in den Journalen der Aufklärungsepoche Raum erhalten, finden sich eher am Rand des Netzwerks. So beispielsweise Topic 7 mit den Kernpunkten elektrische Versuche und der Reflektion über die Erkenntnisse Alessandro

Voltas oder Topic 17, dessen Hauptaugenmerk auf Beschreibungen deutscher Städte, insbesondere ihrer Universitätskultur liegt. Analog dazu können relativ einfach Journale ausgemacht werden, die, da sie eine Vielzahl von Themen bedienen, in zentralen Positionen des Netzwerks gefunden werden können. Dazu zählten zum Beispiel die *Allgemeine historische Bibliothek* (1767-1771), die *Annalen der Physik* (1799-1824) oder die *Deutsche Bibliothek der schönen Wissenschaften* (1767-1770). Diese Periodika stechen dabei nicht zwingend mit einem besonders langen Erscheinungszeitraum hervor, sondern vielmehr mit ihrer Fähigkeit unterschiedlichste Themenbereiche umfangreich zu bedienen. Nicht ohne Grund finden sich unter ihnen zahlreiche Rezensionszeitschriften, ohne expliziten fachlichen Schwerpunkt. Betont werden muss an dieser Stelle noch einmal, dass es sich bei dieser Darstellung aktuell noch um eine Momentaufnahme handelt, da noch nicht alle halleschen Journale des betreffenden Zeitraums in die Topic-Modellierung einbezogen werden konnten.

Nichts desto trotz kann eine Unterscheidung in (zunächst) vier, der zeitlichen Differenzierung von Thema 4 entsprechende Abschnitte eine differenziertere Sichtweise auf die vorliegenden Daten und die Tendenzen der thematischen Entwicklung des halleschen Zeitschriftenmarktes ermöglichen (Abbildung 9-12), um detailliertere Informationen über Verschiebungen, Brüche und Entwicklungen in den Interessenschwerpunkten der halleschen Journale generieren, beziehungsweise einen Vergleich gestatten zu können.

Für den Zeitraum 1689-1725 (Abbildung 9) wird offensichtlich, dass sich der Zeitschriftenmarkt in Halle noch in den Kinderschuhen befand und nur ein sehr dünnes Wissensnetzwerk etablieren konnte. Acht Journale finden sich in diesem Zeitabschnitt wieder, berücksichtigt werden muss dabei eine gewisse Verzerrung, da nur die jeweils für ein Topic relevantesten Zeitschriften in die Visualisierung inbegriffen sind. Journale, die sich in geringerer Intensität an Themen beteiligen, werden zunächst noch ausgeschlossen. Deutlich wird bei den zentralen Journalen dieser Phase aber ihre Positionierung zwischen unterschiedlichen Themen. In Hinblick auf die Themen selbst ist erkennbar, dass eine Vielzahl von Themen noch nicht angesprochen wird, während alle relevanten Themen theologisch geprägt erscheinen. Im zweiten Zeitabschnitt zwischen 1726 und 1756 (Abbildung 10) ist eine zentrale Zeitschrift erkennbar, die insbesondere durch ihre intensive Vernetzung zu einer Vielzahl unterschiedlicher Themen heraussticht, während die *Gundlingiana*, mit einer thematischen Verbindung zu vier Topics, zwar mit der Intensität ihrer Relationen beeindruckt, sich aber dennoch eher am Rand des Netzwerkes findet. Inhaltlich rücken Themen zu Staatswissenschaften, Geschichte und Geographie (Topic 27) und Rezensionsthemen mit staatswissenschaftlichen Schwerpunkten (Topic 15) in den Mittelpunkt.

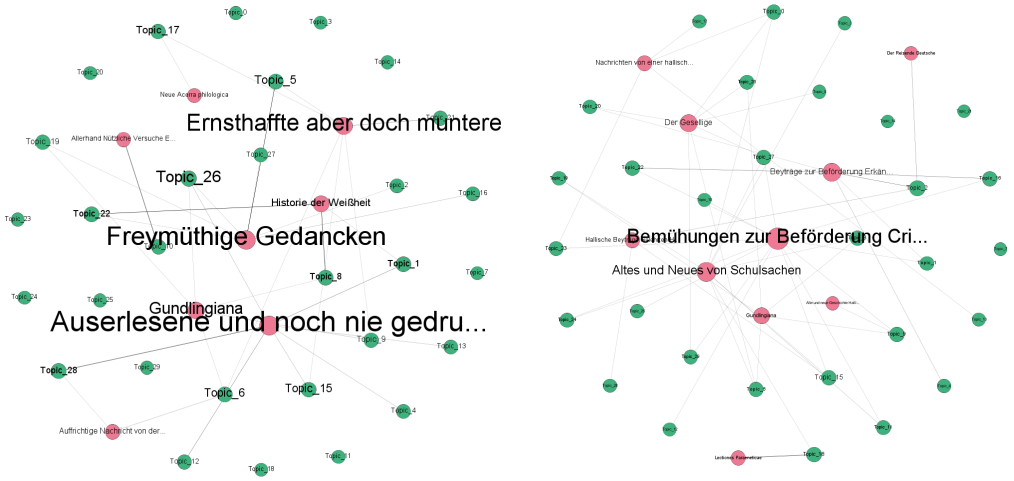

Abbildung 9 und 10: Themen und Journale 1689-1725 und 1726-1756. Grün = Topic, pink = Journal, Knotengröße = gewichteter Grad, Schriftgröße = Betweenness Centrality. Abbildung 9: Netzwerkdurchmesser = 6, mittlerer gewichteter Grad = 11,4. Abbildung 10: Netzwerkdurchmesser = 7, mittlerer gewichteter Grad = 16,6.

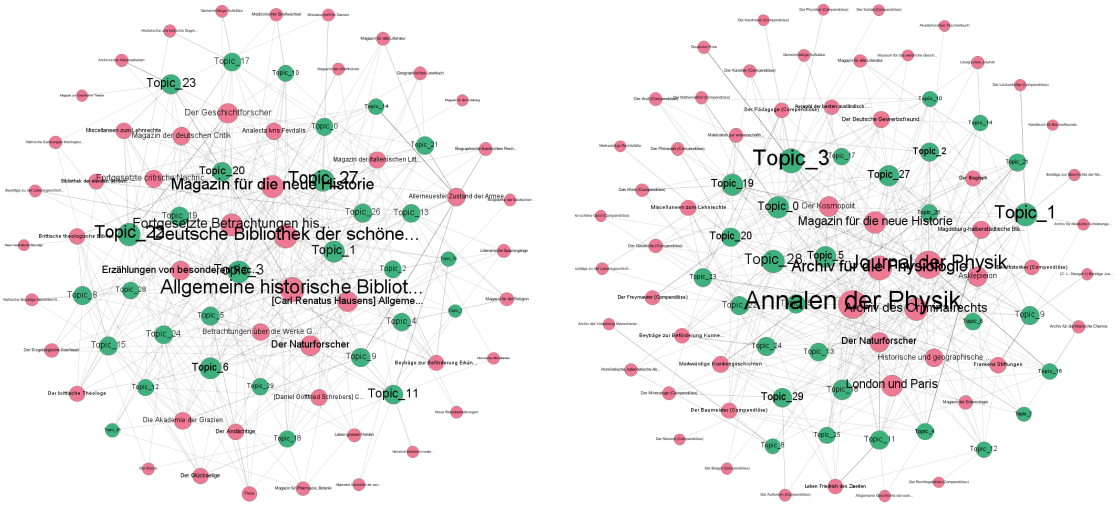

Abbildung 11 und 12: Themen und Journale 1757-1787 und 1788-1817. Grün = Topic, pink = Journal, Knotengröße = gewichteter Grad, Schriftgröße = Betweenness Centrality. Abbildung 11: Netzwerkdurchmesser = 4, mittlerer gewichteter Grad = 59,2. Abbildung 12: Netzwerkdurchmesser = 4, mittlerer gewichteter Grad = 60,4.

Für den dritten Abschnitt zwischen 1757-1787 (Abbildung 11) muss das Kantengewicht relativiert dargestellt werden, denn das Netzwerk verdichtet sich auf allen Ebenen, obwohl noch einzelne hervorragende Journale auszumachen sind, scheint der hallesche Zeitschriftenmarkt eine Vielfalt angenommen zu haben, die unterschiedlichsten periodischen Publikationen eine Existenz bieten konnte. Auf den ‚ersten Plätzen' (nach Betweenness Centrality) der Journale finden sich zwei allgemeine Rezensionsorgane (*Allgemeine historische Bibliothek*, Betweenness 260.9 und *Deutsche Bibliothek der schönen Wissenschaften*, Betweenness 254.2) und mit schon deutlicherem Abstand eine historische Zeitschrift (*Magazin für die neue Historie*, Betweenness 216.6). Zu diesem Zeitpunkt scheinen nahezu alle Themen bedient zu werden, wenn auch eine staatswissenschaftliche (Topic 27 und Topic 3) und eine literarisch-kulturelle (Topic 20 und Topic 6) Fokussierung auffällt.

In der letzten Phase zwischen 1788-1817 (Abbildung 12) dünnt sich das Netzwerk deutlich aus, verstärkt tritt zudem eine Differenzierung zwischen Themen und Journalen hervor, sodass von der relativen Ausgewogenheit insbesondere zwischen den Themen der vorangegangenen Phase nicht mehr gesprochen werden kann. Naturwissenschaftliche Journale drängen ins Zentrum des Netzwerkes und weisen einen deutlichen Vorsprung, in Hinblick auf ihre Zentralität, vor Zeitschriften mit anderen fachlichen Schwerpunkten auf. Ähnlich, wenn auch nicht in einer solchen Intensität und Vielzahl präsent sind unterhaltende Journale. Insbesonder theologische Themen und Journale werden an den Rand des Netzwerkes gedrängt. In diesem letzten Zeitabschnitt sind alle Themen integriert, das heißt, dass im Zeitverlauf trotz vielfältiger thematischer 'Neuerungen' und Erweiterungen keine Topics wieder vollkommen aus dem Netzwerk herausfallen. Auch wenn Sie deutlich seltener aufgegriffen wurden, blieben sie grundlegend bestehen. Die Themen dieser Phase werden dominiert von Landesbeschreibungen unterschiedlicher fachlicher Ausrichtung, physikalischen Abhandlungen aber auch (über)regionalen Nachrichten.

Die Gegenüberstellung der beiden letzten Zeitschnitte erweckt dabei den Eindruck eines relativ klaren Übergangs von historisch-literarisch-kulturellen Themen hin zu einer klaren Dominanz der Naturwissenschaften am Ende des 17. Jahrhunderts. Gerade diese klar erscheinenden Konzentrationsprozesse auf bestimmte Themen werden jedoch immer wieder durchbrochen von Zeiträumen, die durch eine Vielzahl von Themen und Journalen geprägt waren. Diese Beobachtung legt die Überlegung nahe, dass nach Abschluss eines relativ intensiv thematisierten Gegenstandes, innerhalb der periodischen Presse die Suche nach neuen, das Publikum und die Autoren interessierenden, Fragestellungen begann. Gerade hier bietet die Flexibiltät der Netzwerkanalyse gute Ansatzpunkte. Betrachtet man den ‚Zwischenzeitraum' 1777 bis 1797, also die jeweils 10 Jahre vor und nach dem obigen Schnitt ergibt sich ein ausgewogenes Bild (Abbilung 13). Bereits die Visualisierung lässt

erkennen, dass die Themen sich stärker aneinander annähern und von einer deutlich höheren Anzahl von Journalen ‚umgeben' werden. Für diesen Zeitraum erhöht sich die Anzahl der relevanten Themen deutlich. Im Vergleich weisen nur elf Themen für den Zeitraum 1757-1787 eine Betweenness >100 auf, im anschließenden Abschnitt 1788-1817 sind dies vierzehn Themen. Im Zwischenzeitraum jedoch finden sich neunzehn Themen mit einer solchen Betweenness im Netzwerk wieder, was auf eine relative ‚Ausgewogenheit' des thematischen Spektrums in diesem Zeitabschnitt hinweist.

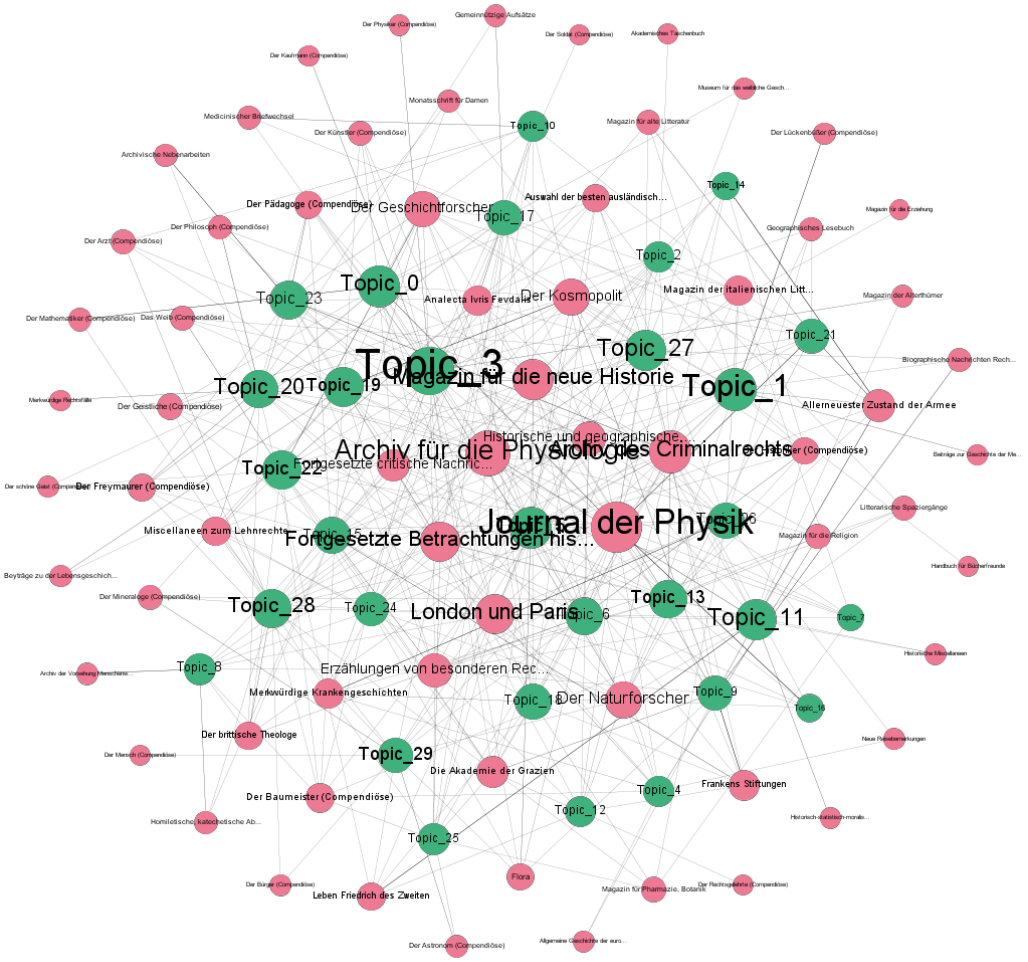

Abbildung 13: Themen und Journale 1777-1797. Grün = Topic, pink = Journal, Knotengröße = gewichteter Grad, Schriftgröße = Betweenness Centrality. Abbildung 11: Netzwerkdurchmesser = 4, mittlerer gewichteter Grad = 56,7.

Die in unseren ersten Versuchen generierten Themen geben Hinweise darauf, worüber, wann und in welchem Umfang in den halleschen Zeitungen und Zeitschriften gesprochen wurde. Ersichtlich wird aber auch, dass Topic-Modelle Themen hervorbringen können, die nach Themen aussehen, sich aber bei einer näheren Betrachtung eher als Relationen anderer Art entpuppen, so zum Beispiel geographische oder territoriale Eigennamen. Gleichzeitig wurden auch Gruppen generiert, die zunächst keinen semantischen oder thematischen Zusammenhang erkennen lassen, während andere, erwartete Themen (zum Beispiel das Thema 'Aufklärung') zu fehlen scheinen. Doch auch in diesen Fällen werden die Daten nicht mehr nur selektiv als 'Beispielspender' genutzt, sondern in allen ihren Widersprüchlichkeiten und zunächst auch einmal mangelnden Plausibilitäten ernst genommen. Für die spätere Analyse der erhaltenen Themen muss darauf Wert gelegt werden, dass alle Ergebnisse des Topic-Modeling einbezogen werden, nicht nur die positiv erscheinenden. In sich anschließenden Schritten führt dies dazu, sowohl die Wortlisten konstanter zu pflegen als auch die Voraussetzungen der Modellierung an die spezifischen Bedürfnisse des Nutzers und seiner Texte anzupassen. Einfluss auf die Ergebnisse haben zweifelsohne die Korpusgröße und -zusammenstellung und ebenso die Parameter des Topic-Modells (zum Beispiel die Topic-Anzahl, die Segmentlänge et cetera). Ein weiteres Problem stellen die zahlreichen erforderlichen Vorarbeiten dar. Diese beschränken sich nicht nur auf die Bereinigung der Wortlisten, sondern sollten für weitere Versuche ebenso Übersetzungen beinhalten. Hier erscheint es als durchaus sinnvoll und wünschenswert, zumindest einige dieser Vorarbeiten direkt in den TopicExplorer zu integrieren, um seine grundlegende Anwendbarkeit auf historische Dokumente weiter zu verbessern.

Als durchaus belastbar hat sich die Einbeziehung von Metadaten in die Datengrundlage des TopicExplorers erwiesen. SQL-Abfragen ermöglichen eine Filterung der mit den Topics in Beziehung stehenden Dokumente und die Überführung in eine Netzwerkanalyse, die zudem auf gewichtete Beziehungen anhand der Topic-Modellierung aufbauen kann. Perspektivisch sollen jedoch nicht nur diese Informationen in die Netwerkanalyse Eingang finden. Eine große Relevanz kommt den persönlichen Beziehungen zwischen den Akteuren zu, wobei Akteure in diesem Zusammenhang nicht länger nur als Autoren verstanden werden dürfen (wie im aktuellen Beispiel geschehen). Beabsichtigt ist vielmehr, die persönlichen Relationen einer Vielzahl von Akteuren in die Netzwerkmodellierung einzubeziehen, dazu gehören beispielsweise Drucker, Verleger oder Kolporteure. Die Betreffenden sollen in unserem Projekt dabei nicht länger über ihren sozialen Status oder ihre familiäre Herkunft, sondern vielmehr über ihr Wissen, die Möglichkeiten dieses zu generieren und andere Akteure an ihm partizipieren zu lassen definiert werden, dadurch rückt ihr

symbolisches Kapital, verstärkt durch die Theorie der Wissensnetzwerke,[34] in den Vordergrund. Denn die Journale der Aufklärung müssen zwingend als öffentliches Netzwerk verstanden werden, in dem sich inhaltliche und strukturelle (persönliche) Verbindungen überlagerten und ergänzten. Der von ihnen dadurch, teilweise neu definierte öffentliche Raum war die entscheidende Voraussetzung für die Zirkulation von Wissen während dieser Epoche.

Literatur

Anthes, Gary. 2010. „Topic models vs. unstructured data." In Commun. ACM 53 (12), 16–18. DOI: 10.1145/1859204.1859210.

Benkler, Yochai; Roberts, Hal; Faris, Robert; Solow-Niederman, Alicia; Etling, Bruce. 2015. „Social Mobilization and the Networked Public Sphere: Mapping the SOPA-PIPA Debate." In Political Communication 32 (4), 594–624.

Blei, David M. 2012. „Probabilistic Topic Models." In Commun. ACM 55 (4), 77–84. DOI: 10.1145/2133806.2133826.

Blei, David M.; Lafferty, John D. 2006. „Dynamic Topic Models." In Proceedings of the 23rd International Conference on Machine Learning. New York, ACM (ICML '06), 113–120.

Breuel, Thomas M. (2008). „TheOCRopusOpenSourceOCRSystem." In Document Recognition and Retrieval XV ed. Berrin A. Yanikoglu; Kathrin berkner, San Jose, 68150F-68150F-15.

Crane, Gregory. 2006. „What Do You Do with a Million Books." D-Lib Magazin (12).
Online: http://www.dlib.org/dlib/march06/crane/03crane.html, zuletzt geprüft am 19.12.2017.

Chang, Jonathan; Gerrish, Sean; Wang, Chong; Boyd-Graber, Jordan L.; Blei, David M. 2009. „Reading Tea Leaves: How Humans Interpret Topic Models." In Advances in Neural Information Processing Systems 22: Curran Associates, ed. Y. Bengio, D. Schuurmans, John D. Lafferty, C. K. I. Williams und A. Culotta, 288–296. Red Hock, NY: Curran.

Dahlgren, P. 2005. „The Internet, Public Spheres, and Political Communication. Dispersion and deliberation". In Political Communication 22, 147–162.

Eymer, Wilfried. 2004. Eymers Pseudonymen-Lexikon. Realnamen und Pseudonyme in der deutschen Literatur. Wiesbaden: Marix-Verlag.

Fruchterman, Thomas M. J.; Reingold, Edward M. 1991. „Graph drawing by force-directed placement." In Softw: Pract. Exper. 21 (11), 1129–1164.

---

[34] Vgl. Henningsson und Hanseth, „The essential".

Gierl, Martin. 2001. „Kompilation und die Produktion von Wissen im 18. Jahrhundert." In Die Praktiken der Gelehrsamkeit in der frühen Neuzeit, ed. Helmut Zedelmaier und Martin Mulsow, 63–94.

Henningsson, Stefan; Hanseth, Ole. 2011. „The essential dynamics of information infrastructures." In International Conference on Information Systems ICIS; 2011.3.

Hinneburg, Alexander; Preiss, Rico; Schröder, Renè. 2012. „TopicExplorer: Exploring Document Collections with Topic Models." In Machine Learning and Knowledge Discovery in Databases. European Conference, ECML PKDD 2012, ed. Peter A. Flach, 838-841, Bristol, UK.

Hinneburg, Alexander; Rosner, Frank; Pessler, Stefan; Oberländer, Christian. 2014. „Exploring Document Collections with Topic Frames." In Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, ed. Jianzhong Li. 2084-2086. New York.

Holzmann, Michael. 1906. Deutsches Pseudonymen-Lexikon. Wien und Leipzig: Akadem. Verlag.

Jacobi, Carina; van Atteveldt, Wouter; Welbers, Kasper. 2016. „Quantitative analysis of large amounts of journalistic texts using topic modelling." In Digital Journalism 4 (1), 89–106.

Jäger, Margarete; Jäger, Siegfried. 2007. Deutungskämpfe. Theorie und Praxis Kritischer Diskursanalyse. Wiesbaden: VS Verlag für Sozialwissenschaften.

Jurish, Bryan. 2016. „Canonicalizing Historical Text with CAB. Searching Linguistic Patterns in Large Text Corpora for Digital Humanities" Research July, 2016. Online verfügbar unter http://kaskade.dwds.de/~moocow/mirror/pubs/esu2016-jurish-cab.pdf, zuletzt geprüft am 04.01.2018.

Kataria, Saubrah; Mitra, Prasenjit; Caragae, Cornelia; Giles, C. Lee. 2011. „Context sensitive topic models for author influence in document networks." In Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence, IJCAI-11. Barcelona, Catalonia, Spain, 16 - 22 July 2011 ed. Toby Walsh, 2247-2280, Menlo Park, Calif.: AAAI Press/Internat.

Kleinen-von Königslöw, Katharina. 2010. Die Arenen-Integration nationaler Öffentlichkeiten. Der Fall der wiedervereinten deutschen Öffentlichkeit. Wiesbaden: VS Verlag für Sozialwissenschaften.

Lim, Kar Wai; Buntine, Wray; Chen, Changyou; Du Lan. 2016. „Nonparametric Bayesian topic modelling with the hierarchical Pitman–Yor processes." In International Journal of Approximate Reasoning 78 (Supplement C), 172–191. DOI: 10.1016/j.ijar.2016.07.007.

Link, Jürgen. 2013. „Diskurs, Interdiskurs, Kollektivsymbolik." In Zeitschrift für Diskursforschung (1). Online verfügbar unter http://www.content-select.com/index.php?id=bib_view&doi=10.3262/ZFD1301007. zuletzt geprüft am 03.05.2018.

Mimno, David; Blei, David. 2011. „Bayesian Checking for Topic Models." In Proceedings of the Conference on Empirical Methods in Natural Language Processing. Stroudsburg, 227-237. PA, USA: Association for Computational Linguistics (EMNLP '11).

Mladenova, Yana. 2016. Automatisierung der Identifikation von Autoren mit Pseudonymlexika am Beispiel der Halleschen Zeitschriften von 1688-1815. Bachelorarbeit. Martin-Luther Universität, Halle. Informatik. (Unveröffentlicht).

Papacharissi, Zizi. 2010. A private sphere. Democracy in a digital age. Cambridge: Polity (Digital media and society series).

Papilloud, Christian; Hinneburg, Alexander (2018): Qualitative Textanalyse mit Topic-Modellen. Eine Einführung für Sozialwissenschaftler. Wiesbaden: Springer Fachmedien Wiesbaden (Studienskripten zur Soziologie).

Parthey, Gustav. 1973. Die Mitarbeiter an Friedrich Nicolai's Allgemeiner Deutscher Bibliothek nach ihren Namen und Zeichen in zwei Registern geordnet. Ein Beitrag zur deutschen Literaturgeschichte. Hildesheim: Gerstenberg.

Riddell, Allen: „Text Analysis with Topic Models for the Humanities and Social Sciences." Online verfügbar unter https://de.dariah.eu/tatom/, zuletzt geprüft am 20.06.2017.

Röder, Michael; Both, Andreas; Hinneburg, Alexander. 2015. „Exploring the Space of Topic Coherence Measures." In Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, 399-408. New York, NY, USA: ACM (WSDM '15).

Sarasin, Philipp. 2011. „Was ist Wissensgeschichte?" In Internationales Archiv für Sozialgeschichte der deutschen Literatur : IASL 36 (1), 159–172.

Schmid, Helmut. 1994 „Probabilistic Part-of-Speech Tagging Using Decision Trees." In International Conference on New Methods in Language Processing, Manchester UK, 44–49.

Sievert, Carson; Shirley, Kenneth. 2014. „LDAvis. A method for visualizing and interpreting topics." In Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces, 63-70. Baltimore, Maryland, USA: Association for Computational Linguistics.

Springmann, Uwe. 2016. „OCR für alte Drucke". In Informatik Spektrum 39 (6), 459-462.

Stier, Sebastian; Bleier, Arnim; Lietz, Haiko; Strohmaier, Markus. 2018. „Election Campaigning on Social Media: Politicians, Audiences, and the Mediation of Political Communication on Facebook and Twitter." In Political Communication 35 (1), 50–74.

Stollwerk, Christoph. 2016. „Machbarkeitsstudien zu Einsatzmöglichkeiten von OCR-Software im Bereich „Alter Drucke" zur Vorbereitung einer vollständigen Digitalisierung deutscher Druckerzeugnisse zwischen 1500 und 1930." In DARIAH-DE working paper 6.

Tangherlini, Timothy R.; Leonard, Peter. 2013. „Trawling in the Sea of the Great Unread: Sub-corpus topic modeling and Humanities research.“ In Poetics 41 (6), 725–749. DOI: 10.1016/j.poetic.2013.08.002.

Weller, Emil. 1886. Lexicon pseudonymorum. Wörterbuch der Pseudonymen aller Zeiten und Völker oder Verzeichnis jener Autoren, die sich falscher Namen bedienten. Regenburg: Coppenrath.