Hochschule für Angewandte Wissenschaften Hamburg
Hamburg University of Applied Sciences

Masterarbeit

Nina Hälker

Teilautomatisierte Erstellung von Dossiers auf der Basis von Textmining-Verfahren

Fakultät Technik und Informatik
Studiendepartment Informatik

Faculty of Engineering and Computer Science
Department of Computer Science

Nina Hälker

Teilautomatisierte Erstellung von Dossiers auf der Basis von Textmining-Verfahren

Masterarbeit eingereicht im Rahmen der Masterprüfung

im Studiengang Next Media
am Department Informatik
der Fakultät Technik und Informatik
der Hochschule für Angewandte Wissenschaften Hamburg

Betreuender Prüfer: Prof. Dr. Kai von Luck
Zweitgutachter: Dr. Gerd Kamp

Eingereicht am: 27.11.2015

Nina Hälker

Thema der Arbeit

Teilautomatisierte Erstellung von Dossiers auf der Basis von Textmining-Verfahren

Stichworte

Dossier, Archiv, Kulturzeitschriften, Textmining

Kurzzusammenfassung

Redakteurinnen und Redakteure sehen sich täglich vor die Aufgabe gestellt, mit stetig wachsenden Archiven umgehen zu müssen. Diese MA-Abschlussarbeit widmet sich der Frage, wie Textmining-Verfahren, die teilautomatisierte Dossiers aus semi-strukturierten Archiven erstellen, dazu beitragen können, diese Aufgabe zu bewältigen. Nachdem eine Definition von Dossiers entwickelt worden ist, werden die Anforderungen an Dossiers in einem spezifischen Archiv – dem des europäischen Kulturzeitschriftennetzwerks Eurozine – vorgestellt. Auf Grundlage der Analyseergebnisse wird erprobt, wie dieses zentral gestellte Archiv durch verschiedene Verfahren erschlossen werden kann.

Nina Hälker

Title of the paper

The semi-automated creation of dossiers that relies on text-mining procedures

Keywords

dossier, archive, cultural journals, text mining

Abstract

This MA thesis is a preliminary investigation. It aims to support editors in their daily work who are faced with the problem of ever-growing archives. It focusses on the question how text-mining procedures can help to create semi-automatised dossiers from semi-structured archives. After developing a definition of „dossiers“, one particular archive will be analysed: the archive of the european netmagazine Eurozine. This analysis will show which demands need to be met when creating dossiers. The outcome of the analysis will be tested by investigating how accessible the archive will be when using different access procedures.

Inhaltsverzeichnis

1 Einleitung

> Es sind viele, aber die Anzahl der Artikel ist natürlich nur die eine Sache. Es ist die Menge an Text, die es schwer macht. [...] Es sind komplexe Themen und die Texte sind lang, lang, lang. Das bedeutet, dass eine manuelle Sichtung viel Zeit braucht. Und ich sehe nicht, wie Eurozine diese Arbeit noch länger leisten kann, wenn wir nicht den nächsten Schritt machen. Und wie tief man da hineingeht, wie sehr man die Ontologien oder semantischen Konstruktionen durchdringt, das ist eine Frage von Ressourcen. Aber wir müssen radikale Schritte in diese Richtung machen – so viel ist klar. (Tjark L., 13:47)

Mit der Einführung neuer Technologien im Medienbereich sind papiergepflegte Archive im Verlauf der vergangenen 20 Jahre sukzessive digitalisiert worden. Darüber hinaus sind mit dem Internet auch neue Arten von Archiven entstanden. Da viele Archive historisch gewachsen sind, sind sie nicht oder nur teilweise systematisch erschlossen. Mittels der Volltextsuche sind sie nun zwar besser erschließbar, aber die Digitalisierung allein kann das Problem der fehlenden systematischen Aufbereitung nicht lösen. Die Volltextsuche erlaubt nur Recherchen allgemeinerer Art. Differenziertere Abfragen müssen weiterhin in Form von Handarbeit vorgenommen werden.

Hier setzt die vorliegende Arbeit an. Sie untersucht am Beispiel des europäischen Kulturzeitschriftennetzwerks Eurozine, ob der Einsatz von Textmining als Ergänzung zur derzeit genutzten Volltextsuche ein geeigneter Schritt sein kann, die in Redaktionen vorhandenen Ressourcen effektiver zu nutzen. Bislang entscheiden RedakteurInnen[1] in einem durch Volltextsuche unterstützten, zumeist individuellen Sichtungsprozess über die Zusammenstellung von Dossiers. Das bedeutet für Redaktionen im Wesentlichen zwei Dinge: Erstens ist die Dossiererstellung sehr zeitintensiv, und zweitens bleibt das Wissen über den differenzierten Inhalt des Archivs bei den einzelnen RedakteurInnen. Eine Verbesserung und Arbeitserleichterung könnte darin bestehen, mithilfe von Textmining-Verfahren Empfehlungsstrukturen im Sinne der Bereitstellung potenziell passender Artikel zu entwickeln. Die Annahme dieser Arbeit ist,

[1] In dieser Arbeit wird in großen Teilen statt der hintereinander gesetzten weiblichen und männlichen Schreibweise das große Binnen-I verwendet.

dass durch solche Verfahren Ergebnisse effektiver geliefert werden können, als es die bisherige Arbeitsweise ermöglicht. Zudem könnten zukünftige Arbeiten zu Archiven an dieser Stelle ansetzen und das Wissen, das durch Empfehlungsstrukturen angesammelt wird, zur Entwicklung lernender Archiverschließungssysteme nutzen.

Im Folgenden wird zunächst mithilfe von Experteninterviews[2] herausgearbeitet, was typischerweise unter einem Dossier verstanden wird und welche Kriterien bei der Erstellung von Dossiers berücksichtigt werden müssen.

Um diese Kriterien an einem konkreten Beispiel überprüfen zu können, wird im nächsten Kapitel das Kulturzeitschriftennetzwerk Eurozine mit seiner spezifischen LeserInnenschaft und seinem konkreten Bedarf an die Erstellung von Dossiers vorgestellt. Der Fokus liegt dabei auf der Frage, was die bisher von Eurozine erstellten Dossiers charakterisiert. Herausgestellt wird dabei, welchen konkreten Fragestellungen man bei einer automatisierten Erschließung von semistrukturierten Textkorpi begegnet.

Im nächsten Schritt wird Textmining als Methode zur Erschließung und Analyse goßer textbasierter Datenbanken vorgestellt. Ergänzend werden einzelne Projekte, die Textmining-Verfahren einsetzen, exemplarisch vorgestellt.

Anschließend werden anhand der vorher entwickelten Kriterien Methoden zur Erschließung des englischsprachigen Teils des Eurozine-Archivs getestet. Alle eingesetzten Methoden verwenden die zum Werkzeugkasten des Textmining gehörenden Bag-of-Words-Ansätze. Die so entstandenen Parameter sind eingeflossen in die Arbeit von Marcel Schöneberg, der sich parallel zu und im Dialog mit dieser Arbeit den Möglichkeiten teilautomatisierter Dossiererstellung mittels komplexer Mining-Ansätze annäherte. Gemeinsam wurde der zentralen Frage nachgegangen, wie sich die semistrukturierten Daten des Eurozine-Archivs als einfach zu handhabende Ressource für den redaktionellen Alltag nutzen lassen können.

Die Arbeit versteht sich als vorbereitende bzw. Mittlerarbeit. Sie ist im Kontext eines interdisziplinären Dialogs mit Informatikern über die Schnittstellen von *Content* und *Technology* entstanden. Gegenstand dieser Arbeit ist daher die Erörterung der Vorbedingungen für die Entwicklung von Algorithmen, nicht aber die Entwicklung von Algorithmen selber.

[2] Der Begriff der Expertin / des Experten wird in Kapitel 2.2.1 präzisiert.

2 Dossiers

Derzeit beschäftigen sich viele Redaktionen mit der Frage, wie Zeitungs- und Zeitschriftenarchive durch Verwendung von Textmining-Verfahren aufbereitet, sogenannter „related content“ automatisiert zusammengestellt und LeserInnen angeboten werden kann. Die Vermarktung dieser quasi-neuen Produkte, die oftmals eine Zweitverwertung von Content sind, bedeutet für Verlage neben einer größeren Sichtbarkeit nicht zuletzt auch die Möglichkeit, zusätzliche Einnahmequellen zu schaffen.

Bislang werden Dossiers überwiegend redaktionell erstellt, was mit einem hohen zeitlichen und personellen Aufwand einhergeht und außerdem von den RedakteurInnen ein sehr differenziertes Wissen über das eigene Archiv erfordert. Einzelne Verlage haben sehr früh mit der Verschlagwortung ihrer Inhalte angefangen und spezialisierte Systeme entwickelt. Dazu gehören außer einigen größeren Zeitungen u.a. auch Nachrichtenagenturen. Sie benutzen ein gewachsenes, oftmals sehr differenziertes und auf den eigenen Content spezialisiertes Schlagwortsystem, bei dem jeder neue Artikel durch eine entsprechende Verschlagwortung such- und auffindbar ist. Andere, zumeist kleine Publikationen verfügen so gut wie gar nicht über ein systematisch erfasstes Archiv. Eine inhaltliche Strukturierung für ein Archiv nachträglich händisch vorzunehmen erfordert außer Zeit und Geld ein immenses Know-how über Archivierung und Verschlagwortung. Solange jedoch (vor allem) kleine Zeitschriften keinerlei Verschlagwortung ihrer Archive haben, läuft das angehäufte Wissen über die Inhalte des Archivs häufig in einer Person zusammen: der Redakteurin oder dem Redakteur, die/der am längsten in der Redaktion arbeitet. Im schlechtesten Fall geht deshalb bei einem Personalwechsel der Zugang zu den Archivinhalten verloren – was einem Gedächtnisverlust der Redaktion bzw. der Zeitschrift gleichkommt.

So fest der Begriff „Dossier“ im umgangssprachlichen Gebrauch zu sein scheint, so sehr verliert er an Konturen, wenn, wie in dieser Arbeit, versucht wird, daraus verbindliche Regeln für die Erstellung von Dossiers abzuleiten. Die Formulierung solcher Regeln könnte jedoch die Voraussetzung dafür zu schaffen, in einem weiteren Schritt zu präzisieren, welche Anforderungen Textminingverfahren erfüllen müssen, um teilautomatisiert Dossiers zu erstellen. Aus diesem Grund wird in diesem Kapitel das Konzept Dossier näher untersucht. Dazu werden

zunächst herkömmliche lexikalische Definitionen herangezogen. Dabei stellt sich rasch eine gewisse Unschärfe des Begriffs heraus, sodass in einem weiteren Schritt zur Empirie, in diesem Fall zur qualitativen Methode des Experteninterviews, gegriffen wird: Für diese Arbeit wurden Experteninterviews mit Fachleuten aus dem Medienbereich durchgeführt, transkribiert und ausgewertet.[1] Vorab wird die Methode vorgestellt und eine Arbeitshypothese formuliert.

Dabei stehen folgende Fragen im Vordergrund:

- Wie wird der Begriff verwendet?
- Welche unterschiedlichen Facetten und Entscheidungen werden bei der Erstellung von Dossiers berücksichtigt ?
- Wie sind Dossiers aufgebaut?
- Welche Rolle spielen Archive bei der Erstellung von Dossiers?

Darauf aufbauend werden die Experteninterviews ausgewertet: Nachdem für jedes Interview kurz skizziert wird, was die hauptsächlichen Aspekte des Gesprächs waren – um eine erste Verdichtung vorzunehmen und das Spektrum der Themen der Interviews zu vermitteln –, werden die einzelnen herausgearbeiteten Facetten der Themen „Dossier“ bzw. „Dossiererstellung“ nacheinander beschrieben. Gezeigt wird dabei, was aus ExpertInnensicht Dossiers kennzeichnet und welche Kriterien bei der Erstellung berücksichtigt werden. Im letzten Teil dieses Kapitels wird dann bilanziert, welchen Beitrag die gewonnenen Erkenntnsse für die weitere Arbeit bedeuten.

2.1 Allgemein gebräuchliche Definitionen von Dossiers

Der Begriff „Dossier“ kommt aus dem Französischen und bedeutet ‚Aktenbündel‘, ‚Sammelmappe‘, ‚Akte(n)‘. Gängige Sichtweisen auf Dossiers, die ebenfalls auf diese recht unspezifische Wortbedeutung eingehen, finden sich im Duden, bei Wikipedia und in anderen Lexika. Diese werden hier kurz skizziert. Geht man der Wortbedeutung genauer nach und befragt man ExpertInnen nach ihren Sichtweisen auf Dossiers, kommen, wie im weiteren Verlauf des Kapitels gezeigt wird, sehr viel differenziertere Definitionen und Aspekte zur Sprache.

[1] Der Vorbereitung der Experteninterviews dienten zudem sondierende Gespräche mit BibliothekarInnen und RedakteurInnen. Die Erkenntnisse dieser Gespräche berücksichtigend (vgl. Kap. 2.3), wurde anschließend ein leitfadengestützter Fragebogen für die Experteninterviews entwickelt.

Der Duden unterscheidet zwei Bedeutungen des Begriffs: Ein Dossier sei sowohl eine „umfänglichere Akte, in der alle zu einer Sache, einem Vorgang gehörenden Schriftstücke gesammelt sind“, als auch eine „(besonders in der Presse in Form einer [Sonder]beilage o.Ä.) Dokumentation zu einem bestimmten Thema“. (Duden [2013]) Als Synonyme für Dossier nennt der Duden darüber hinaus

> [...] Aktensammlung, Archivalien, Dokumente, Schriftstücke, Unterlagen, Vorgang; (bildungssprachlich) Faszikel, Konvolut. (Duden [2013])

Auf weitere Kriterien eines Dossiers geht die Definition von Wikipedia ein, der zufolge ein Dossier

> eine Sammlung von Dokumenten zu einem bestimmten Thema [ist]. Meist werden die Schriftstücke in einer festen Hülle zusammengefasst. Daneben steht der Begriff „Dossier“ einerseits für eine Kategorie von Zeitungsartikeln, wird andererseits im politisch-sozialen Bereich im Internet aber auch als Artikelsammlung verstanden. (Wikipedia [2014])

Spezifiziert wird diese Definition unter dem Begriff „Zeitungsdossier“:

> Der Begriff „Dossier“ wird auch für eine Kategorie von Zeitungsartikeln verwendet, in denen die Informationen aus Akten aufbereitet sind. Nicht zuletzt werden mit „Dossier“ gemeinsam veröffentlichte Bündelungen von mehreren Artikeln, Hintergrundberichten, Interviews, Porträts etc. in Zeitschriften zu einem Themenschwerpunkt bezeichnet. Bei Onlineausgaben von Zeitungen, wie Süddeutsche.de oder jungewelt.de, werden insbesondere Sammlungen alter, ehemals frei verfügbarer Onlineartikel, die jetzt käuflich zu erwerben sind, als „Dossier“ bezeichnet. (Wikipedia [2014])

Es klingt bereits an, wie breit der Begriff verwendet wird: Während das Dossier in Gestalt einer Akte auf Vollständigkeit abzielt, ist eine [Sonder-]Beilage eine meist lockere Zusammenstellung von Material zu einem Thema. Auch die Art der Zusammenstellung sowie Struktur und Gliederung der unterschiedlichen Arten von Dossiers dürften sich voneinander unterscheiden. Gleiches kann von den Quellen, die für eine Dossiererstellung hinzugezogen werden, vermutet werden.

In beiden Definitionen bleibt weitestgehend offen, wodurch sich die Zusammengehörigkeit von Artikeln zu einem Dossier auszeichnet. Im Wikipedia-Beitrag kommen zwar neben der möglichen Form eines Dossiers auch Aspekte wie die Erstellung und die mögliche Verwendung eines Dossiers zur Sprache, jedoch bleiben auch hier Details ungenannt.

Der Blick auf den deutschsprachigen Zeitungs- und Zeitschriftenmarkt zeigt, dass es analoge wie auch digitale Dossiers gibt. Als Dossiers werden dabei oftmals recht allgemein Artikel bezeichnet, die gebündelt zu einer spezifischen Fragestellung angeboten werden. Diese Dossiers können sowohl ausschließlich aktuelle Artikel beinhalten als auch eine Zusammenstellung von Artikeln sein, die über mehrere Jahre hinweg publiziert worden sind.

Die bereits erwähnte *Süddeutsche Zeitung* stellt zum Beispiel Online-Dossiers zusammen, mittels derer man sich gezielt über ein Thema informieren kann (Ebitsch [2014]). Ebenfalls den Titel „Dossier“ tragen ein regelmäßiges Buch, das die Wochenzeitung *Die Zeit* heraugibt, und eine Rubrik in der Wochenzeitung *Jungle World*. Beide erläutern die Verwendung des Begriffs jedoch nicht näher. Zeitschriften wie die deutsche Ausgabe der *Le Monde diplomatique* (*LMd*) und die monatlich erscheinenden *Blätter für deutsche und internationale Politik* (*Blätter*) veröffentlichen neben ihren Ausgaben Sondereditionen, die sie, als kuratierte Artikelsammlung zu einem bestimmten Themenkomplex, als Dossiers verkaufen. *Le Monde diplomatique* beschreibt ihr Online-Dossierangebot folgendermaßen: „In den Dossiers finden Sie ausgewählte LMd-Artikel, Karten und Grafiken über zeitlose Themen und aktuelle Konflikte.“ (LMd [2015a]) *Le Monde diplomatique* veröffentlicht zudem zweimal jährlich die gedruckte „Edition LMd“, eine Zusammenstellung bereits publizierter sowie neuer Artikel zu thematischen Schwerpunkten, die durch ein Editorial eingeleitet werden. (LMd [2015b]) Die *Blätter* bieten auf ihrer Website Themen-„Dossiers“ aus etwa 20 Artikeln als Artikelsammlungen an. (Blätter [2015]) Sowohl *LMd* als auch die *Blätter* praktizieren so eine Zweitverwertung ihres Contents: Die neue redaktionelle Zusammenstellung schafft einen Mehrwert für LeserInnen und für die Redaktion.

Deutlich wird, dass Dossiers – entsprechend der dargestellten Verwendung – einen verdichteten Zugang zu einem Thema ermöglichen sollen. Sie können sowohl einem überblicksartigen Einstieg als auch einer vertiefenden Recherche und Information dienen. Dossiers sind insbesondere bei Online-Magazinen (und -Archiven) eine gute Möglichkeit zur strukturierten und lesefreundlichen Erschließung eines Archivs.

Ein Dossier ist, gemäß der bisherigen Beschreibungen, das Ergebnis eines Selektions- und Filterprozesses: Eine bestimmte Anzahl an Artikeln wird entsprechend einer Fragestellung einem Archiv entnommen und zu einer Artikelsammlung zusammengestellt. Solch eine Zusammenstellung von Artikeln aus einem umfangreichen Archiv schafft einen gezielten Zugang zu und eine differenzierte Nutzung von Zeitungs- und Zeitschriftenarchiven.

Unbeantwortet bleiben bisher jedoch Fragen wie bspw. die folgenden:

- Welche Kriterien werden herangezogen, um eine Zugehörigkeit von Artikeln zueinander zu konstatieren?

- Welche Bedeutung kommt der Fragestellung eines Dossiers oder/und dem Archivbestand, also dem Textkorpus, zu, wenn eine Zugehörigkeit von Artikeln zu einem Dossier festgestellt werden soll?
- Lassen sich diese Kriterien gemäß Regeln benennen?

Um Antworten auf diese Fragen zu bekommen, wurde entschieden, sich mithilfe des „Experteninterviews“, einer qualitativen Methode der Sozialforschung, der Definition von Dossiers und den Aspekten der Erstellung von Dossiers detaillierter zu widmen.

2.2 Die Methode

2.2.1 Das Experteninterview

Das Experteninterview ist eine Methode der qualitativen Sozialforschung. Experteninterviews ermöglichen eine Analyse der „Strukturen und Strukturzusammenhänge des ExpertInnenwissens und -handelns“ (Bogner u. a. [2005], S. 76): Durch die Interviews werden die Situationsdefinition der ExpertInnen, ihre Strukturierung des Gegenstands und seine Bewertungen erfasst (Bogner u. a. [2005], S. 72). Im Fokus des Interviews steht nicht die Gesamtperson, sondern ihr organisatorischer und institutioneller Zusammenhang. Im besten Fall wird das Wissen, das vor der Durchführung der Interviews über eine Thematik bestand, durch das Erfahrungswissen (Hintergrundwissen und fachspezifisches Interesse), das die ExpertInnen auszeichnet, angereichert.

Wer gilt als Expertin/Experte? Die InterviewpartnerInnen werden beim Experteninterview aufgrund ihrer Expertise in einem bestimmten Bereich angefragt. Der ExpertInnenstatus ist ein „relationaler Status“ (Bogner u. a. [2005], S. 73), der temporär verliehen wird und sich explizit auf das Thema und die Fragestellung, die im Interview erörtert werden sollen, bezieht. Der Vorteil der Methode ist, dass die Interviewten selber „Teil des Handlungsfelds sind, das den Forschungsgegenstand ausmacht“ (Bogner u. a. [2005], S. 73). Für die vorliegende Arbeit sind dies Berufsgruppen, die mit dem Handlungs- und Themenfeld, in dem sich die Erstellung und Nutzung von Dossiers verorten lässt, vertraut sind, wie z. B. RedakteurInnen, JournalistInnen, BibliothekarInnen und ArchivarInnen.

Dem Experteninterview liegt kein standardisierter Fragebogen zugrunde. Vor dem vereinbarten Interviewtermin bekommen die ausgewählten GesprächspartnerInnen einen Leitfaden zugeschickt: Er soll Gelegenheit bieten, sich auf das Thema des Gesprächs vorzubereiten. Durch dieses Vorgehen soll sichergestellt werden, dass die Befragten möglichst umfassend von ihrer Expertise berichten und auf Nachfragen fachlich kompetent reagieren können. Die

Expertise und Einschätzung der ExpertInnen schafft durch dieses Vorgehen eine wertvolle und valide Grundlage für die weitere Arbeit.

Ziel der im Rahmen dieser Arbeit durchgeführten Experteninterviews war, die verschiedenen Aspekte der Fragestellung realitätsnah und detailliert zu erfassen. Das Erkenntnisinteresse der qualitativen Erhebung durch Experteninterviews bestand darin zu erfahren, welche Kriterien bei der Erstellung von Dossiers berücksichtigt bzw. welche Entscheidungen während der Erstellung von Dossiers getroffen werden. Die durch die Interviews gesammelten Informationen wurden im Anschluss verdichtet, zu Oberthemen zusammengeführt und zur Präzisierung der Definition hinzugezogen.

2.2.2 Planung und Durchführung

Insgesamt wurden sieben Interviews mit neun ExpertInnen geführt und transkribiert.[2] Der vorab erstellte Gesprächsleitfaden wurde den ExpertInnen zeitnah vor den Interviewterminen zugeschickt.[3] Die Interviews dauerten durchschnittlich 40 Minuten. Ein Interview beanspruchte deutlich mehr Zeit, da nach den Themen des Leitfadens noch über die Entstehung und Entwicklung von Eurozine gesprochen wurde. Die auf diese Weise gewonnenen Informationen sind in Kapitel 3 eingeflossen, in dem das europäische Kulturzeitschriftennetzwerk Eurozine vorgestellt wird.

Die GesprächspartnerInnen wurden in den Transkripten anonymisiert. Wo nötig, wurden Informationen, die Rückschlüsse auf eine konkrete Person zulassen könnten, zum Zweck der Anonymisierung verändert.

Alle befragten ExpertInnen befassen sich in ihrem beruflichen Alltag mit der Erstellung oder professionellen Rezeption von Dossiers.[4] Sie erstellen Dossiers, Schwerpunkthefte, Focal Points für KollegInnen, JournalistInnen, Studierende, LeserInnen. Die so erstellten Dossiers ermöglichen die Einarbeitung in ein Thema oder eine vertiefende Beschäftigung. Ein Teil der Befragten rezipiert Dossiers beruflich für die eigene journalistische oder redaktionelle Arbeit. Die ExpertInnen äußerten aufgrund ihrer unterschiedlichen beruflichen Herkunft und ihres Arbeitsumfeldes sowohl voneinander abweichende Definitionen als auch sehr differenzierte Kriterien für die Auswahl, den Umfang und die Sortierung von Artikeln eines Dossiers.

[2] Die Datenbasis besteht aus sieben Transkripten, da zwei Interviews mit jeweils zwei Personen gleichzeitig geführt wurden. Auf den Wunsch einzelner InterviewpartnerInnen sind die Transkripte nicht Teil des Anhangs dieser Arbeit oder anderweitig online einsehbar, werden aber bei Bedarf von der Verfasserin vorgelegt.

[3] Der Gesprächsleitfaden für die Experteninterviews findet sich im Anhang.

[4] Unter den Begriff Dossiers fallen dabei auch anverwandte Formate, wie Themen- oder Schwerpunkthefte, die Artikel zu einem Oberthema beinhalten. Solche veröffentlichen z. B. einige Partnerzeitschriften des im nächsten Kapitels vorgestellten Kulturzeitschriftennetzwerks Eurozine. Auch die bei Eurozine unter dem Titel „Focal Points“ zusammengestellten Artikelsammlungen fallen unter den Oberbegriff Dossier.

Die Interviewten waren im Einzelnen

- die leitende Bibliothekarin einer wissenschaftlichen Bibliothek (Eva A.),
- die Pressereferentin eines wissenschaftlichen Instituts (Christina R.),
- der leitende Redakteur einer Nachrichtenagentur (Samuel T.),
- der Dokumentar einer politischen Wochenzeitung (Andre B.),
- eine Redakteurin und ein Redakteur einer Partnerzeitschrift von Eurozine (Karen S. und Christian K.) sowie
- drei GründerInnen von Eurozine (Anke T., Tjark L. und Ralf C.).

Letztere sind mit den Zielen, der Entwicklung und der derzeitigen Situation des Kulturzeitschriftennetzwerks Eurozine sehr vertraut und arbeiten in Redaktionen der Partnerzeitschriften bzw. bei Eurozine selber. Einhergehend mit ihrer Eurozine-Expertise beschäftigen sich die GründerInnen des Netzwerks – wie einige der anderen Interviewten auch – seit langer Zeit mit dem Wandel der Medienwelt und seinen Auswirkungen nicht zuletzt auf die Rezeption von Nachrichten und Hintergrundberichten.

2.2.3 Auswertung

Die Auswertung der Interviews orientiert sich an der von Mayring [2010] beschriebenen qualitativen Inhaltsanalyse, deren Ziel die Zusammenfassung, die Explikation und die Strukturierung des erhobenen Materials mithilfe verschiedener Techniken ist (Mayring [2010], S. 65). Die Aufgaben, die die qualitative Analyse im vorliegenden Fall erbringen soll, sind Hypothesenfindung und Theoriebildung (vgl. Mayring [2010], S. 22). Das Anliegen ist, durch die Auswertung der Interviews die vorab getätigten Vermutungen bestätigt oder widerlegt zu sehen und die Ergebnisse der Analyse in Form theoretischer Ansätze in die Arbeit am Anwendungsfall einzubeziehen. Im Zentrum der Analyse stehen die Zusammenfassung des Materials und die induktive Kategorienbildung (vgl. Mayring [2010], S. 64ff.).

Im Anschluss an jedes Interview wurde eine Kurzzusammenfassung in Form eines Gedächtnisprotokolls angefertigt. So konnten für die Fragestellung der Arbeit zentrale, von den GesprächspartnerInnen zur Sprache gebrachte Aspekte festgehalten werden. Die Zusammenfassungen, die in Abschnitt 2.4 zu lesen sind, geben Aufschluss über die konkrete Praxis der Befragten sowie ihren Bedarf hinsichtlich Dossiers und Dossiererstellung. Der Vergleich der

Gedächtnisprotokolle ermöglichte außerdem, als ersten Verdichtungsschritt, eine gute überblicksartige Einschätzung hinsichtlich der für die Dossiererstellung von den ExpertInnen als relevant erachteten Aspekte.

Ergänzend zur Zusammenfassung der subjektiven Eindrücke der Gespräche durch die Autorin wurden Rückmeldungen von der Person, die die Transkripte angefertigt hat, eingeholt.[5] Beides bildete die Basis für eine erste Erstellung von Kategorien. Dabei zeigte sich z. B., dass die einzelnen Aspekte der Erstellung von Dossiers bzw. des Filterprozesses – die Fragestellung bzw. Eingrenzung eines Themas, das Vorwissen über die vermuteten oder bekannten LeserInnen etc. – jeweils für sich betrachtet wichtig sind. Beim Lesen der Transkripte wurden die Interviews dann an diesen Aspekten orientiert weiter verdichtet. Mittels eines Codesystems konnten einzelnen Abschnitten auf diese Weise Kategorien bzw. Oberbegriffe zugeordnet werden (z. B. „Definitionen“, „Zweck“ oder „Leserschaft“).

Durch diesen Schritt der Verdichtung (das Lesen und teilweise Paraphrasieren der Transkripte sowie das Zuordnen von Oberbegriffen) wurde es möglich, Überschneidungen, Mehrfachnennungen und Priorisierungen einzelner Aspekte in den Interviews zu erkennen. Als Nächstes wurden konkrete Zitate bzw. Paraphrasen aus den Interviews den Kategorien zugeordnet. Dadurch konnten die verschiedenen Definitionen, die die InterviewpartnerInnen für den Begriff Dossier gegeben hatten, gebündelt werden. Gleiches geschah mit den in den Gesprächen genannten Kriterien für ein „gutes“ Dossier und für Such- und Selektionskriterien. Zusammen können diese Aspekte Antworten darauf geben, wie aus einer großen Menge von Artikeln diejenigen herausgefiltert werden können, die für ein bestimmtes Dossier relevant erscheinen.

2.3 Hypothesenbildung mittels sondierender Gespräche

Um die Fragestellung für die Empirie zu präzisieren, wurden sondierende Gespräche geführt und mittels Notizen protokolliert.[6] Der Zweck war, in kurzen Gesprächen mit ExpertInnen eine grobe Einschätzung darüber zu gewinnen, welche Aspekte für die Erstellung von Dossiers – sowohl in redaktioneller Hinsicht als auch bezogen auf Automatisierungsmöglichkeiten – für relevant erachtet würden. Auf genau diese Aspekte sollte in den Interviews ein besonderes Augenmerk gelenkt werden.

Mehrfach genannte Stichworte in den sondierenden Gesprächen waren

[5] Vielen Dank an dieser Stelle an Therese Roth, die die Interviews in sagenhaftem Tempo transkribiert hat.

[6] Vielen Dank an die KollegInnen, die zu diesem Zweck den einen oder anderen Kaffee mit der Autorin getrunken haben.

- die Bedeutung des Kontexts von Dossiers und
- die Bedeutung verschiedener Kriterien bei der Erstellung von Dossiers.

Kontext

Hingewiesen wurde auf die Relevanz und die Heterogenität des Kontextes. So wurde zu bedenken gegeben, dass Archivtexte vergangener Jahr(zehnt)e nicht unbedingt mit heutigen Parametern les- oder bewertbar seien. Die Erstellung von Dossiers würde aus diesem Grund die unbedingte Berücksichtigung des thematischen und zeitlichen Kontexts erfordern. So sei bspw. oftmals die Vorgeschichte eines Artikels wichtig, um einen Artikel zu verstehen.

Kriterien des Erstellungsprozesses

Die Sinnhaftigkeit eines Dossiers entstehe, so die GesprächspartnerInnen, durch die Entschlüsselbarkeit der Zusammenstellung des Materials. Zwei entscheidende Aspekte seien dabei Nachvollziehbarkeit und Transparenz der Suchhistorie. Die Genese eines Dossiers könnte gleichsam als Code betrachtet werden, mit dem es entschlüsselt werden könne. Dafür müsse vor der Erstellung möglichst präzise geklärt werden, wer mit dem spezifischen Dossier angesprochen werden soll und was die inhaltlichen Algorithmen seien. Angefangen beim Auftrag, ein Dossier zu erstellen, über die Sondierung des zur Verfügung stehenden Materials bis hin zur Überprüfung des Materials anhand der gegebenen Themenstellung verlaufe der Suchprozess oft in mehreren Schleifen der Überprüfung von Themen (bzw. der Übereinstimmung mit einem Thema), der Korrektur der Hypothese und der Eingrenzung und Auswahl des Materials.

Die GesprächspartnerInnen bezweifelten, dass eine (teil-)automatisierte Dossiererstellung ihren redaktionellen Anforderungen genügen könnte. Ihre Annahme war, dass der Kontext und die Kriterien, die in die redaktionelle Erstellung von Dossiers immer einbezogen würden, durch eine Automatisierung praktisch reduziert würden.

Die Ergebnisse der sondierenden Gespräche und die vorab beschriebenen Verständnisse und Verwendungsarten von Dossiers ließen vermuten, dass der Begriff Dossier auch in weiteren Expertenkreisen sehr breit gefasst und sehr unterschiedlich verwendet wird. Angesichts des über diese Arbeit hinausgehenden Ansinnens, Möglichkeiten der teilautomatisierten Dossiererstellung zu eruieren, wurde der Gesprächsleitfaden für den empirischen Teil dahingehend ausgerichtet, möglichst differenziert zu erfragen, welche Entscheidungen bei der Erstellung von Dossiers zum Tragen kommen. Insofern war das leitende Interesse weniger, was genau ein Dossier auszeichnet, als vielmehr, wie Dossiers erstellt werden.

2.4 Zusammenfassung der Experteninterviews

Beim Lesen der transkribierten Interviews bestätigte sich die Annahme, dass eine Definition von Dossiers die differenzierte Beschäftigung mit verschiedenen Aspekten erfordern würde. Die in den sondierenden Gesprächen erwähnten Themen – die Bedeutung des Kontextes (von Archiven, Artikeln und von Suchanfragen) und die verschiedenen Kriterien für die Dossiererstellung – fanden sich in allen Interviews wieder.

Zentrale Punkte im Gespräch mit Eva A. waren die Themen Materialauswahl, Objektivität und Bewertung. Den Prozess der Erstellung einer Materialsammlung beschrieb Eva A. als Prozess: von einer groben Orientierung – beginnend mit dem Sammeln und Betrachten vieler Dokumente, um sich einen Überblick über ein Thema zu verschaffen – hin zum Eingrenzen der Ergebnisse, dem verstärkten Aussortieren und dem Beschränken auf wenige, explizit zum Thema gehörende Dokumente. Das Ergebnis sei dann im besten Fall eine übersichtliche Anzahl von Dokumenten. Dossiers sollten, so Eva A., einen guten Überblick zu einem Thema geben, alle entscheidenden Aspekte darstellen und entsprechend spezifische Dokumente bzw. Beiträge beinhalten. Die Zusammenstellung sollte möglichst objektiv und die Auswahl der Dokumente nicht bereits durch eine eigene Bewertung des Themas einseitig ausgerichtet sein. Bezogen auf die Zusammenstellung und Struktur eines Dossiers betonte Eva A. die Bedeutung seiner Bewertung bzw. „Einordnung“. Diese bezwecke, den LeserInnen ein schnelles Erfassen des Spektrums der Sammlung zu ermöglichen.

Ralf C. und Tjark L. beschrieben im Interview den Aspekt des Unabgeschlossenen, Erweiterbaren. Eine Art von Archiven, beispielsweise dem Archiv von Eurozine, sei die einer prinzipiell unabgeschlossenen, über die Zeit größer und umfangreicher werdenden Sammlung. Neue Artikel mit neuen Themen, Informationen, Hintergründen und Positionen kämen hinzu, andere Themen würden nicht weiter fortgeführt etc. Diese Art des Archivs sei ständig in Bewegung, was eine andere Art der Erschließung erfordere als bei geschlossenen Archiven mit unveränderlichem Umfang. Dies habe Auswirkungen auf die Erstellung von Dossiers. Beide Experten betonten, dass auch die Dossiers, die aus der beschriebenen Art von Archiven erstellt würden, prinzipiell offen sein sollten, damit sowohl Ergänzungen als auch Löschungen vorgenommen werden könnten. Denn die Bewegung des Materials habe u.a. zur Folge, dass Artikel über die Zeit hinweg an Relevanz verlieren wie auch gewinnen könnten und der Kontext, innerhalb dessen Themen rezipiert würden, sich durch gesellschaftspolitische Entwicklungen verändere.

Zentral im Gespräch mit Samuel T. war die Notwendigkeit einer Systematik und strengen Struktur eines Archivs: Die systematische Zuweisung von Stichworten zu jedem neuen Artikel

sei maßgeblich dafür, dass Artikel auch in einem großen Archiv über gezielte Suchanfragen schnell gefunden werden können. Eine Hierarchisierung von Stichworten entsprechend einer Ontologie sei so funktional und zielführend wie ein Barcode, mithilfe dessen es möglich sei, sich über die vorgegebene Struktur von der groben Einordnung hin zum spezifischen Artikel vorzuarbeiten.

Auch Christina R. kam im Interview auf Übersichtlichkeit und Eindeutigkeit zu sprechen. Veröffentlichungen der MitarbeiterInnen ihres Instituts ebenso wie Presseberichte mit Bezug zum Institut u.ä. würden systematisch verschlagwortet – vergleichbar mit der bereits erwähnten Stichwortvergabe. Dies gewährleiste die Auffindbarkeit der Publikationen zu einem späteren Zeitpunkt. Schlagworte würden nicht beliebig vergeben, sondern auf Basis der in der Schlagwortdatenbank des Instituts vorliegenden Begriffe – wobei die Liste ergänzbar ist. Die verschlagworteten Publikationen des Instituts würden themenabhängig jeweils einem oder mehreren Dossiers zugeordnet und auf der hauseigenen Website LeserInnen zugänglich gemacht. Die Erstellung der Dossiers sollte nach klaren, einfachen Regeln erfolgen. Die Themen der Dossiers sollten im besten Fall knapp formuliert sein – die Dossierthemen des Instituts bestünden überwiegend aus einem einzigen Begriff – und strahlten neben einem gewissen Grad an Pragmatismus Klarheit und Übersichtlichkeit aus.

Andre B. betonte im Gespräch, dass es „das“ Dossier nicht gebe, vielmehr unterschiedliche Arten von Dossiers existieren würden: Abhängig vom Verwendungszweck unterschieden sie sich sowohl hinsichtlich der Kriterien der Inhaltsauswahl als auch der Struktur, und zwar im Erstellungsprozess wie im Produkt. Der wichtigste Aspekt bei der Dossiererstellung sei der Dialog mit den RezipientInnen – vor allem, um die Frage beantworten zu können, was für ein Dossier erstellt werden soll. Sie seien diejenigen, die ihre Frage spezifizieren müssten. Ein gutes Dossier zeichne sich durch die Übersichtlichkeit und Nachvollziehbarkeit der Zusammenstellung der Dokumente aus (z. B. eine Vorstrukturierung in Interviews, Reportagen, Texten von einer bestimmten Person, das Markieren wichtiger Textpassagen etc.). Auch eine Einordnung oder ein kurzer Kommentar im Sinne eines kommentierten Inhaltsverzeichnisses gehört für Andre B. zu einem guten Dossier – damit Erkenntnisse aus der Recherche den RezipientInnen nicht verborgen blieben und sie sich die Arbeit, die sich einE DokumentarIn oder RechercheurIn schon gemacht hat, nicht selber machen müssten.

Für Karen S. und Christian K. war eine wichtige Frage, welche Möglichkeiten digitale Dossiers gegenüber analogen Dossiers zur Verfügung stehen. Dies könnten u.U. bessere wie auch vielfältig nutzbarere LeserInnenführungen sein, aber auch die Möglichkeit, schneller von einer horizontalen Themenerfassung zu den vertikalen, vertiefenden Themenangeboten zu gelangen. Für beide kam bei der Frage des Kontexts, in dem Dossiers rezipiert werden, das Thema

Bildung zur Sprache. Die Erstellung von Dossiers geschehe nie im kontextfreien oder luftleeren Raum, sondern sei eng mit den jeweiligen Voraussetzungen und dem vorhandenem Vorwissen verknüpft.

2.5 Definition und Aspekte der Dossiererstellung aus Sicht der ExpertInnen

Durch die Auswertung der Experteninterviews bestätigte sich die Annahme, dass die einzelnen Kriterien und Aspekte, die bei der Dossiererstellung eine Rolle spielen, vom Kontext der Dossiererstellung abhängen und sich gegenseitig bedingen. Den ExpertInnen zufolge bedeutet das z. B., dass ohne die Kenntnis des Nutzungszwecks eines Dossiers der Umfang und die Struktur (bspw. Inhaltsverzeichnis, Priorisierungen von Teilen des Materials etc.) nicht zufriedenstellend bestimmt werden können. Zudem ist die Fragestellung bzw. die thematische Eingrenzung eines Dossiers stark von der prospektiven LeserInnenschaft bzw. Zielgruppe (konkreter: deren Interessen, Zeit, Vorwissen etc.) abhängig.

Welche Bedeutung dies für die Planung einer teilautomatisierten Dossierstellung hat, soll später erörtert werden. Um der Komplexität der verschiedenen Aspekte der Dossiererstellung Rechnung zu tragen, werden in den folgenden Abschnitten zunächst verschiedene Aspekte als isolierte beschrieben. Im Anschluss daran wird zusammenfassend bilanziert, was aus dem Dargestellten hinsichtlich der Planung einer teilautomatisierten Dossiererstellung zu schlussfolgern ist und welche Aspekte dabei berücksichtigt werden müssen.

2.5.1 Was ist ein Dossier?

Entsprechend den Definitionen und den Merkmalen bei den auf dem Zeitungsmarkt existierenden Dossiers beschreiben alle ExpertInnen Dossiers als Materialzusammenstellung zu einem bestimmten Thema bzw. einer Fragestellung. Umfang und Spezialisiertheit können sich dabei allerdings stark unterscheiden. So sind u.a. die folgenden, sich voneinander unterscheidenden Dossierformate für die ExpertInnen denkbar:

- Ein Dossier ist ein Metaformat, das ein Thema auf einen Blick erfassen lässt. (Samuel T., 18:25)
- Ein Dossier ist eine Materialzusammenstellung, die zu einem bestimmten Thema oder zu einer bestimmten Fragestellung einen guten Überblick geben sollte (Eva A., 4:30, und Andre B., 1:06).

- In seiner einfachsten Form ist ein Dossier als eine „qualifizierte Liste“ denbkar: „Man nimmt einen Vorgang wie ‚Der Bürgerkrieg in Syrien‘ und sagt: So, jetzt gehe ich mal durch die Datenbank und stelle erstmal alles zusammen, was wir in der Zeit dazu gemacht haben“. (Samuel T., 1:30) Aufwändiger gestaltet wird ein Dossier ggfs. zu einem analytischen, redaktionell gewichteten Themendossier, in das durch die Reihenfolge des Inhalts sogar eine erzählerischer Perspektive integriert werden kann. (Samuel T., 2:21)

- Die loseste Form eines Dossier wurde als „feste Papphülle “ beschrieben – „weil drin liegen lauter Fetzen. Und damit die Fetzen nicht durch die Gegend fliegen, muss man sie eben in eine feste Hülle“ legen. (Ralf C.,17:14)

Diese Auflistung von Paraphrasen und Zitaten der Erhebung vermittelt, wie breit der Begriff Dossier gefasst wird. Demnach können Dossiers von sehr einfach bis sehr aufwändig kuratiert sein, linear (alphabetisch, chronlologisch etc.) oder konzentrisch aufgebaut sein oder völlig andere Gliederungsarten verfolgen. Sie können nur einige wenige Artikel oder Archivquellen enthalten oder aber sehr umfangreich sein. Mit den knappen Worten eines Experten lässt sich zusammenfassen, dass die Experteninterviews bestätigen, was die Definitionen aus Lexika bereits vermuten ließen: dass der Begriff Dossier ein „relativ fuzzy Konzept“ ist. (Ralf C., 15:45)

2.5.2 Zweck und Funktion von Dossiers

Der Zweck von Dossiers wird – ähnlich den Definitionen – sehr verschieden bestimmt und praktiziert. Er kann darin bestehen, bereits veröffentlichte Artikel zu bündeln und so einer Zweitverwertung zuzuführen, er kann aber auch darin bestehen, eine neue Veröffentlichung vorzubereiten – ein Dossier also, das Recherchematerial bereitstellt. Dossiers, die die kompletten Quellen beinhalten, haben häufig eine andere Funktion als Dossiers, die vornehmlich aus kurzen Teasertexten oder Links bestehen. Während erstere einer detaillierten Recherche dienen, beinhalten letztere die Möglichkeit einer solchen Recherche, bezwecken aber zunächst, einen Überblick über mögliches Material zu geben. Während manche Dossiers als Forschungs- oder als „Entscheidergrundlage“ zusammengestellt werden, bezwecken andere, einen allgemeinen Überblick zu einem Thema zu geben.

Im ersten wie auch im zweiten Fall sollen Dossiers

- eine „Gatekeeper-Funktion“ erfüllen (Karen S., 14:02), also einen (neuen) Zugang zum Thema ermöglichen,

- „Welt strukturieren“ (Christian K., 59:58) und

- LeserInnen ermöglichen, die zentralen Argumentationsstränge (eines Themas) nachzuvollziehen (Eva A., 10:55).

Die ExpertInnen wiesen – auch angesichts der Spannbreite möglicher Verwendungszwecke und Funktionen – darauf hin, dass der Zweck wiederum von der Fragestellung für ein Dossier abhängt. Diese wiederum ist eng mit der Leserschaft verknüpft. Ein Dossier, das einer Politikerin eine Entscheidungsgrundlage bieten soll, um gut vorbereitet in eine Verhandlung über ein geplantes Großprojekt zu gehen, wird eine gänzlich andere Zusammenstellung von Artikeln und Dokumenten beinhalten, als ein Dossier, das eine Nachbarschaftsinitiative erstellt hat, die die Planung eines Großprojekts kritisiert und sich dagegen organisiert. Wiederum anders als diese beiden Dossiers sähe das Dossier eines Journalisten aus, der eine Reportage über das Großprojekt und die zugehörigen Debatten vorbereitet. Die Verschiedenheit der Dossiers bezöge sich in diesem Fall abgesehen von der unterschiedlichen Funktion auch auf den Inhalt, die Länge und die Art der Zusammenstellung, also die innere Chronologie eines Dossiers.

Der Zweck bzw. die Funktion von Dossiers müssen zudem auf verschiedenen Ebenen betrachtet werden. Um ein Beispiel zu nennen: Während der Zweck von Dossiers aus Sicht einer Redaktion in der Zweitverwertung von veröffentlichten Artikeln bestehen kann, sehen LeserInnen deren Funktion möglicherweise darin, ein konkretes Thema in einer Form dargeboten zu bekommen, die Struktur und Übersichtlichkeit bietet – neben weiteren möglichen Aspekten wie Aktualität oder Vertiefung in ein Thema.

2.5.3 Die diverse LeserInnenschaft

Dem Umstand folgend, dass die ProduzentInnen von Dossiers nicht zugleich die RezipientInnen sind, formulierten die ExpertInnen, dass im besten Fall vor der Erstellung ein Wissen um die späteren RezipientInnen und ihre Bedürfnisse für die Dossiererstellung existiert. Im besten Fall (aber in der Realität im seltensten Fall) sollte ein Dialog mit den prospektiven NutzerInnen stattfinden – z. B. über den Verwendungszweck des Dossiers, den Umfang und die zeitliche oder thematische Ausrichtung.

Für Andre B. gleicht dieses Vorgehen einem „Pingpong-Spiel“, bei dem diejenigen, die ein Dossier zu einem Thema anfordern, zunächst genauer zum Thema befragt werden (was ist schon bekannt, was ist neu, wie tief soll ins Thema eingestiegen werden, sind bestimmte Facetten eines Themas wichtiger als andere?). Dabei sollte auch der Umfang des Dossiers vorab festgelegt werden.

2.5.4 Struktur und Umfang von Dossiers

Dossiers sind, so die ExpertInnensicht, für gewöhnlich absichtsvoll sortiert und nicht zufällig zusammengeworfen. Häufig sind Dossiers chronologisch oder thematisch sortiert. Die Bündelung von Artikeln zu einem Dossier durch vorgegebene Selektionskriterien, welche Artikel aufgenommen werden (und welche nicht), führt im besten Fall zu einer ersten Strukturierung des Dossiers. Ähnliches kann von Zeitungen behauptet werden, die jeden Tag erneut nicht nur Inhalte produzieren und gebündelt anbieten, sondern dabei eine Kuratierung der Inhalte erzeugen.

Der Aufbau von Dossiers wurde in den meisten Interviews nicht nur hinsichtlich der Frage thematisiert, ob und welche unterschiedlichen Formate in einem Dossier gebündelt werden (können), sondern auch bezüglich der Frage, ob zum Aufbau eines Dossiers auch eine Einleitung (Editorial) oder Bewertung gehört. Die Einordnung einer Zusammenstellung durch einen redaktionell eigens für diesen Zweck erstellten Beitrag wird sowohl als strukturelle wie auch als inhaltliche Navigation durch ein Dossier verwendet. So werden Entscheidungen über die Auswahl der Artikel nachvollziehbar. Auch kann auf den Zweck eines Dossiers hingewiesen und vermerkt werden, ob ein Dossier eher einen Überblick vermitteln soll oder spezieller auf einzelne Aspekte eines Themas eingeht. Insbesondere redaktionell gewichtete, thematisch fokussierte Dossiers, die (durch den Aufbau) eine erzählerische Perspektive einnehmen, erfordern eine komplexere analytische Kompetenz und Herangehensweise an die Zusammenstellung von Dossiers.

Neben den genannten Möglichkeiten tragen den ExpertInnen zufolge auch andere Aspekte zur Strukturierung von Dossiers bei: Ob sie

- in „schlanker“ Form als analoge oder digitale Listen, z. B. als Linksammlung, bestehen,
- mit Teasern zu jedem Beitrag angereichert sind oder
- die vollständigen Dokumente beinhalten,

ist die Folge der Entscheidung für einen speziellen Umfang und einen speziellen Zugang, da die drei Varianten einen unterschiedlich schnellen und umfassenden Grad an Informiertheit ermöglichen.

Als praktizierte Navigationshilfen wurden ein beschreibender Titel, die Nennung des Themas und die Einbettung eines Inhaltsverzeichnis genannt, die zudem der Einordnung dienen und Übersichtlichkeit schaffen. Diese und andere Aspekte einer Navigationshilfe und NutzerInnenfreundlichkeit befriedigen LeserInnenbedürfnisse nach Effizienz und Relevanz (Samuel

T., 32:22) und geben eine Orientierung, mithilfe der LeserInnen erkennen können, auf welchem Weg sie schnell zum Ziel kommen bzw. welche Teile eines Dossiers für sie relevant sind (Samuel T., 35:41).

Ein Aspekt, der ebenfalls bei der Erstellung von Dossiers berücksichtigt werden sollte (aber noch kaum realisiert wurde), sind Wahlmöglichkeiten für LeserInnen: Beispielsweise könnten LeserInnen Tools zur Verfügung gestellt bekommen, mittels derer sie das Thema eines Dossier individuell weiter eingrenzen oder selber weiter auffächern können. Konkret könnte das bedeuten, als LeserIn selber Inhalte zu filtern, Fragestellungen zu präzisieren etc. (Karen S., 38:51) Hier werden allerdings bereits Fragen behandelt, die nicht nur die Struktur, sondern auch wieder den Zweck, die Leserschaft etc. mit einbeziehen.

Entsprechend ihres Verwendungszwecks haben Dossiers sehr unterschiedliche Umfänge. Auch müssen dabei offene, also erweiterbare Dossiers von geschlossenen, also im Umfang begrenzten Dossiers, unterschieden werden. Mit analogen Dossiers wird zudem oft anders gearbeitet als mit digitalen, und mit langen Dossiers anders umgegangen als mit kurzen.

2.5.5 Signaturen und Verschlagwortung

Verschlagwortung, Ressortzuordnungen und die Zuordnung zu Signaturen dienen in einem Archiv der Vorsortierung des Materials. So entstehen (virtuelle) Ordner, in denen eine Sammlung von Artikeln zu einem Oberthema abgelegt ist (Andre B., 10:33) und die die Erstellung von Dossiers erheblich unterstützen. DokumentarInnen, ArchivarInnen und BibliothekarInnen legen alle Artikel, die für ein Thema relevant sind, unter der gleichen Signatur ab. Signaturen in Archiven, Dokumentationen und Bibliotheken können insofern analog zu Schlagworten als Grundlage für die Erstellung von Dossiers herangezogen werden.

2.5.6 Von der Fragestellung zum Inhalt

Um zu entscheiden, was in ein Dossier aufgenommen wird, bedarf es eines Themas bzw. einer konkreten Fragestellung. Das Thema eines Dossiers kann sehr offen gewählt sein oder aber eng gefasst. Oft können erst durch die Eingrenzung eines Themas (beispielsweise „Demokratie“) und die zeitliche wie thematische Kontextualisierung durch weitere Begriffe (z. B. „Transition“ oder „Marktwirtschaft“) bestimmte Aspekte bei der Materialauswahl fokussiert werden. Durch die Einordnung eines Themas wird es möglich, die Relevanz von Artikeln einzuschätzen. So ist auch die Qualität eines Dossiers zu einem wesentlichen Teil von der Sinnhaftigkeit der Zusammenstellung der Inhalte abhängig, welche neben dem Thema in vielen Fällen erst in Abhängigkeit von der LeserInnenschaft und dem Zweck des Dossiers zu bestimmen ist.

Das Spektrum möglicher Inhalte eines Dossiers zeigt sich in der folgenden Auflistung aus den Experteninterviews. Ein Dossier sollte demnach enthalten:

- „die wichtigsten Schlüsselartikel zu einem bestimmten Thema“ – ohne Anspruch auf Vollständigkeit (Ralf C., 3:32) – oder aber, so ein anderer Experte, inklusive dem Anspruch auf Vollständigkeit: jeden Artikel, der zum Thema erschienen ist, „weil, es kann ja sein, dass in dem siebzehnten Artikel zum gleichen Thema irgendeine Wendung drin ist, die interessant ist“ (Andre B., 34:03),
- unterschiedliche Positionen und Aspekte, um auf diese Weise „ein breites Feld, lange Debatten knackig und fokussiert“ darzustellen (Eva A., 6:03),
- eine Chronologie (Samuel T., 18:37),
- nicht zu viele Details (Karen S., 13.35),
- verdichtete Information (Christian K., 5:51),
- offizielle, historische Dokumente, ministeriale Beschlüsse oder Dekrete, „gewissermaßen als Hintergrundinformationen zu den Artikeln“ (Ralf C., 12:13),
- einen ergänzenden, einordnenden Text: „eine kurze Bewertung, Einordnung und vielleicht auch Positionierung zur Materialsammlung (Eva A., 4:52).

Die Materialbasis

Basis einer Dossiererstellung ist der Pool bzw. das Datenset, aus dem Artikel und anderes Material gesichtet und ausgewählt werden sollen. Klar sein muss zunächst: Handelt es sich um das Archiv einer Zeitung, um den Bestand einer Bibliothek oder den eines Archivs? Oder handelt es sich um die eigenen Publikationen einer Einrichtung, die thematisch zusammengefasst werden sollen, um NutzerInnen den Zugang zu dem Material zu vereinfachen? In der vorliegenden Arbeit handelt es sich um eine wiederum andere Grundlage: Das Eurozine-Archiv ist ein Artikel-Archiv, das eine Auswahl der Publikationen der am Netzwerk beteiligten Zeitschriften umfasst. Es ist prinzipiell unvollständig, mehrsprachig, thematisch divers und semistrukturiert.

Vollständigkeit versus selektive Auswahl

Aus der Themenstellung eines Dossiers sollte hervorgehen, ob und wie vollständig die Zusammenstellung des Materials sein soll. Den ExpertInnen zufolge zeichnen sich Dossiers dadurch

aus, dass sie immer selektiv sind und nicht beanspruchen, vollständig zu sein. Die Auswahl der Inhalte eines Dossiers sollte nach vorab festgelegten Kriterien erfolgen. Die Selektivität geht in der Konsequenz im besten Fall mit einer guten Qualität einher.

Dossiers sollten frei sein von Redundanzen und sich auf das Wichtigste beschränken. „Das Wichtigste“ wiederum ergäbe sich, so die ExpertInnen, durch die Verknüpfung von Fragestellung, Zweck und LeserInnenschaft. Eine Gefahr bei der Auswahl für ein Dossier bestünde darin, viel zu viel Material aufzunehmen (Andre B., 19:36). Bei der Erstellung von Dossiers sei daher Strenge und Konzentration gefordert. Wichtige Inhalte eines Dossiers bestünden z. B. in einem historischen Abriss zu einem Thema, besonderen Entscheidungen, beteiligten Personen, Parteien etc. Dossiers sollten „Wegmarken“ eines Themas abbilden.

Eng verknüpft mit der Frage nach der Vollständigkeit oder Selektivität eines Dossiers ist die Frage, ob ein Dossier als offenes, also erweiterbares Dossier geplant ist oder als geschlossen betrachtet wird.

Dossiers, die wie bei Christina R. aus Publikationen erstellt werden, die die hauseigenen MitarbeiterInnen verfasst haben, sind, ähnlich den Focal Points von Eurozine, auf die später noch eingegangen wird, erweiterbare Dossiers: Neue Beiträge eines Archivs können oftmals schon bestehenden Dossiers zugeordnet werden. Diese Form des Dossiers wird meistens unspezifischer verwendet als Dossiers, die für eine individuelle Recherche erstellt werden.

Die Schritte der Inhaltsauswahl

Um Dossiers zu erstellen, muss entweder individuell für jedes Dossier oder für eine Gruppe von Dossiers bestimmt werden, worin die Aufgabe genau besteht: „Wie kann ich sie in Einzelaufgaben unterteilen, was ist die Reihenfolge, wie die Aufgaben abgearbeitet werden müssen?“ (Christian K., 46:39)

Wenn das Thema und der Zweck eines zu erstellenden Dossiers feststehen, beginnt die Recherche: Dabei gehören die Erhebung des Stands der Forschung und des Bestands (im Archiv) zu den ersten Schritten im Prozess der Inhaltsauswahl. „Man hat in der Bibliothek recherchiert, was es gibt zu einem bestimmten Thema. [...] Man hat in Bibliografien geschaut. [...] Man hat sich Literatur bestellt. Man hat geguckt, was es so gibt.“ Ein Aspekt bei diesem Vorgehen kann darin bestehen, dass „man den Überblick nicht wirklich oder nur mit großem Zeitaufwand gewinnen“ kann. Von Vorteil ist dabei, wenn das zur Verfügung stehende Material bzw. das Archiv bereits aufbereitet ist und durch Verschlagwortung, Signaturen, die Einteilung in Rubriken oder die Möglichkeit, eine Volltextsuche mit einzubeziehen, besser und ggfs. schneller passende Treffer gefunden werden. Den ExpertInnen zufolge werden häufig erst im Verlauf der Dossiererstellung scharfe Kriterien entwickelt, mittels derer es gelingt, „die relevanten

Sachen auszusieben und nicht in Suchresultaten zu ersticken“. Gut aufbereitete, strukturierte Archive sind dabei eine hilfreiche Unterstüzung. Bei nicht-verschlagworteten Archiven und unzureichend strukturierten Daten stellt sich die Frage, wie diese Sortierung im Nachhinein stattfinden kann. Darauf erste Antworten zu geben ist das Interesse dieser Arbeit.

2.5.7 Bilanz

Die im vorangegangenen Kapitel erarbeiteten Antworten auf die Frage, was Dossiers kennzeichnet und welche Kriterien für die Erstellung von Dossiers zu berücksichtigen sind, haben bestätigt, dass ExpertInnen mit einem breiten Spektrum möglicher Arten von Dossiers arbeiten. Gezeigt werden konnte weiterhin, dass es Schnittstellen zwischen den verschiedenen Formen von Dossiers gibt:

Dossiers sind eine kuratierte Auswahl aus einer größeren Menge an Material. Zur Erstellung von Dossiers gehört immer eine Entscheidung darüber, was dazu gehört und was nicht. Entschieden wird anhand von Nähe- bzw. Distanz-Kriterien, mittels derer das Verhältnis von Texten zueinander bestimmt wird. Als Dossier wird eine Sammlung dessen zusammengestellt, was als relevant für eine gegebene Fragestellung betrachtet wird.

Bezogen auf die Planung einer Teilautomatisierung für die Erstellung von Dossiers ergeben sich aus diesen Erkenntnissen die folgenden Konsequenzen:

- Die Erstellung von Dossiers findet unter den jeweils individuellen Bedingungen eines konkreten Archivs und anhand einer konkreten Aufgabenstellung statt.
- Es reicht für diese Arbeit daher nicht aus, ein generelles Verfahren zur Teilautomatisierung zu planen. Stattdessen bedarf es eines konkreten Archivs, an dem exemplarisch über Algorithmen für eine Teilautomatisierung nachgedacht werden kann.
- Die Konzepte zur Teilautomatisierung, die in einem solchen Test an einem exemplarischen Archiv gute Ergebnisse erzielen, müssen im Anschluss daran auf ihre Übertragbarkeit auf andere Archive getestet werden.
- Auf diese Weise kann perspektivisch die Entwicklung allgemein gültiger Algorithmen für eine teilautomatisierte Dossiererstellung erfolgen.

Im folgenden Kapitel wird das Zeitschriftennetzwerk Eurozine vorgestellt, dessen Archiv den Anwendungsfall dieser Arbeit bildet. Die Beschreibung nicht nur der Datenbasis des Archivs, sondern auch des weiteren Rahmens (dem Netzwerk, den beteiligten Zeitschriften, der LeserInnenschaft, dem Zweck der thematischen Zusammenstellung von Artikeln etc.)

bezweckt, hinsichtlich der geplanten teilautomatisierten Dossiererstellung konkrete Überlegungen für speziell dieses Archiv formulieren zu können.

3 Das europäische Kulturzeitschriftennetzwerk Eurozine

In diesem Kapitel werden das Archiv des europäischen Kulturzeitschriftennetzwerks Eurozine sowie das Netzwerk selber vorgestellt.[1] Eurozine veröffentlicht regelmäßig „Focal Points"– thematisch fokussierte Artikelsammlungen bzw. Dossiers aus durchschnittlich 20 bis 60 Artikeln verschiedener Zeitschriften des Netzwerks. Wie viele über die Jahre gewachsene Archive steht Eurozine nun der Herausforderung gegenüber, trotz der Größe und dem weiteren Wachsen des Archivs einen inhaltlichen Überblick über das Vorhandene zu behalten, obwohl es bislang keinerlei Verschlagwortung der Artikel gibt. Überlegungen hinsichtlich Möglichkeiten teilautomatisierter Dossiererstellung könnten für Eurozine eine Chance bieten, das Artikelarchiv systematisch aufzubereiten und die redaktionelle Arbeit dadurch zu entlasten.

Ausgehend von der fachlichen Ausgangssituation der vorliegenden Arbeit (Contentperspektive) wird zunächst die Außensicht von www.eurozine.com gezeigt. Anschließend wird die innere Struktur von www.eurozine.com beschrieben. Letztere dient als Hinführung zur technologischen Perspektive und erste Auseinandersetzung mit den Voraussetzungen und Anforderungen des Archivs an eine teilautomatisierte Dossiererstellung.

3.1 Kulturzeitschriften

Zeitschriften lassen sich grob in zwei Sparten unterteilen: in Publikumszeitschriften, welche die auflagen- und umsatzstärksten Zeitschriften sind,[2] und Fachzeitschriften, die titelreicher, aber in deutlich niedrigeren Auflagen erscheinen. Noelle-Neumann u. a. [1994] weisen darauf hin, dass es eine unüberschaubare Anzahl an „periodischen Publikationen [gibt], die mit dem

[1] An dieser Stelle möchte ich mich bei Eurozine, dem Vorstand und den Eurozine-MitarbeiterInnen für die Erlaubnis bedanken, mit dem Eurozine-Archiv in diesem Rahmen arbeiten zu dürfen. Außerdem bedanke ich mich ganz herzlich für das mir entgegengebrachte Interesse am Thema dieser Arbeit sowie an der Bereitschaft und Zeit, mir in Interviews von der Geschichte und der aktuellen Praxis des Netzwerks zu berichten. Die Informationen zu Eurozine in diesem Kapitel basieren größtenteils auf den Gesprächen, die ich im Dezember 2014 mit den GründerInnen des Netzwerks geführt habe. Die Gespräche wurden aufgezeichnet und transkribiert. Bei Bedarf kann Einblick in die Transkripte genommen werden.

[2] Publikumszeitschriften wiederum lassen sich in General-Interest- und Special-Interest-Zeitschriften einteilen.

Sammelbegriff Zeitschrift bezeichnet werden. [...] Ihre Gesamtzahl lässt sich kaum vollständig ermitteln. Bei den meisten handelt es sich um Fachzeitschriften." (S. 401)

Lehnert [2012] verortet Kulturzeitschriften, und damit auch Eurozine, im Feld der Fachzeitschriften, die ihm zufolge Magazine sind,

> die weniger einen unterhaltenden, als vielmehr einen informativen Charakter haben. Sie werden in der Regel nur von Fachpublikum gelesen, das eine entsprechende Ausbildung hat und häufig auch beruflich in einem bestimmten Fachbereich tätig ist. Beispiele für solche Fachzeitschriften sind Computer- oder Kulturzeitschriften sowie Magazine mit wissenschaftlichem Hintergrund.

Der überwiegende Teil der Eurozine zugehörigen Kulturzeitschriften erscheint in niedrigen Auflagen von weniger als 1.000 bis hin zu 10.000 Exemplaren.

Zur Geschichte von Kulturzeitschriften in Deutschland schreibt Noelle-Neumann u. a. [1994]:

> Der Mangel an Zeitungen während der Lizenzperiode und zumal das Bedürfnis, den politischen Umbruch nach dem Ende der nationalsozialistischen Herrschaft publizistisch zu reflektieren, bildeten die wichtigsten Voraussetzungen für die Gründung politischer und kulturpolitischer Zeitschriften in der unmittelbaren Nachkriegszeit. Unter den Kulturzeitschriften lassen sich wiederum verschiedene Typen unterscheiden: vergleichsweise politikfreie Organe (z. B. Merian, art), „klassische" Rundschauzeitschriften (Die Neue Rundschau, Merkur), Organe der Linksintellektuellen (Das Argument, Kursbuch, Freibeuter). [...] Solche politischen oder kulturpolitischen Zeitschriften sprechen lediglich kleine intellektuelle Minderheiten an. (S. 407f.)

Ähnliche definiert Anke T., Mitgründerin von Eurozine, die Charakteristika von Kulturzeitschriften:

> Kulturzeitschriften sind traditionell nicht auflagenstark und bedienen ein ganz bestimmtes Segment der intellektuellen Debatte in unterschiedlichen kulturellen Kontexten. Sie haben eine lange Tradition. Sie sind nicht elitär, aber bedienen auch nicht Tagesdebatten oder kurzfristigere Interessen, sondern sind so ein bisschen Luxus. Ein Austausch über Ideen, Vorstellungen – immer auch ein bisschen intermediär – und auch, natürlich, zwischen verschiedenen Bereichen intellektueller Aktivitäten angesiedelt. Kulturzeitschriften sind einer der wenigen Orte, an denen Kunstschaffende mit wissenschaftlichen Intellektuellen [...] in einen Austausch zueinander treten. Also eigentlich eher eine langatmige, gar nicht so aktuell zeitgebundene Form der Kommunikation. (Anke T., 00:26)

Zeitschriften unterscheiden sich von Zeitungen, da die Zeitschrift

> im Gegensatz zur Zeitung nicht zur Aktualität verpflichtet ist. Sie kann also auch Themen aufgreifen, die schon seit Jahren, Jahrzehnten oder Jahrhunderten vorbei sind und zu denen es auch nicht zwangsläufig neue Erkenntnisse gibt. [...] Nicht selten sind die einzelnen Artikel hier auch ressortübergreifend gestaltet und stehen allesamt unter einem gemeinsamen Titelthema [...]. (Lehnert [2012])

Das Gabler-Wirtschaftslexikon formuliert es in ähnlicher Weise:

> Das Merkmal der Aktualität (Gegenwartsbezug) ist bei Zeitschriften nur unter bes. Bedingungen nachweisbar; grundsätzlich sind Zeitschriften nicht primär auf Aktualität ausgerichtet. (Sjurts [2012])

Manche Presseorgane sind in ihrem „äußeren Erscheinungsbild den Zeitungen ähnlich, doch sind diese nach ihren publizistischen Merkmalen genau genommen den Zeitschriften zuzurechnen. Sie dienen weniger der aktuellen Berichterstattung als der Hintergrundinformation und der tagesübergreifenden Meinungsbildung“. (Noelle-Neumann u. a. [1994], S. 400) Anders als bei Zeitungen, die überwiegend Artikel (Kurzmeldungen, Nachrichten und Berichte) veröffentlichen, bei denen die wichtigsten Informationen am Anfang eines Beitrags genannt werden, veröffentlichen Zeitschriften mit Reportagen, Essays, Analysen, Porträts u.a. sehr unterschiedliche Formate.

Bezüglich der Konzeption und Zielsetzung von Eurozine als europäischem Netzwerk weist Anke T. darauf hin, dass Debatten, die über Kulturzeitschriften nachvollziehbar sind, „besonders dann spannend werden, wenn sie nicht nur disziplinäre und andere Bereichsgrenzen wie mediale Grenzen überschreiten, sondern auch nationale Räume überschreiten“. (Anke T., 00:26)

3.2 Der kulturelle und strukturelle Kontext von Eurozine

Eurozine wurde 1998 gegründet, um einen vormals losen Zusammenschluss europäischer Kulturzeitschriften in ein kontinuierlich arbeitendes Netzwerk zu überführen.[3] Eurozine hat Partnerzeitschriften in nahezu allen europäischen Ländern und damit ein besonderes Potenzial in europäischen Debatten.

[3] Seit 1983 existierte, als Vorläufer von Eurozine, ein informelles Netzwerk westeuropäischer Kulturzeitschriften, zu dem nach dem Zusammenbruch des Ostblocks und dem Ende des Kalten Kriegs zunehmend auch osteuropäische Zeitschriften gestoßen waren.

3.2.1 Idee und Status quo

Die Idee von Eurozine wird in der Selbstbeschreibung auf der Website folgendermaßen formuliert:

> By providing a Europe-wide overview of current themes and discussions, Eurozine offers a rich source of information for an international readership and facilitates communication and exchange between the journals themselves. By presenting the best articles from its partners and their countries, as well as original texts on the most pressing issues of our times, Eurozine opens up a new space for transnational debate. (Eurozine [2014a])

Abbildung 3.1: Politische Europakarte (http://geodressing.de/freie-karten/politische-europakarte, CC-BY 3.0 Unported)

Derzeit gehören mehr als 80 Partnerzeitschriften aus 32 Ländern dem non-profit-organisierten Netzwerk an. Zentraler Teil ist die Website www.eurozine.com, die von dem in Wien ansässigen Redaktionsbüro betrieben wird. Das Büro kümmert sich auch um die Kommunikation

mit den Partnerzeitschriften, die Übersetzung von Beiträgen, die Vorbereitung von Konferenzen und die Finanzierung des Projekts durch Fördermittel.

Wichtige Funktionen nehmen außer der Redaktion ein Editorial Board und ein Advisory Board ein: Das Editorial Board besteht aus vier (alle paar Jahre wechselnden) RedakteurInnen der Partnerzeitschriften, die die Belange der Zeitschriften vertreten, Vorschläge für Konferenz- und Schwerpunktthemen erarbeiten etc. Dem Advisory Board, das beratende Funktion hat, gehören u. a. JournalistInnen und WissenschaftlerInnen an. Beide Gremien gewährleisten neben der zentralen Eurozine-Redaktion und den Redaktionen der Partnerzeitschriften die Qualität der Artikel und Debatten, die Eurozine der Öffentlichkeit präsentiert.

Eurozine veröffentlicht auf seiner Website regelmäßig Artikel der Partnerzeitschriften, die sich inhaltlich in den Bereichen Kultur, Politik und Literatur verorten. Die Artikel erscheinen in der Regel in der Originalsprache und, wenn es sich dabei nicht um Englisch handelt, ergänzend oft in englischer Übersetzung. Die Artikel sind, bedingt durch die unterschiedliche Ausrichtung der Partnerzeitschriften, von sehr unterschiedlichem Format – sowohl hinsichtlich ihrer Länge als auch bezogen auf die Art und Komplexität, wie Themen aufgegriffen werden. Sie sind eine Auswahl aktueller zeitgenössischer europäischer Debatten – zu Erinnerungskulturen, zu Integrations- und Exklusionsprozessen u.a. Dabei werden kaum tagesaktuelle Themen aufgegriffen:

> Eurozine setzt Themen, die manchmal erst zwei Jahre später breiter diskutiert werden. [...] Eurozine fasst aber auch Debatten zusammen und greift Themen auf, die selber schon eine sehr lange Geschichte haben. (Tjark L., 11:47)

Die Website von Eurozine, die in ihrer Aufmachung einem Online-Magazin entspricht, fungierte bereits in den Anfängen des Netzwerks als multifunktionales „Tool“:[4]

- Sie eröffnete den Zeitschriften mit den bis heute als Teil der Seite bestehenden Zeitschriftenprofilen die Möglichkeit einer eigenen Webpräsenz zur weltweiten Verbreitung ihrer Inhalte.[5] Dazu gehört auch, die Inhaltsverzeichnisse der Zeitschriftenausgaben zu veröffentlichen – in den Originalsprachen und oftmals ergänzt um eine englischsprachige Übersetzung.

- Die Veröffentlichung von Artikeln auf der Seite vereinfachte für die Redaktionen der Partnerzeitschriften, insbesondere durch die Übersetzung vieler Artikel ins Englische, die Möglichkeit, Artikel miteinander auszutauschen.

[4] Der Fokus dieser Arbeit liegt auf dem englischsprachigen Teil des Artikelarchivs von Eurozine.

[5] Ende der 1990er Jahre waren eigene Webseiten für (Kultur-)Zeitschriften noch die Ausnahme und das Konzept Eurozines avantgardistisch. Erst vier Jahre zuvor, im Herbst 1994, ging das Nachrichtenmagazin *Der Spiegel* als weltweit erste Zeitschrift online. Vgl. http://de.wikipedia.org/wiki/Internet-Zeitung.

- Eurozine schuf durch den kostenlosen Zugang zu den Artikeln eine insgesamt größere Öffentlichkeit für die Zeitschriften und das Netzwerk.

Die Absicht von Eurozine, einen europaweiten Austausch über kultur- und gesellschaftspolitische Debatten zu etablieren, wurde durch die Möglichkeiten des Internets unterstützt. Darüber hinaus wird die Sichtbarkeit der transnationalen Debatte durch den zeitschriftenübergreifenden Austausch von Artikeln und die Übersetzung von Artikeln in andere Sprachen befördert: Themen und Debatten sollen ebenso wie die AutorInnen, die darüber schreiben, kulturelle, intellektuelle und nicht zuletzt geografische Grenzen überwinden.[6] Menschen mit Interesse an kultur- und gesellschaftspolitischen europäischen Debatten können diese in kuratierter Form bei Eurozine verfolgen.

3.2.2 Abbildung und Herstellung einer europäischen Debatte

Das Eurozine-Archiv bildet stichprobenhaft Debatten ab, die in den verschiedenen Zeitschriften geführt werden. Es ist kein Spiegel aller Themen und Artikel der Partnerzeitschriften. Eurozine zeigt aber durch die redaktionelle Zusammenstellung der Themen – beispielsweise zu Focal Points, aber auch zu Konferenzthemen etc. – ein Spektrum europäischer Themen in Form einer europäischen Stimmenvielfalt, die unterschiedliche Positionen beinhaltet. Eurozine lädt dazu ein, das zu lesen, was in unterschiedlichen Gegenden von Europa diskutiert wird. Im besten Fall gelingt es Eurozine auf diese Weise, Europa als eine Vielfalt europäischer Stimmen und Perspektiven abzubilden. Gerade in einer Zeit, in der in Europa ein gravierender Rechtsruck auszumachen ist, ermöglicht ein Netzwerk wie Eurozine, das Zeitschriften verbindet, die eine liberale Idee Europas verfolgen, aufgeklärte Positionen über europäische Themen aus verschiedenen nationalen Perspektiven zu verfolgen.

Die Qualität der Inhalte soll durch die Auswahl der Zeitschriften, den Dialog über die angebotenen Artikel und die Kuratierung der Inhalte für die Website (z. B. durch die Zusammenstellung von Focal Points) gewährleistet werden. Das breite Angebot ermöglicht, sich darüber zu informieren, welche Themen wo und wie diskutiert werden. Das zentrale Anliegen von Eurozine besteht darin, möglichst viele Artikel in englischer Sprache zu veröffentlichen, um sie so einer möglichst großen Leserschaft zugänglich zu machen. Auf diese Weise wird ein Dialog ermöglicht, der deutlich erschwert würde, wenn die Artikel nur in den jeweiligen Sprachen der einzelnen Zeitschriften publiziert werden würden.

Eine Gründerin von Eurozine formuliert die Chance, die in einem Netzwerk wie Eurozine liegt:

[6] Vgl. Interview mit Tjark L., 11:47

> Wirklich spannend wird es, wenn man die kulturelle Debatte in einem Raum in Beziehung zu einer kulturellen Debatte in einem ganz anderen Denkklima und Denkraum setzen kann – historisch, aber auch mental und auch von der Art kreativer Produktionen in ganz unterschiedlichen geografischen Bereichen. [...] Das Spannende an Eurozine ist, dass man durch die Transferleistung, die dieser Austausch zwischen den Kulturzeitschriften erbringt, und auch durch Übersetzungsleistungen, über dieses Portal tatsächlich nachvollziehen kann, was in der Slowakei diskutiert wird oder in Schweden oder in Räumen, die nicht so im Fokus sind und die auch durch sprachliche Grenzen nicht so zur Verfügung stehen [...]. Es geht nicht so sehr um Zeitgeistgeschichten und Public Intellectuals, die sich eine große Prominenz erobern konnten, sondern es geht auch um Verborgeneres. (Anke T., 0:26)

Ist es durch den europäischen Kulturdebatten-Aggregator Eurozine (das Netzwerk und die Website bzw. die Artikel) also möglich, Einschätzungen zu Europa zu gewinnen? Die Antwort muss wohl lauten: ja und nein. Auf der einen Seite bildet Eurozine ab, über welche Themen in welchem Land wie diskutiert wird. Auf der anderen Seite sind die Themen und Diskussionen, die bei Eurozine nachvollzogen werden können, nur ein Ausschnitt dessen, was die Zeitschriften insgesamt veröffentlichen. Und die am Netzwerk beteiligten Zeitschriften wiederum bilden nur einen Ausschnitt der Publikationen und Debatten in einzelnen Ländern ab.

3.2.3 Die Partnerzeitschriften

Die Rubrik „Partner journals“ auf der Eurozine-Website enthält, unter Nennung des Erscheinungslandes, eine alphabetische Liste aller am Netzwerk beteiligten Zeitschriften: Für jede Partnerzeitschrift findet sich dort in englischer Sprache eine Kurzportrait, die Themen und Inhaltsverzeichnisse der erschienenen Heftausgaben sowie ein Link zur zeitschrifteneigenen Website. Auch die Artikel, die bei Eurozine veröffentlicht wurden, sind auf den Partnerseiten verlinkt. Die Inhaltsverzeichnisse der Hefte gibt es sowohl in der Originalsprache als auch in englischer Übersetzung. Gleiches gilt für viele Artikel. Einige Zeitschriften veröffentlichen ergänzend zu den Inhaltsverzeichnissen der Hefte regelmäßig Abstracts zu den Artikeln der neuen Ausgaben. LeserInnen können auf diese Weise nicht nur nachvollziehen, aus welchen Ländern und Bereichen (z. B. Kunst, Politik, Literatur) die Zeitschriften kommen und welche Schwerpunkte sie verfolgen, sondern sich auch erschließen, welche Themen gerade in einem Land von Interesse sind, wie sie kontextualisiert und rezipiert werden (Interview mit Anke T., 7:13).

Die inhaltliche Ausrichtung der Partnerzeitschriften ist sehr unterschiedlich: Sie reicht von Literaturzeitschriften über politische Magazine bis zu Zeitschriften, die sich mit spezielleren Themen befassen, wie z. B. die österreichische Zeitschrift *derive*, deren Schwerpunkt auf stadtsoziologischen und politischen Diskussionen liegt. Die Eurozinepartner, die sich alle als Kulturzeitschriften begreifen, eint das Interesse an transnationalen Debatten und einer „translation of cultures“ (vgl. Eurozine [2014a]).

Anke T. ergänzt, die Partnerzeitschriften würden sich nicht nur dadurch unterscheiden,

> dass sie in unterschiedlichen nationalen Kontexten entstehen, sondern dass sie auch unterschiedliche Begriffe haben von sich selbst und auch eine fachliche Breite repräsentieren. [...] [D]ie Vielfalt [der Zeitschriften] ist nicht nur eine Vielfalt kultureller und geografischer, nationaler Räume, sondern auch eine Vielfalt des Zugangs, kulturellen Zugangs. (Anke T., 10:00)

Wichtiges Kriterium bei der Auswahl und Zusammenstellung der Partnerzeitschriften ist eine möglichst ausgewogene geographische Verteilung der Zeitschriften, um zu vermeiden, dass eine Region, z. B. Nordeuropa, die Debatten bestimmt.[7] Neue Partnerzeitschriften kommen zu Eurozine, indem sie entweder von sich aus Interesse bekunden, oder gezielt von Eurozine angesprochen werden.

3.3 Die Website – Aufbau der Seite, Focal Points und Artikel

3.3.1 Der Aufbau der Seite

Auf der Startseite von Eurozine werden unter der Rubrik „Headlines“ die neuesten Artikel angekündigt: Ein Schlagwort (beispielsweise „Rights“, „Debate“, „Urban politics“), der Titel des Artikels, der Name der AutorInnen, ein mehrzeiliger Teasertext und das Veröffentlichungsdatum und die Informationen, ergänzt um einen Link am Ende jedes Teasers, der zum vollständigen Artikel führt. Thematisch passende Fotos bzw. Banner ergänzen die Teaser der Contentspalte.

Anders als vermutet werden könnte, werden die Teaser der Artikel nicht automatisch, sondern redaktionell erstellt. So sind z. B. die in Kapitälchen an den Anfang der Teaser gesetzten Schlagworte (International Politics, Essay, Art and Politics, Literature, Social Movements, Gender etc.) nicht den einzelnen Artikeldatensätzen zugeordnet. Auch die Schlagworte werden nicht einer festgelegten und erweiterbaren Schlagwortliste entnommen, sondern mehr

[7] Die finanziellen und personellen Beschränkungen vieler Zeitschriften setzen diesem Anspruch Grenzen: Für viele nordeuropäische Zeitschriften ist es weitaus einfacher, stabile Produktionsbedingungen und Ressourcen für eine regelmäßige redaktionelle Arbeit zu garantieren, als für einige Zeitschriften aus Südeuropa.

oder weniger unsystematisch nach individueller Entscheidung eingesetzt. Aus diesem Grund ist es nicht möglich, systematisch nach Artikeln zu suchen, die das Schlagwort „Essay“ oder „Politics“ zugewiesen bekommen haben. Die Zuweisung existiert nur auf der Webseite, ist aber nicht mit den Artikeldatensätzen in der Datenbank verknüpft.

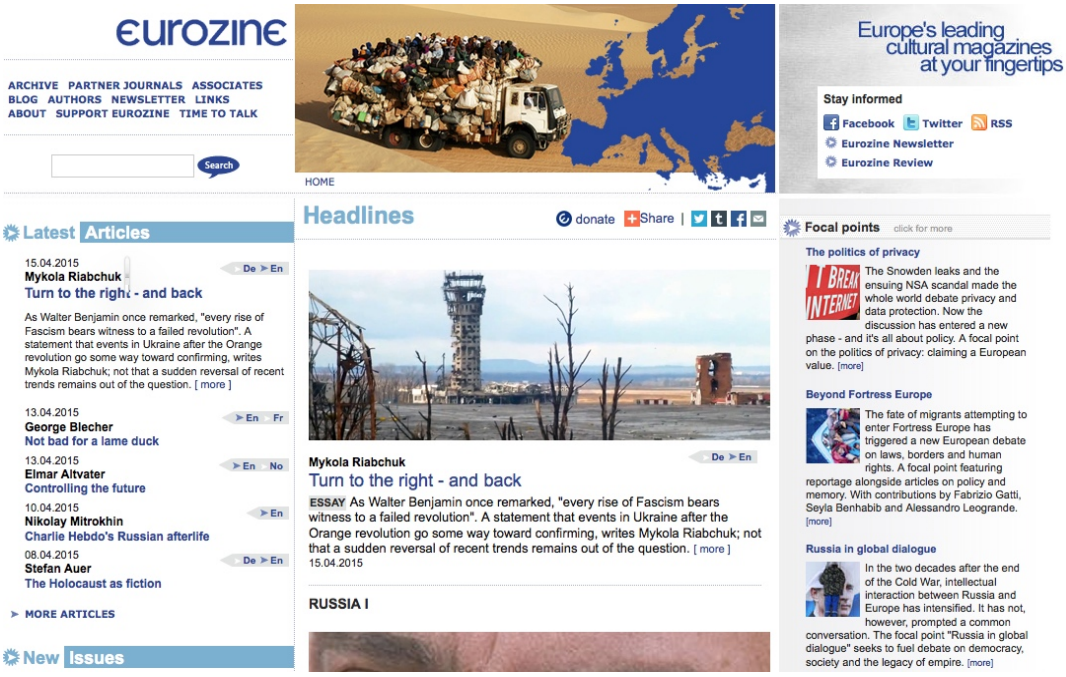

Abbildung 3.2: Die Startseite von www.eurozine.com (Stand: 16.04.2015)

Die linke und die rechte Spalte der Seite beinhalten außer der Seiten-Navigation („Archive“, „About Eurozine“, „Partner journals“, „Associates“ etc.) noch weitere Rubriken. Die Sortierung der Artikel in den verschiedenen Rubriken folgt unterschiedlichen Sortierkriterien: Während „Latest articles“, „New issues“ und „New reviews“ chronologisch sortiert sind und automatisiert ausgegeben werden, erfolgt die Zusammenstellung und Sortierung der Artikel in der Rubrik „Focal Points“ redaktionell, das heißt „händisch“.

Die Zugehörigkeit der Artikel zu Zeitschriften ist über die Partnersection-Seiten nachvollziehbar. Bis auf die Zuordnung von Artikeln zu den Partnerzeitschriften, die automatisiert stattfindet, sind alle diese Zuordnungen Ergebnis redaktioneller Entscheidungen. Das bedeutet neben dem zeitlichen Aufwand für die Redaktion, dass die Artikel nicht systematisch wiederfindbar sind.

Auch die Inhalte anderer Rubriken sind redaktionell zusammengestellt: „Editor’s choice“ präsentiert von der Redaktion empfohlene Beiträge; „Time to talk“, „Debate series“ und „Con-

ferences“ weisen auf Veranstaltungen und Berichte hin, und unter der Rubrik „Literature“ finden sich Artikel mit Bezug zu zeitgenössischer Literatur. Die Rubrik „Popular tags“ präsentiert, in der Optik einer Tagcloud, Schlagworte zu zentralen Themen und Debatten, zu denen Eurozine Artikel veröffentlicht hat. Auch die Tagcloud ist redaktionell erstellt.

Es gibt eine Volltextsuche, über die nach Artikeln gesucht werden kann. Die Treffer können über ein Dropdown-Menü nach den Feldern AutorInnenname, Erscheinungsjahr und Sprache der Artikel weiter eingegrenzt werden.

Der zweiwöchentlich erscheinende „Eurozine Review“ und der einmal pro Monat erscheinende „Eurozine Newsletter“ informieren über die neuesten Heftausgaben und thematischen Schwerpunkte der Partnerzeitschriften. Editorials in den Reviews, häufig mit tagespolitischen Bezügen, geben einen kurzen Überblick über die neuen Hefte, bevor sie einzeln vorgestellt werden. Der Newsletter bewirbt regelmäßig einen besonders erwähnenswerten Artikel und weist auf neu erschienene Artikel hin (Eurozine [2014b]). Der Eurozine-Review wie auch der Newsletter bieten die Möglichkeit, thematische Verknüpfungen zwischen unterschiedlichen Partnerzeitschriften vorzunehmen. Ähnlich den Focal Points, auf die im Abschnitt über die Artikel und am Ende des Dossier-Kapitels genauer eingegangen wird, zeichnen Review und Newsletter facettenartig eine Landkarte aktueller europäischer kultur- und gesellschaftspolitischer Debatten nach.

3.3.2 Focal Points

Die Rubrik „Focal Points“ bündelt Artikel unterschiedlicher Zeitschriften, AutorInnen und Erscheinungsjahre und stellt unterschiedliche Sichtweisen eines Themas länderübergreifend dar. So werden verschiedene Facetten eines Themas abgebildet und vertieft sowie unterschiedliche Zugänge zu einem Thema geschaffen. Eingeleitet wird jeder Focal Point durch ein Editorial, das außer einem Überblick gezielte Verweise auf einzelne Artikel beinhaltet.[8] Es handelt sich bei den Focal Points neben der Zweitverwertung der von bereits in den Partnerzeitschriften erschienen Artikeln auch um deren neue Kontextualisierung, durch die sich LeserInnen ein Thema strukturiert erschließen können.

[8] Die Editorials spielen, anders als die einzelnen Artikel der Focal Points, für den weiteren Verlauf dieser Arbeit keine Rolle.

Abbildung 3.3: Ausschnitt aus dem Focal Point „The ends of democracy“

Ein Mitgründer von Eurozine beschreibt Focal Points am Beispielthema „Europäische Integration“ wie folgt:

> Irgendwann hast du die türkische Perspektive auf die europäische Integration dabei, dazu den britischen und den deutschen Blick, und die schwedische Sichtweise auf die europäische Integration noch dazu. Und das nur, weil Eurozine Partner in all diesen Ländern hat. (Tjark L., 18:04)

Legt man die im vorangegangenen Kapitel erarbeiteten Definitionen für Dossiers zugrunde, sind Focal Points als Dossiers zu betrachten. So bildet z. B. der Focal Point „The ends of democracy“, in dem europäische Debatten zur Zukunft und den Grenzen der Demokratie(n) wiedergegeben werden, einen Einblick in eine Debatte zu einem eingegrenzten Thema. Ähnlich sieht man das bei anderen Focal Points, wie z. B. „The politics of privacy“ oder „Ukraine in focus“.

Allerdings handelt es sich bei den Focal Points um eine jeweils selektive Zusammenstellung von Artikeln, die sich von manchen anderen Archiven unterscheidet: Das Eurozine-Archiv bildet nur einen Ausschnitt der von den Partnerzeitschriften veröffentlichten Artikel ab und ist nicht vollständig: Eurozine selber verfügt nicht über die kompletten Archive der Partnerzeitschriften, und selbst die Artikel, die auf der Website von Eurozine stehen, sind nicht immer auf Englisch zu bekommen.

Das Archiv wächst permanent, da die Partnerzeitschriften Eurozine regelmäßig neue Artikel anbieten. Eine Verschlagwortung der Artikel oder eine Vorsortierung zu inhatlichen Rubriken gibt es für das Eurozine-Archiv bislang allerdings nicht. Alle Focal Points werden unter extrem hohen zeitlichen Aufwand redaktionell kuratiert. Die Möglichkeit, durch den Einsatz von automatisierten Verfahren teilautomatisiert Dossiers zu erstellen, böte eine starke Entlastung für die Redaktion. Gleichzeitig stellt sich jedoch die Frage, ob die bereits im Dossierkapitel beschriebene Komplexität, die bei der Erstellung von Dossiers eine Rolle spielt, von automatisierten Verfahren übernommen werden könnte.

In jedem Fall wäre für Redaktionen u.U. ein Vorschlagsystem attraktiv: Durch den Einsatz automatisierter Verfahren könnte evtl. der redaktionell nicht erfassbare große Umfang des gesamten Eurozine-Archivs in sehr kurzer Zeit auf die für eine bestimmte Themenstellung relevanten Artikel durchsucht werden. Die Treffer könnten gleichsam und als Vorschlagsliste für die Erstellung eines neuen Focal Points herangezogen werden. Die Aufgabe der RedakteurInnen bliebe in diesem Fall, die automatisch generierte Vorauswahl weiter einzugrenzen, also eine Qualitätsprüfung der Treffer vorzunehmen und so Artikel zu einem sinnhaften Dossier zusammenzustellen.

Im folgenden Abschnitt werden die Artikel des Archivs, auf das bei der Erstellung von Focal Points zurückgegriffen wird, hinsichtlich Struktur, typischer Strukturelemente und ihrer Verwertungsmöglichkeit für automatisierte Verfahren betrachtet.

3.3.3 Die Artikel

Die meisten Artikel bei Eurozine sind Beiträge, die bereits in den Partnerzeitschriften (meist als gedruckte Version) veröffentlicht und für eine Zweitveröffentlichung freigegeben worden

sind. Die übliche Praxis im Netzwerk ist, dass Zeitschriften aus jeder Ausgabe einen Artikel für diese Art der Veröffentlichung weitergeben. Neben den Artikeln der Partnerzeitschriften gibt es Beiträge, die Eurozine als Netzwerk anfragt bzw. einwirbt. Bei diesen Artikeln handelt es sich entweder um Beiträge, die aus Vorträgen (z. B. bei den jährlich stattfindenden Eurozine-Konferenzen) entstanden sind, oder um Artikel, die einen neuen Focal Point thematisch ergänzen.

Das Artikel-Archiv umfasste zum Zeitpunkt des Beginns dieser Untersuchung knapp 7.600 Artikel.[9] Knapp die Hälfte der Artikel sind englischsprachig. Viele Artikel des Archivs stehen auf der Website sowohl in der Sprache der Originalveröffentlichung (Russisch, Norwegisch, Deutsch, Spanisch etc.) als auch in englischer oder anderen Übersetzungen zur Verfügung.[10] Hinter dem Anliegen von Eurozine, möglichst viele englischsprachige Artikel zu veröffentlichen, steht das Interesse, die Themen und Debatten einer internationalen LeserInnenschaft zugänglich zu machen. Dem entsprechend ist ein inhaltliches Auswahlkriterium für die Artikelvorschläge der einzelnen Redaktionen, dass die Themen auch für eine LeserInnenschaft jenseits des Erscheinungslands der Zeitschrift – also in einem europäischen Kontext – von Interesse sind.

Form und Länge der Artikel bei Eurozine variieren – ähnlich den unterschiedlichen Ausrichtungen der Partnerzetischriften – sehr stark. Es gibt essayistisch geschriebene Artikel, Artikel, die eher in einem klassisch wissenschaftlichen Stil verfasst (und mit Quellenverweisen in Form von Fußnoten ausgestattet) sind, es gibt Interviews und vereinzelte literarische bzw. Lyriktexte. Die Länge der Artikel variiert von weniger als 1.000 Wörtern bis zu Beiträgen mit mehr als 6.000 Wörtern.

Alle Artikel haben eine redaktionelle Begutachtung durch die Partnerzeitschriften hinter sich. Diese Veröffentlichungspraxis spiegelt – ähnlich der Zusammenstellung der Zeitschriften – das Interesse wider, kultur- und gesellschaftspolitische Debatten über Länder- und Sprachgrenzen hinaus zu transportieren und einer europäischen LeserInnenschaft zugänglich zu machen: Auf diese Weise können Debatten aus Südeuropa Debatten in Nordeuropa rezipiert und Themen aus Osteuropa in Westeuropa aufgegriffen werden.

[9] Stand Dezember 2014

[10] Die Anzahl der Übersetzungen ist sowohl abhängig davon, ob ein Artikel ggfs. schon vor seinem Erscheinen in einer Partnerzeitschrift auf Englisch vorlag, als auch von den finanzielle Möglichkeiten Eurozines. EinE RedakteurIn ist für Übersetzungen ins Englische zuständig.

Die Struktur der Artikel

Bei der Einzelartikelansicht (Abb. 3.4) steht oberhalb des Titels jedes Beitrages, in welcher Zeitschrift er veröffentlicht worden ist und in welchen Sprachen der Artikel bei Eurozine gelesen werden kann.

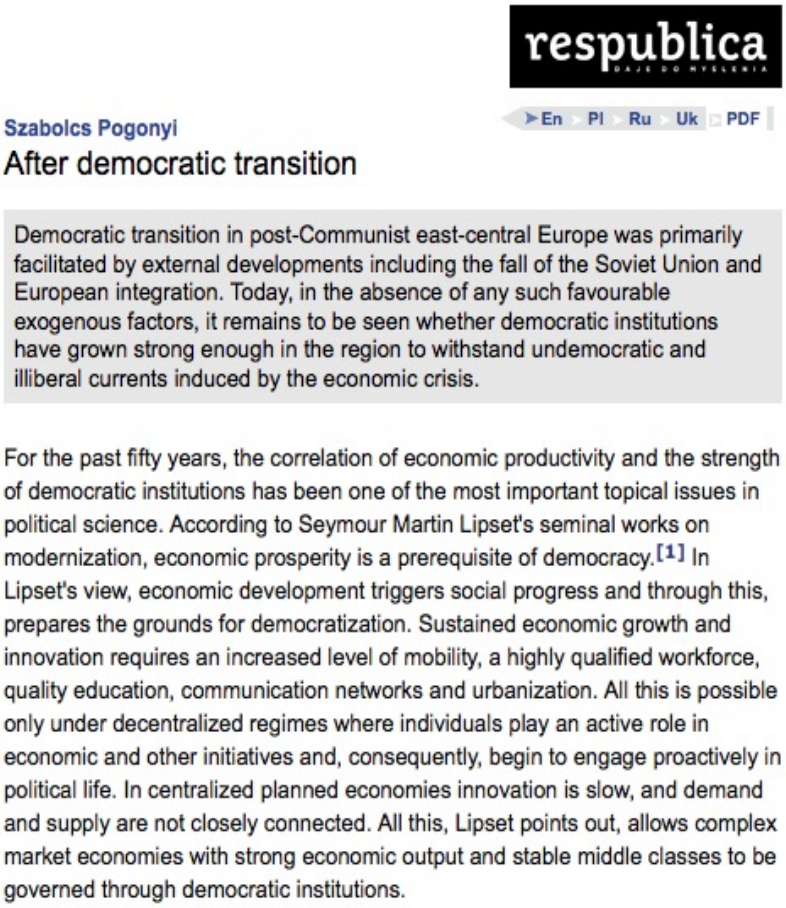

Abbildung 3.4: Einzelartikelansicht

In viele Artikel sind Infokästen (mit Verweisen auf themenverwandte Artikel und andere Zusatzinformationen) und Bilder eingebaut. Am Ende jedes Artikels (ggfs. unterhalb der Fußnoten) finden sich im sogenannten Publikationskasten (Abb. 3.5) die Eckdaten der Veröffentlichung: das Datum der Veröffentlichung des Beitrags bei Eurozine, der Name der erstveröffentlichenden Zeitschrift, die Heftnummer, ein Copyrightverweis und ggfs. der Name der Übersetzerin / des Übersetzers.

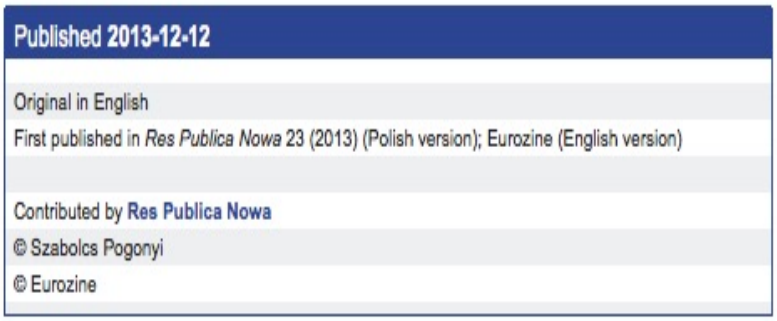

Abbildung 3.5: Publikationskasten

In der Datenstruktur der der Seite zugrunde liegenden XML-Dateien sind für verschiedene Bereiche eines Artikels unterschiedliche XML-Auszeichnungen (Mark-ups) vorgesehen (Abb. 3.6). Auf diese Struktur wird im weiteren Verlauf der Arbeit, im Kapitel zum Anwendungsfall, im Rahmen der Überlegungen, wie durch Textmining-Verfahren teilautomatisierte Dossiers erstellt werden können, zurückgegriffen. Im Folgenden werden nun die Auszeichnungen in der Reihenfolge ihres Auftretens in den Artikeldateien beschrieben:

In den XML-Dateien des Eurozine-Archivs steht zu Beginn „!DOCTYPE article SYSTEM“ sowie ein Verweis auf eine Schemadatei namens „article.dtd“. Diese Schemadatei gibt es leider nicht mehr. Sie wurde dafür verwendet, die häufig unsauberen XML- (oder anderen) Daten, die von den Partnerzeitschriften mit Code-Fehlern an Eurozine weitergegeben wurden, in ein für die Weiterverarbeitung brauchbares Format zu übersetzen.

Die Auszeichnung <article> wird für jeden Artikel als Kennzeichnung des kompletten Datensatzes verwendet. Innerhalb des öffnenden Tags findet sich jeweils die Abkürzung der Sprache, in der der Artikel verfasst worden ist („en“ für Englisch, „de“ für Deutsch etc.).

Der Abschnitt „Impressum“ am Anfang der Datei (<imprint>) beinhaltet die Mark-ups

- <author> (wenn es verschiedene AutorInnen gibt, wird diese Auszeichnung entsprechend der Anzahl der beteiligten AutorInnen verwendet),
- <copyright> (häufig identisch mit <author>, oft ergänzt um ein bei Eurozine liegendes Copyright),
- <contributor> für die Zeitschrift, in der der Artikel erschienen ist (in dem Fall, dass Eurozine einen Artikel einwirbt, steht dort „Eurozine“ – ggfs. ergänzt um <firstin> als Hinweis für den Ort der Erstveröffentlichung), und

- <pubdate> für das Datum, an dem der Artikel bei Eurozine veröffentlicht worden ist.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!DOCTYPE article SYSTEM "article.dtd">
<article lang="en">
<imprint>
<author>Szabolcs Pogonyi</author>
<copyright>Szabolcs Pogonyi</copyright>
<copyright>Eurozine</copyright>
<contributor>Res Publica Nowa</contributor>
<firstin><i>Res Publica Nowa</i> 23 (2013) (Polish version); Eurozine (English
version)</firstin>
<pubdate>2013-12-12</pubdate>
</imprint>
<title>After democratic transition</title>
<blurb>Democratic transition in post-Communist east-central Europe was primarily
facilitated by external developments including the fall of the Soviet Union and
European integration. Today, in the absence of any such favourable exogenous
factors, it remains to be seen whether democratic institutions have grown strong
enough in the region to withstand undemocratic and illiberal currents induced by
the economic crisis.</blurb>
<body>
For the past fifty years, the correlation of economic productivity and the
strength of democratic institutions has been one of the most important topical
issues in political science. According to Seymour Martin Lipset's seminal works
on modernization, economic prosperity is a prerequisite of
democracy.<footnote>Seymour Martin Lipset, <i>Political Man: The Social Bases of
Politics</i> (Heinemann, 1960); Seymour Martin Lipset, "Some Social Requisites of
Democracy: Economic Development and Political Legitimacy", <i>The American
Political Science Review</i> 53, no. 1 (1959): 69-105</footnote> In Lipset's
view, economic development triggers social progress and through this, prepares
the grounds for democratization. Sustained economic growth and innovation
requires an increased level of mobility, a highly qualified workforce, quality
education, communication networks and urbanization. All this is possible only
under decentralized regimes where individuals play an active role in economic and
other initiatives and, consequently, begin to engage proactively in political
life. In centralized planned economies innovation is slow, and demand and supply
are not closely connected. All this, Lipset points out, allows complex market
economies with strong economic output and stable middle classes to be governed
through democratic institutions.<p/>

<figure class="alignleft">
<img
src="http://www.eurozine.com/UserFiles/illustrations/pogonyi_468w.jpg"/>
<figcaption>Warsaw skyline. Photo: Guido Heitkoetter. Source: <a
href="http://www.flickr.com/photos/guidolo/3810755031/in/photostream/">Flickr</
a></figcaption></figure>
```

Abbildung 3.6: Der Anfang einer XML-Datei aus dem Eurozine-Archiv

Nach dem Impressum finden sich zunächst die Auszeichnungen

- <title> für den Titel des Beitrags,
- <subtitle> (falls ein Untertitel vohanden ist) und
- <blurb> für den Abstract, der in wenigen Zeilen den Artikelinhalt zusammenfasst und auf der Einzelartikelansicht der Website grau unterlegt ist.

Unterhalb der genannten Tags steht der eigentliche Artikel – gekennzeichnet mit dem Tag <body>. Innerhalb dieses Tags gibt es weitere Auszeichnungen, die bei Bedarf eingesetzt werden, aber nicht in allen Artikeln vorkommen. Dies betrifft die Tags:

- <motto>, mit dem Zitate o.Ä. einem Artikel vorangestellt werden können,

- <subheadline> für Zwischenüberschriften,
- <footnote> für Fußnoten, die in den Fließtext gesetzt und auf der Website am Ende des Artikels nummeriert ausgegeben werden,
- <figure> für einzufügende Bilddateien, die zugehörigen Referenzen und Bildunterschriften und
- <a> als Anker für z. B. Verweise auf andere Webseiten.

Dazu kommen Tags für Kursivierungen oder Fettungen von Textstellen ebenso wie die Auszeichnung <p> für die Unterteilung eines Artikels in verschiedene Absätze.

Die Auszeichnung <includebox> fügt eine Infobox innerhalb eines Artikels ein. So kann auf weitere Beiträge zum gleichen Thema („Read also“) verwiesen werden. Solche Verknüpfungen werden, ebenso wie die Zusammenstellung der Focal Points, ausschließlich händisch von der Redaktion erstellt. Dies erfordert sowohl einen hohen zeitlichen Aufwand als auch ein umfangreiches Wissen über die Inhalte und Themen des Artikelarchivs. Verschlagwortungen ebenso wie automatische Verknüpfungen existieren derzeit nicht. Nutzbar ist für einen solchen Zweck bisher nur die Volltextsuche auf der Seite, die allerdings keine zufriedenstellenden Ergebnisse liefert. Auch bestehende Rubriken wie „Latest articles“ und die Rubrik „Partner journals“ ermöglichen nur einen eingeschränkten Zugang zum Archiv von Eurozine. Der Großteil der Artikel ist mit diesen Möglichkeiten nur schlecht oder gar nicht auffindbar.

3.4 Bedarf und Perspektiven für die Erschließung des Eurozine-Archivs

Vor dem Hintergrund des wachsenden Umfangs des Archivs und dem damit einhergehenden zunehmenden Zeitaufwand bei der händischen Erstellung von Focal Points stellt sich für diese Arbeit die Frage, ob und wie inhaltliche Verknüpfungen zwischen Artikeln automatisch ermittelt und als „related content“ (zeitschriften- und jahrgangsübergreifend) ausgegeben werden könnte. Dies erfordert die Ermittlung von Distanzen zwischen Artikeln.

Eine schnellere und umfassendere Erschließung des Archivs würde eine bessere Nutz- und einfachere Handhabbarkeit bedeuten. Außer der Eurozine-Redaktion könnten davon sowohl die Partner-Redaktionen als auch die LeserInnen profitieren: Ihnen könnte ein schnellerer und im besten Fall umfangreicherer Überblick über das vorhandene Archiv ermöglicht werden.

Für die weitere Arbeit wird im Folgenden zusammengefasst, für welche Kombination der im Kapitel 2 beschriebenen möglichen Dossierarten eine Automatisierung erfolgen und wem dies als Unterstützungssystem dienen soll.

Die Focal Points bei Eurozine sind, wie gezeigt wurde, Dossiers, die verschiedene Facetten eines Themas abbilden. Dabei geht es nicht um die Zusammenstellung aller bei Eurozine zu einem Thema erschienenen Artikel. Die Focal Points bezwecken vielmehr, als Leseanregung einen Überblick zu einem Thema zu geben. Zudem wird durch die Beiträge verschiedener europäischer Kulturzeitschriften eine europäische Debatte bereichert.

Überlegungen hinsichtlich einer teilautomatisierten Dossiererstellung dienen in erster Linie einer Entlastung der Eurozine-Redaktion. Eine rechnerunterstützte Erschließung des Archivs würde ermöglichen, (bislang verborgenes) Wissen zu bergen. Gedacht werden könnte z. B. an ein Vorschlagsystem: Durch den Einsatz von Textmining (s. Kapitel 4) bzw. durch das Ausprobieren einzelner Verfahren am Eurozine-Korpus wird im weiteren Verlauf der Arbeit (s. Kapitel 5) untersucht, ob eine automatisch erstellte Sammlung von Artikeln, die inhaltliche Ähnlichkeiten aufweisen, generiert werden kann. Diese Artikelvorschläge würden dann einer inhaltlichen Prüfung durch die Redaktion unterzogen und ausgewählte Artikel zu einem Focal Point zusammengestellt. So würde der erste Schritt, die Suche nach zueinander passenden Artikeln im gesamten Archiv, maschinell unterstützt werden, während der zweite weiterhin in Form einer Qualitätsprüfung und engeren Auswahl der Artikel durch ExptertInnen erfolgt. Ziel dessen ist eine maschinelle Unterstützung für Redaktionen ohne Qualitätsverluste.

Zur weiteren Auseinandersetzung mit dem vorliegenden Artikelarchiv können folgende Aspekte festgehalten werden:

Datenstruktur

Für den Einsatz von Mining-Verfahren sind saubere Daten, das heißt Daten mit einer einheitlichen Struktur, sehr wichtig. Je sauberer die Daten vorliegen, desto gründlicher bzw. präziser kann eine Analyse der Daten erfolgen. Entsprechend besser fallen auch die Ergebnisse der Analyse aus. Die Untersuchung der Datenstruktur des Eurozine-Archivs und der beschriebenen Auszeichnungen (Tags) zeigte, dass das XML des Artikelarchivs invalide ist und Unregelmäßigkeiten bzw. Fehler an verschiedenen Stellen aufweist. Ein Nachbearbeiten der Daten ist zeit- und kostenintensiv. Die Unregelmäßigkeiten bestehen größtenteils in fehlenden Tags, die die Datenanalyse erschweren. Beispiele hierfür sind:

- Wenn nur die Abstracts von Artikeln (<blurb>) auf bestimmte Wörter durchsucht werden sollen, manche Artikel aber kein Abstract bzw. keine Auszeichnung für ein Abstract

aufweisen, können nicht alle Artikel in der Analyse berücksichtigt werden, sondern nur Artikel, die das Tag <blurb> aufweisen.

- Wenn nur die in den Artikeln vorkommende Zitate auf ihren Inhalt hin analyisert werden sollen, aber manchen Zitaten das schließende Tag fehlt, das das Ende des Zitats markiert, ist es nicht möglich, das Zitat zu erkennen – und entsprechend unmöglich, eine Analyse der Zitate durchzuführen.

Außer dem Autorennamen, dem Ersterscheinungsort und dem Veröffentlichungsdatum sind bei den meisten Artikeln auch der Titel und, soweit vorhanden, der Untertitel ausgezeichnet. Der überwiegende Anteil der Artikel beinhaltet vor dem Beginn des eigentlichen Beitrags zudem eine kurze Zusammenfassung in Form eines Abstracts („blurb"). Viele Artikel haben Absätze bzw. Zwischenüberschriften, welche allerdings nicht ausgezeichnet sind und somit nicht strukturiert ausgelesen werden können. Viele Artikel haben Fußnoten. Zudem finden sich innerhalb der Beiträge immer wieder vereinzelte Links zu Fotos, anderen Artikeln sowie Zitate. Eine Verschlagwortung oder eine in den XML-Dateien vermerkte Zuordnung zu bestehenden Focal Points exisitert nicht. Es handelt sich also um einen semistrukturierten Textkorpus.

Die Vorbereitung der Archiv-Daten, die in dieser Arbeit untersucht wurden, beinhaltete die Bereinigung der Daten – zugunsten einer teilweisen Vereinheitlichung der Datenstruktur (soweit möglich).[11]

Analyse von Textabschnitten und Gewichtung

Der Umstand, dass die Eurozine-Artikel größtenteils lange Artikel sind und nicht Kurztexte, wie sie z. B. bei Twitter zum Einsatz kommen, lässt eine differenzierte Betrachtung unterschiedlicher Textabschnitte als sinnvoll erscheinen. Angesichts der unterschiedlichen Auszeichnungen sollte untersucht werden, ob eine inhaltliche Erschließung des Archivs auch über einzelne Textabschnitte erfolgen könnte – ob also die Distanz zwischen Artikeln auch über die Wörter aus den Überschriften oder aus den Abstracts ermittelt werden könnte. Kriterien für Distanzen bzw. „related content" müssten dafür noch festgelegt werden.

Die Möglichkeit, dass durch die Auszeichnungen einzelne Teile der Dateien isoliert untersucht werden können, lässt es als sinnvoll erscheinen, beispielsweise die Wörter aus den Artikelüberschriften höher zu gewichten als die Wörter, die im restlichen Artikel vorkommen. Die Annahme dabei wäre, dass Wörter, die in den Überschriften vorkommen, eine hohe

[11] Vielen Dank an Tobias Eichler und Marcel Schöneberg, die die Säuberung der Daten übernommen haben. Dies war notwendig, um die Daten durch ein Programm zur Erstellung einer Tag-Cloud laufen zu lassen und um mit Rapidminer und Overview arbeiten zu können.

Aussagekraft für den Inhalt der Artikel haben und die Charakterisierung von Artikeln sowie die Bemessung des Abstands zwischen Artikeln präziser wird, wenn zentrale Begriffe einen höheren Wert zugemessen kommen. ein Problem, dies umzusetzen, könnte darin bestehen, dass nicht alle Überschriften als Titel gekennzeichnet sind.

Ermittlung von Distanzen anhand von Metainformationen

Um Kriterien für eine teilautomatisierte Erstellung von Dossiers zu entwickeln, bedarf es einer Definition, die ermöglicht, den Abstand bzw. die Distanz von Artikel zueinander zu bestimmen. Die in den XML-Daten verwendeten Tags enthalten Metainformationen, die bereits eine einfache Bestimmung von Distanzen ermöglichen: Bezogen auf die Dimensionalitäten der Merkmale Sprache, Veröffentlichungsdatum, Zeitschrift, AutorIn und ÜbersetzerIn lassen sich Übereinstimmungen bei verschiedenen Artikeln feststellen. Diese Übereinstimmungen können als Zusammengehörigkeit von Artikeln begriffen werden. Konkret: Artikel, deren Metabeschreibung zu entnehmen ist, dass es sich um englischsprachige Artikel handelt, gehören, bezogen auf die Sprache, zusammen. Sie haben, bezogen auf das Kriterium Sprache, einen geringeren Abstand zueinander als Artikel unterschiedlicher Sprachen. Ergänzend dazu könnte man Kriterien entwickeln, um weitere Distanzen aus den Metainformationen zu bestimmen. Dies könnten z. B. die geografische Nähe der jeweiligen Partnerzeitschriften sein (Bsp.: Frankreich und Spanien haben eine geringere geografische Distanz zueinander als Frankreich und Serbien) oder die inhaltliche bzw. thematische Ausrichtungen der Zeitschriften (Bsp.: Kunst und Theater versus Politik).

Im Folgenden wird für die verschiedenen Auszeichnungen kurz skizziert, wie Distanzen ermittelt werden könnten bzw. welche Kriterien für Distanzen herangezogen werden könnten. Dies dient einer Hinführung zur Vergleichbarkeit von Artikeln.

- Sprache: Eine Annahme könnte sein, dass Artikel gleicher Sprache einem Land zugeordnet werden können. Dabei kommt es jedoch sofort zu einem Zuordnungsproblem, da viele Artikel bei Eurozine auf Englisch veröffentlicht worden sind. Auch die Tatsache, dass manche Artikel in verschiedene Sprachen bei Eurozine zu finden sind, zeigt, dass das Kriterium Sprache bei der Erstellung teilautomatisierter Dossiers höchstens in einem Punkt relevant sein könnte: für die Einteilung des Archivs in englische und anderssprachige Artikel. So kann über das Mark-up Sprache ein Datenset zusammengestellt werden, das, wie für die vorliegende Arbeit geschehen, ausschließlich englische Artikel enthält. Die beiden Vorteile eines solchen Vorgehens sind: 1. Die mögliche Dopplung von Artikeln durch das Vorkommen verschiedener Übersetzungen wird durch den

Filter „Sprache“ vermieden, und 2. müssen bei der Erschließung eines sprachlich einheitlichen Korpus keine zusätzlichen Übersetzungsprogramme in den Textmining-Prozess integriert werden. Ergänzend zum Aspekt Sprache kann gesagt werden, dass eine inhaltliche Erschließung des Archivs durch das Kriterium Sprache nicht möglich ist.

- AutorIn: Eine andere Annahme könnte sein, dass Artikel der gleichen AutorInnen inhaltliche Übereinstimmungen aufweisen und dadurch eine Distanz zwischen Artikeln ermittelt werden könnte. Ein Blick in das Archiv von Eurozine bestätigte die Annahme, dass AutorInnen oftmals über mehrere Jahre hinweg bestimmte Themen und Entwicklungen verfolgen (wie beispielsweise die koratische Schriftstellerin und Journalistin Slavenka Drakulic). Das Kriterium <Autor> könnte also sowohl aus dem genannten Grund für die Erstellung teilautomatisierter Dossiers relevant sein als auch vor dem Hintergrund, dass ein Kriterium bei der Erstellung selbiger sein könnte, keine Doppplung von AutorInnen aufzuweisen. Ein Filter über die Auszeichung <Autor> würde dies ermöglichen.

- ÜbersetzerIn: Für das Merkmal „Übersetzer“ kann, anders als für das Merkmal „Autor“, angenommen werden, dass es wenig Relevanz hinsichtlich der Erstellung teilautomatisierter Dossiers hat, da ÜbersetzerInnen keinen im engeren Sinne eigenen inhaltlichen Beitrag zum Artikel leisten. Jedoch weist das Merkmal darauf hin, dass der Artikels ursprünglich in einer anderen Sprache verfasst worden ist. Eine Distanz zwischen Artikeln anhand des Kriteriums „Übersetzer“ zu ermitteln, erscheint für den Zweck dieser Arbeit allerdings nicht zweckdienlich.

- Zeitschrift: Für Artikel, die in der gleichen Zeitschrift erschienen sind, könnte angenommen werden, dass sie eine inhaltliche Nähe in thematischer Hinsicht aufweisen – davon ausgehend, dass sich viele Kulturzeitschriften einem Bereich wie Literatur, Politik, o.a. zuordnen lassen. eine solche Zuordnung würde eine Erschließung des Archivs durch die Vorsortierung der Zeitschriften ermöglichen und könnte ein erster Schritt in Richtung einer Kategorienbildung sein. Gleichzeitig kann über das Merkmal „Zeitschriftgrqq als Filter verwendet werden, um zu verhindern, mehrere Artikel der gleichen Zeitschrift in ein Dossier aufzunehmen. Gerade für ein Kulturzeitschriftennetzwerk wie Eurozine könnte ein Kriterium bei der teilautomatisierten Dossiererstellung sein, Debatten zeitschriftenübergreifend abzubilden.

- Datum: Bislang ist auf der Website eine zeitliche Suche nach dem Jahr der Veröffentlichung von Artikeln möglich. Ein Eingrenzen der Treffer auf Monate oder einzelne Tage

kann nicht vorgenommen werden. Das Merkmal „Veröffentlichungsdatum“ könnte bei einer teilautomatisierten Dossiererstellung dabei unterstützen, Debattenverläufe chronologisch abzubilden. Das Datum allein gibt jedoch noch keine verlässliche Auskunft darüber, ob der Inhalt eines Artikels ein aktueller ist oder ob es sich um einen Rückblick handelt. Auch muss bezweifelt werden, dass Artikel, die in zeitlicher Nähe zueinander erschienen sind, eine inhaltliche Nähe zueinander aufweisen, die bei der Dossiererstellung von Bedeutung sein könnte. Verlässliche Aussagen über den Inhalt der Artikel können auf diese Weise nicht getroffen werden.

- Titel und Untertitel: Während die gerade beschriebenen Auszeichnungen kaum bzw. wenige Informationen über die Inhalte von Artikeln geben, kann angenommen werden, dass die Mark-ups für Titel und Untertitel eines Artikels (<title> und <subtitle>), für den Abstract (<blurb>) und für den eigentlichen Artikeltext (<body>) deutlich aussagekräftiger hinsichtlich der Themen von Artikeln sind. Überlegungen zur Ermittlung von Distanzen zwischen Artikeln können hier jedoch in ähnlicher Weise angestellt werden: Ein geringer Abstand zwischen Artikeln könnte angenommen werden, wenn es Übereinstimmungen vieler Wörter in den Artikeln gibt oder wenn identische Wörter ähnlich häufig in unterschiedlichen Artikeln vorkommen. Eine komplexere Aufgabe würde darin bestehen, nicht nur identische, sondern ähnliche Wörter auszumachen – also Wörter, die zu einem Themenfeld gehören. Anders als es bei dem Großteil der zuvor beschriebenen Kriterien der Fall sein wird, wird diesen Annahmen im Kapitel 5 genauer nachgegangen: Mithilfe verschiedener Verfahren zur Erschließung des Archivs wird die Verteilung von Worthäufigkeiten in den Artikeln eines Focal Points und im gesamten Korpus ermittelt, um auf dieser Grundlage Distanzen zwischen Artikeln feststellen zu können. Die Ergebnisse sollen dabei unterstützen, Kriterien für eine teilautomatisierte Dossiererstellung weiter zu präzisieren.

  Die bereits existierenden Focal Points dienen in der Untersuchung als Referenzpunkt für Qualität – ausgehend von der Annahme, dass ihrer Erstellung redaktionelle Entscheidungen zugrunde liegen und davon auszugehen ist, dass die Focal Points inhaltlich sinnhaft zusammengesetzt sind. Aus diesem Grund wird im weiteren Verlauf der Arbeit ein bestehender Focal Point und der englischsprachige Teil des Archivs untersucht. So soll der Versuch unternommen werden, Kriterien zu extrahieren, die hinsichtlich einer teilautomatisierten Erstellung von Artikelsammlungen hilfreich sind.

4 Textmining – Grundlagen

Die Menge der weltweit produzierten Daten verdoppelt sich inzwischen innerhalb von 20 Monaten (vgl. Runkler [2015], Marr [2015] u.a.). Durch moderne Technologien können nicht nur in immer kürzeren Zeitintervallen größer werdende Datenmengen produziert, sondern auch dauerhaft gespeichert werden. Diese Entwicklung hat zu einer Informationsflut (auch als „Information Overload“ bezeichnet) geführt, angesichts der Menschen die vorhandenen Datenmengen nicht mehr mit herkömmlichen, analogen Methoden untersuchen bzw. die von ihnen gesuchten Informationen finden können.

Digitale Daten liegen in den Datenbanken von Bibliotheken, Archiven, Nachrichtenstreams und Webseiten in unterschiedlicher Struktur vor. Sie bedürfen unterschiedlicher Herangehensweisen für die Erschließung und strukturierte Aufbereitung, damit die gezielte Suche nach Informationen Erfolg haben kann. Während bspw. für Bibliotheken das Management von Dokumenten eine zentrale Aufgabe darstellt und bestehende Signaturen und Verschlagwortungen beim Auffinden von Dokumenten unterstützen, steht bei der Informationssuche auf Webseiten oder der Analyse unverschlagworteter Archive, wie dem Eurozine-Archiv, oftmals deutlich weniger Struktur zur Verfügung, die eine solche Unterstützung bieten könnte.

Um den Herausforderungen angesichts der wachsenden Datenmengen und dem Interesse, aus den Daten nützliches Wissen zu bergen, zu begegnen, sind spezielle Methoden für die Analyse großer Datenmengen entwickelt worden. Sie werden unter Begriffen wie Dataming und Textmining subsumiert (vgl. Aggarwal [2015], S. 429, u.a.). Grundlegende Begriffe zum Verständnis dieser Methoden werden in diesem Kapitel dargestellt: Außer Begriffsdefinitionen findet die Beschreibung des klassischen Vorgehens bei einer maschinellen Analyse großer Datenbestände, dem sogenannten KDD-Prozess (KDD bedeutet Knowledge Discovery in Databases) statt. Im Anschluss daran werden einige Arbeiten vorgestellt, die sich mit Möglichkeiten der automatisierten Erschließung textbasierter Datenbanken auseinandersetzen.

Die Ausgangssituation dieser Arbeit berücksichtigend, handelt es sich bei den Ausführungen dieses Kapitels in erster Linie um eine sich den technologischen Voraussetzungen und

Erfordernissen beim Einsatz von Textmining-Verfahren annähernde Perspektive und weniger um eine Auseinandersetzung aus klassisch informationstechnologischer Sicht.[25]

4.1 Begriffsdefinitionen

Die Bedeutungen der Bezeichnungen Datamining und Textmining überschneiden sich teilweise bzw. werden nicht immer trennscharf verwendet. Im gleichen Zusammenhang fallen auch die Begriffe Information Retrieval (IR) und Knowledge Discovery in Databases (KDD) (vgl. Lewandowski [2005], Manning u. a. [2009], u.a.). Alle genannten Methoden bedienen sich Techniken und Werkzeuge unterschiedlicher Disziplinen: Neben Statistik, Datenbanken und Computergrafik kommen auch Forschungsergebnisse zu Künstlicher Intelligenz, Computerlinguistik, Mustererkennung, Clusteranalyse und Klassifikation zum Einsatz (vgl. Cleve u. Lämmel [2014], S. 11, und Witte u. Mülle [2006], S. 43).

Grob kann unterschieden werden, dass beim Einsatz von Datamining- und Textmining-Verfahren sowie bei Knowledge Discovery in Databases nach neuen, noch unbekannten Informationen gesucht wird (Kroeze u. a. [2003]), während sich Information Retrieval dem Wieder- bzw. Auffinden prinzipiell bekannter Informationen widmet (Jain u. a. [1999]).

4.1.1 Textmining

Das Ziel von Textmining besteht darin,

> aus einem einzelnen oder einem Satz von Dokumenten neues Wissen zu extrahieren, etwa durch automatische Textzusammenfassungen, die Erkennung und Verfolgung benannter Objekte oder die Aufdeckung neuer Trends in Forschung und Industrie. (Witte u. Mülle [2006], S. IX)

Textmining zielt damit auf die Beschaffung von neuem, nützlichem, bedeutsamen und gültigen Wissen durch eine komplexe automatisierte Analyse und den Vergleich der Dokumente bzw. Dateien ab.

Data- wie auch Textmining-Verfahren und KDD versuchen durch die Analyse großer Datenbestände Muster in selbigen zu erkennen. Bei Textmining-Verfahren werden, anders als beim Datamining, keine strukturierten, sondern zumeist Textdateien, die in semi- oder unstrukturierte Daten vorliegen, analysiert (vgl. Hippner u. Rentzmann [2006]): „The discovery

[25] Die Darstellung von Textmining-Verfahren aus informationstechologischer Perspektive kann bei Marcel Schöneberg nachgelesen werden. Vgl. dazu: https://users.informatik.haw-hamburg.de/ ubicomp/arbeiten/master/schoeneberg.pdf

of knowledge from databases sources containing free text is called text mining.“ (vgl. Kroeze u. a. [2003]) Anders als in strukturierten Datenbanken kann aus diesem Grund in Artikeldatenbanken nicht jeder Teil eines Artikels exakt zugeordnet werden. Textmining, in dessen Fokus die Analyse von Dokumentensammlungen steht, wird deshalb als ein Teilbereich des Datamining angesehen.

Textdateien verfügen überwiegend „über eine implizite Struktur, die aus der Grammatik resultiert, und – je nach Textdokument – über eine explizite Struktur, die sich z. B. aus Titel/Untertitel und Absätzen erschließen lässt“. (Cleve u. Lämmel [2014], S. 65) Die explizite Struktur ist im vorangegangenen Kapitel bereits beschrieben worden: Es handelt sich dabei um die Auszeichnungen in den XML-Dateien, die im weiteren Verlauf der Arbeit noch mal zur Sprache kommen werden.

Feldman u. Sanger [2007] beschreiben Textmining als

> a new and exciting research area that tries to solve the information overload problem by using techniques from data mining, machine learning, natural language processing (NLP), information retrieval (IR), and knowledge management. Text mining involves the preprocessing of document collections (text categorization, information extraction, term extraction), the storage of the intermediate representations, the techniques to analyze these intermediate representations (such as distribution analysis, clustering, trend analysis, and association rules), and visualization of the results. (S. X)

Bei der Analyse wird eine Sammlung von Dokumenten untersucht, um auf diese Weise auf Informationen zu stoßen, die nicht aus eine einzelnen Dokument gezogen werden können, sondern erst durch einen artikelübergreifenden Blick auf eine Sammlung von Dokumenten. Dies erklärt, warum zu Beginn eines Mining-Prozesses noch nicht klar ist, was gefunden werden könnte. Einzig Hypothesen können formuliert werden.

Für Journalisten bietet Textmining eine gute Möglichkeit zur Erschließung von Archiven (vgl. Hälker [2014]). Bezogen auf den Artikelkorpus des Eurozine-Archivs, der im Fokus der vorliegenden Arbeit steht, bedeutet das: Es ist bekannt, dass es sich um ein semistrukturiertes, dynamisches Archiv handelt (Feldman u. Sanger [2007], S. 2).[1] Titel, Abstracts, Sprache, etc. können aufgrund der bereits beschriebenen Auszeichnungen isoliert analysiert werden. Zudem ist aufgrund der Zeitschriftenprofile, den Newsletter und die redaktionell erstellten Focal Points bekannt, zu welchen Themenbereichen Artikel in Form von ausgezeichneten XML-Dateien im Archiv liegen. Diese Ausgangssituation ist typisch für den Einsatz von Textmining-

[1] Ein dynamisches Archiv ist im Gegensatz zu einem statischen Archiv unabgeschlossen: Es wird stetig um neue Archivinhalte erweitert.

Verfahren: Die Größe und fehlende Strukturiertheit des Archivs begründen das Interesse, automatisierte Verfahren zur Erschließung einzusetzen.

Aufgabe des Textmining ist es in diesem Fall, Muster im Eurozine-Archiv zu entdecken und daraus Vorschläge für neue Themendossiers zu generieren. Bezweckt wird, Ähnlichkeiten zwischen Artikeln zu finden – z. B. in Form thematischer Überschneidungen. Um diese Ähnlichkeiten messen zu können, müssen Distanzen zwischen Artikeln ermittelt werden. Das kann u.a. durch den Einsatz von Distanzfunktionen geschehen.[2] In der vorliegenden Arbeit geschieht dies durch die Auswahl von Selektoren bzw. Features: Als Selektor oder Feature werden die Kriterien bezeichnet, die als Grundlage für die Ermittlung von Distanzen herangezogen werden. Ein Feature kann bspw. ein ausgewähltes Wort sein, dessen Häufigkeit in den zu analysierenden Artikeln ermittelt wird, um die Distanz der Artikel anhand der Häufigkeitsverteilung des Wortes zu bestimmen.[3] Würde man die Distanz zwischen Artikeln hinsichtlich der Sprachen, in denen sie geschrieben worden sind, ermitteln wollen, wäre der Selektor oder das Feature die Sprache. Die ermittelten Zahlen werden als Feature-Vektoren abgebildet, mittels derer die Werte der Artikel miteinander verglichen und so Distanzen ermittelt werden können. Besteht ein Feature nur aus einem Wort, spricht man von einem eindimensionalen Featurevektor, der ermittelt wird; werden alle Wörter eines Artikels hisichtlich ihrer Häufigkeit ermittelt, handelt es sich um einen n-dimensionalen Feature-Vektor.

Durch die Berechnung von Distanzen wird es möglich, einander ähnliche Artikel als Dossiervorschläge für Redaktionen zu extrahieren. Aus diesen Vorschlägen können dann wiederum redaktionell neue Focal Points erstellt werden. Der Vorteil für die RedakteurInnen besteht dabei darin, nicht eigenhändig das Archiv durchsuchen zu müssen, sondern automatisch generierte Artikelvorschläge für ein Thema ausgegeben zu bekommen.

4.1.2 Information Retrieval, Datamining und Knowledge Discovery in Databases

Im folgenden Abschnitt werden Überschneidungen und Abgrenzungen zwischen Information Retrieval (IR), Datamining und Knowledge Discovery in Databases (KDD) und ihr Verhältnis zu Textmining skizziert.

[2] Zum Einsatz von Distanzfunktionen vgl. die Masterarbeit von Marcel Schöneberg: https://users.informatik.haw-hamburg.de/ ubicomp/arbeiten/master/schoeneberg.pdf, S. 20

[3] Z. B. kann auf diese Weise herausgefunden werden, welche Artikel wie häufig das Wort „Demokratie“ beinhalten. Vgl. dazu Kap. 5. Es können auch alle im Dokument vorkommenden Wörter in den Feature-Vektoren eingerechnet werden. Durch diese n Dimensionen wird allerdings der Feature-Raum deutlich größer und unübersichtlich. Vgl. dazu Feldman u. Sanger [2007], S. 6.

IR ist „traditionell auf einen leichten Zugang zu Informationen fokussiert, nicht auf deren Analyse“. (Seidel, S. 24) und verfolgt damit allgemeinere Aufgaben als Textmining. Das Ziel von IR besteht darin, NutzerInnen dabei zu unterstützen, ihr Informationsbedürfnis zu befriedigen: „The problem isn't so much that the desired information is not known, but rather the desired information coexists with many other valid pieces of information.“ (Hearst [1999], S. 3) Der Zugang zu und die Erschließung von Informationen durch IR-Verfahren ist die Vorbedingung für eine differenziertere Analyse eines Archivs durch den Einsatz von Textmining-Verfahren.

Beim IR ist zu Beginn der Analyse meistens bekannt, wonach gesucht wird bzw. welche Informationen (wieder-)gefunden werden sollen. Demgegenüber steht am Anfang von Data- bzw. Textmining sowie bei KDD nicht fest, was genau gesucht wird (Manning u. a. [2009]). KDD, Data- und Textmining bezwecken, „to discover or derive new information from data, finding patterns accross datasets, and/or separate signal from noise“. (Hearst [1999], S. 3)

Kroeze u. a. [2003] weisen auf die Erweiterung des Forschungsfeldes hin, die sich im Bereich der Knowledge Discovery vollzogen habe: Vom (nahezu) ausschließlichen Auslesen strukturierter Datenbanken habe sich das Feld in jüngerer Zeit um das Auslesen von (semi- bis unstrukturierten) Textdatenbanken erweitert und sei damit zu einem umfassenderen Forschungsfeld geworden:

> Until recently computer scientists and information system specialists concentrated on the discovery of knowledge from structured, numerical databases and data warehouses. However, much, if not the majority, of available business data are captured in text files that are not overtly structured.

Eine wichtige Voraussetzung für ein erfolgreiches Mining-Projekt besteht darin, ein Verständnis für den jeweiligen Gegenstandsbereich und die zu bearbeitenden Daten zu haben. Ebenso gehört ein Verständnis der unterschiedlichen einsetzbaren Verfahren (vgl. Cleve u. Lämmel [2014], S. VI) dazu. Erst dieses Wissen um die speziellen Aspekte der verschiedenen Verfahren ermöglicht, vor Durchführung von Verfahren qualifiziert entscheiden zu können, welches sich für genau die gestellte Fragestellung eignet.

Für das Eurozine-Archiv kann festgehalten werden, dass die Zusammenstellung von Artikeln zu Dossiers durch den Einsatz von Textmining-Verfahren neues Wissen generieren soll. Der Zugewinn könnte z. B. darin bestehen, Dossiers zusammenzustellen, die verschiedene Blickrichtungen auf eine Fragestellung bündeln. Dies ist ein Fernziel, da Algorithmen, die diese Aufgabe bewältigen, zu einem nicht unwesentlichen Teil mit Verfahren arbeiten müssten, die künstliche Intelligenz einsetzen. Im Rahmen dieser Voruntersuchung wird der Fokus dar-

auf liegen, eine Idee davon zu bekomen, wie die Zusammenstellung von Artikeln überhaupt ansatzweise austomatisiert werden könnte und welche Verfahren sich für diese Aufgabe eignen könnten.

4.2 Der Ablauf einer Datenanalyse – Der KDD-Prozess

Während als Datamining häufig nur die Datenanalyse im engeren Sinne bezeichnet wird, wird unter dem Begriff KDD der gesamte Prozess gefasst: die Datenhaltung und -vorbereitung ebenso wie die Datenanalyse (vgl. Fayyad u. a. [1996]). Der folgende Abschnitt dient der genaueren Beschreibung des Prozesses. Cios u. a. [2007] beschreiben KDD als den „nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data“. Weiter führen sie aus, KDD sei

> the entire knowledge extraction process, including how data are stored and accessed, how to use efficient and scalable algorithms to analyze massive datasets, how to interpret and visualize the results, and how to model and support the interaction between human and machine. It also concerns support for learning and analyzing the application domain (Cios u. a. [2007]).

Cleve u. Lämmel [2014] (S. 5, 65) unterteilen den Ablauf einer Datenanalyse in die Schritte

- Datenselektion und -extraktion (Auswahl der zu analysierenden Datensätze),
- Datenreinigung und -vorbereitung (Säuberung der Daten, Korrektur fehlerhafter Datensätze etc.),
- Datentransformation (die Umwandlung in die benötigten Datenformate),
- Datamining (Exploration, Merkmalsextraktion und -selektion und Finden von Mustern) und
- Interpretation bzw. Evaluation (Auswertung der Ergebnisse) (vgl. auch Boersch).

Bei Textmining erfolgt die Datenanalyse analog zum Ablauf des KDD-Prozesses:

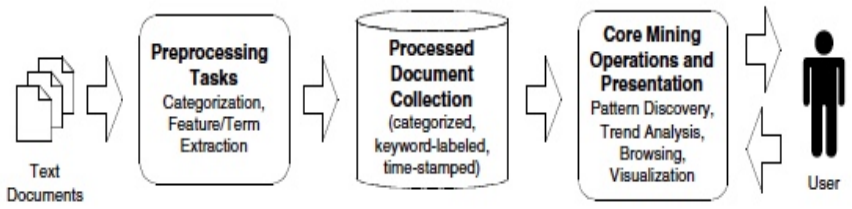

Abbildung 4.1: High-level text mining functional architecture. Abbildung aus Feldman u. Sanger [2007], S. 15

Die Datenvorbereitung

Eine gute Datenvorbereitung („Preprocessing“) bildet eine zentrale Voraussetzung für eine gute Datenanalyse. Sie umfasst

> all those routines, processes, and methods required to prepare data for a text mining system's core knowledge discovery operations. [...] Preprocessing tasks generally convert the information from each original data source into a canonical format before applying various types of feature extraction methods against these documents to create a new collection of documents fully represented by concepts. (Feldman u. Sanger [2007], S. 12f.)

Zur Datenvorbereitung gehören u.a. Tokenization, POS-Tagging, Stemming sowie das Löschen von Stopwords (vgl. u.a. Cleve u. Lämmel [2014], S. 65).

- Bei der Tokenization wird Text in Sätze und Wörter unterteilt. Satzgrenzen müssen dabei sicher erkannt werden (vgl. Feldman u. Sanger [2007], S. 60), was bedeutet, dass Wörter wie Mr. oder Mrs. aber auch andere Abkürzungen, die mit einem Punkt enden, vom Punkt, der das Satzende markiert, unterschieden werden müssen.

- Beim POS-Taggen findet eine Annotation der Wörter eines Textes statt: Wörter werden entsprechend ihres Kontextes im Satz kategorisiert. Die Tags geben Informationen über den semantischen Inhalt eines Wortes wieder. Gebräuchlich ist ein Set von sieben verschiedenen Tags, wozu unter anderem die Tags noun, preposition, verb und adjective gehören (vgl. Feldman u. Sanger [2007], S. 60). Bei der vorliegenden Arbeit, bei der insbesondere die Substantive der Artikel analysiert werden sollen, um Rückschlüsse auf den Inhalt der Artikel ziehen zu können, ist die Annotation als nouns im Rahmen des

POS-Tagging enorm wichtig. Bei anderen Mining-Verfahren, bei denen unterschiedliche Positionen und Meinungen oder Emotionen aus Texten analysiert werden sollen, gewinnen auch die Tags für Adjektive und Präpositionen an Relevanz.

- Stemming bedeutet, das Wörter auf ihren Wortstamm reduziert werden. Dieser Schritt ist für die meisten Analysen bedeutsam, damit konjugierte Wörter als identische Wörter erkannt werden.

- Das Löschen von Stopwords dient dazu, die Analyse der Texte auf Wörter zu beschränken, denen eine Signifikanz beigemessen wird. Stopwords sind Wörter wie „der“, „die“, „das“, „aber“, „ob“, „weil“ etc., die nicht erlauben, auf den Inhalt von Texten zu schließen und aus diesem Grund vor der Analyse der meisten Texte als irrelevant gelöscht werden.

4.3 Exemplarische Arbeiten zur automatisierten Erschließung digitaler Dokumentensammlungen

> Der Grundgedanke des datengestützten Journalismus ist es ja, dass die Journalisten dem Publikum relevante, konzentrierte Information verschaffen, die für das Leben der Menschen nützlich ist und ihnen hilft, die Welt zu verstehen. [...] Es geht darum, dass mit der Computertechnik und dem Zugriff auf große Datensätze Journalismus exakter und wissenschaftlicher wird.
>
> Emily Bell (Interviewed von Marti [2015])

Die Frage, wie und welche Textmining-Verfahren eingesetzt werden, um einen gegebenen Textkorpus hinsichtlich bestimmter Fragestellungen zu erschließen und ein Mehrwissen durch die Analyse der Daten zu generieren, kann auf verschiedene Arten beantwortet werden. Sowohl auf Konferenzen über die Auswirkungen des digitalen Wandels auf Medien (Anderson u. a. [2012]) als auch bei der Suche in Bibliotheken[4] findet man diverse Studien und Projekte zum Themenspektrum „automated document collection“, „recommender systems“ und „dossier creation“. Viele bieten Auseinandersetzungen mit verschiedenen Ansätzen zur systematischen Erschließung digitaler, textbasierter Datenbanken sowohl aus journalistischer Perspektive als auch aus der Sicht der Informatik. Exemplarisch werden im Folgenden einzelne Projekte kurz vorgestellt.

[4] Für Ersteres vgl. z. B. die bisherigen Talks der jährlich stattfindenden re:publica: https://re-publica.de, für Letzteres beispielsweise ACM-Library (Association for Computing Machinery): http://dl.acm.org.

Named entity disambiguation in the news domain

Das Verfahren „IdentityRank“ von Fernández u. a. [2012] geht der Frage nach, wodurch – bezogen auf die Analyse von Nachrichtenseiten – die hohe Anzahl an Wörtern, die verschiedene Bedeutungen haben können, unterschieden werden können bzw. wie es gelingen kann, sie ihrem richtigen Kontext zuzuordnen. Bisher besteht das Problem darin, dass Analysemöglichkeiten bzw. -ergebnisse teilweise nicht zufriedenstellend sind, weil sich in den Treffern regelmäßig Texte finden, die aufgrund einer doppelten Bedeutung von Wörtern falsch zugeordnet werden. Das Projekt zur „named entity disambiguation“ setzt auf die Unterstützung von speziell für den Newsbereich entwickelten Ontologien (Gruber [2007]), mithilfe derer eine semantische Analyse und Zuordnung über eine „precise identification of the concept represented by each named entity occurring in a document“ ermöglicht werden und zu besseren Trefferquoten führen soll. IdentityRank bedient sich für die Analyse auch der Metadaten, wie z. B. dem Zeitpunkt der Veröffentlichung, der wichtige Kontextinformationen für die Unterscheidung und Zuordenbarkeit doppeldeutiger Wörter beinhaltet.

Blogosphere: Research issues, tools, and applications

Die inhaltliche Erschließung von Blogs ist Thema der Arbeit von Agarwal u. Liu [2008]. Der Annahme folgend, dass in Blogs ExpertInnenwissen liegt, das geborgen werden sollte, um es einer größeren Öffentlichkeit zugänglich zu machen, haben die Autoren zunächst versucht, eine inhaltliche Erschließung von Blogs anhand der händisch vergebenen Tags vorzunehmen. Auf diese Weise konnte zwar in Form von Kategorien das thematische Spektrum von Blogs erfassbar gemacht werden, eine inhalltiche Erschließung der Themen einzelner Blogposts war so jedoch nicht möglich. In einem nächsten Schritt wurden daraufhin mittels TF-IDF[5] die „top three famous words“ der einzelnen Blogposts analysiert. Das Ergebnis zeigte, dass mit dieser Methode der Inhalt der einzelnen Artikel sehr viel präziser erschlossen werden konnte.

Open domain event extraction from Twitter

Die Arbeit von Ritter u. a. [2012] befasst sich mit der Frage, welche Verfahren dabei helfen können, im Nachrichtenstrom von Twitter wichtige von unwichtigen Ereignissen zu unterscheiden – wie also die gesuchten Informationen aus dem sogenannten „Rauschen“ (engl.:

[5] TF-IDF wird häufig als Maß im Bereich des Information Retrieval eingesetzt, um die Vorkommenshäufigkeit von Wörtern zu ermitteln. Während sich TF, die term frequency, auf die Vorkommenshäufigkeit eines Wortes im zu analysierenden Dokument bezieht, bezieht sich IDF, die inverse document frequency, auf das Vorkommen bzw. die Relevanz eines Wortes für den gesamten Textkorpus.

„noise“) herausgefiltert und kategorisiert werden können. Die kurzen und einfach strukturierten Texte sowie der Umstand, dass es sich um einen „most-up-to-date and inclusive stream of information and commentary of current events“ mit einem sehr hohen Datenaufkommen handelt, ermöglicht es, auf relativ einfache Weise Redundanzen, also identische bzw. häufig wiederholte Informationen, aus dem Nachrichtenstrom herauszufiltern. Das von Ritter u. a. [2012] entwickelte Verfahren, TwiCal, nutzt für die die Kategorisierung von Ereignissen eine „4-tupel-representation“: Indem einem Ereignis vier Objekte zugeordnet werden (named entity, event phrase, calendar date und event type), wird der Versuch unternommen, die Struktur, in der Ereignisse auf Twitter benannt werden, nachzustellen – um auf diese Weise die Texte zu erschließen.

Außer den exemplarisch skizzierten Untersuchungen gibt es unzählige weitere. Viele große Medienhäuser unterhalten zudem seit Jahren eigene Labs, in denen sie speziell auf die Bedürfnisse der eigenen Publikationen ausgerichtete Digitalstrategien entwickeln. Stellvertretend für andere können in diesem Zusammenhang z. B. das Lab der *New York Times* (NYT) und das der *Neuen Zürcher Zeitung* (NZZ) genannt werden[6] sowie das Projekt der *Süddeutschen Zeitung*, automatische Klassifizierung und Visualisierung zur Archiverschließung einzusetzen (Schek [2005]). Im Zusammenhang mit Möglichkeiten der Aufbereitung von Archiven können auch einzelne Projekte von Neofonie [2013] genannt werden: Mit der von ihnen entwickelten „Zeitmaschine“ lässt sich das Digitalarchiv der Wochenzeitung *Die Zeit* gezielt durchsuchen. Man kann sich entsprechend seinen persönlichen Interessensgebieten Artikel anzeigen lassen (http://labs.neofonie.de/zeitmaschine/). Mit WATT, dem Web-Annotations-Tool Demonstrator ist es möglich, eine semantische Analyse von Artikeln bzw. Webseiten vornehmen zu lassen.[7]

[6] Einen Eindruck von den Prototypen der NYT kann man sich hier verschaffen: http://nytlabs.com. Eine Linkliste des NZZ-Labs zum Thema „Structured Journalism“, das Thema auf der re:publica in Berlin im Mai 2014 gewesen ist, kann hier eingesehen werden: http://bit.ly/structured-journalism-links.

[7] Vgl.: http://labs.neofonie.de/watt. Das Verfahren enthält verschiedene Features, wie z. B. eine thematische Zuordnung und Verschlagwortung sowie die Einbindung von Zusatzinformationen.

5 Der Anwendungsfall

Zu Beginn dieser Arbeit stand die Annahme, dass der Einsatz von Textmining-Verfahren eine Unterstützung für Redaktionen bei der Erstellung von Dossiers bieten könnte. Um die Anforderungen präzisieren zu können, wurden ExpertInnen über ihre Definition von und Arbeitsweise mit Dossiers befragt. Dabei zeigte sich, dass es sowohl unterschiedliche Arten von Dossiers als auch entsprechend unterschiedliche Anforderungen an die Erstellung selbiger gibt. Für die Planung einer teilautomatisierten Dossiererstellung mithilfe von Mining-Verfahren muss dementsprechend vorab klar sein, welche Art Archiv bzw. Dokumentenkorpus für eine Dossiererstellung verwendet wird und was für Dossiers erstellt werden sollen. Für Eurozine, dessen englischsprachiges Artikelarchiv als Anwendungsfall dieser Arbeit dient, stellte sich heraus, dass die Spezifik des Archivs u.a. in seiner Diversität besteht: Das Archiv beinhaltet vielfältige Themen, die unterschiedlich aufgegriffen werden – von essayistisch-literarisch bis wissenschaftlich-analytisch. Zudem variiert die Länge der Artikel stark. Das Archiv bildet thematische Ausschnitte der Debatten der Partnerzeitschriften ab, nicht aber die vollständigen Debatten. Es gibt keine Verschlagwortung des Archivs und keine anderweitig bestehende inhaltichen Erschließung des Archivs durch das vorhandene CMS.

Die bestehenden Focal Points entsprechen einer weiten Definition von Dossiers (vgl. Kap. 2.5.7 und 3.3.2): Abhängig von einer Fragestellung beleuchten die Focal Points *entweder* einen einzelnen Aspekt eines Themas *oder* bieten einen historischen Abriss *oder* bündeln konträre Positionen eines Themas. Entsprechend sind die Focal Points Leseangebote, die einen ausdifferenzierten Überblick zu einem Thema abbilden, der bewusst über den Horizont einer einzigen Zeitschrift hinausgeht.

Die Erstellung der Focal Points ist für Eurozine bislang sehr (zeit-)aufwändig – unter anderem, da die Artikel redaktionell zusammengestellt werden. Ein automatisiertes Vorschlagsystem würde eine Unterstützung und Arbeitserleicherung für die Redaktion bedeuten. Auf diese Weise könnte perspektivisch ein Übergang von redaktioneller zu teilautomatisierter Dossiererstellung vollzogen werden, bei dem gewährleistet bleibt, dass bei der Dossiererstellung keine Qualitätsverluste in Kauf genommen werden müssen.

In diesem Kapitel werden die Grundlagen für eine teilautomatisierte Dossiererstellung für das Eurozine-Archiv erarbeitet. Das Ziel ist die Entwicklung eines automatisierten Vorschlagsystems – eine algorithmengesteuerte Artikelauswahl aus dem Gesamtkorpus des Archivs –, das die redaktionelle Entscheidung für Artikelzusammenstellungen vereinfacht.

Die Verfasserin dieser Arbeit hat gemeinsam mit Marcel Schöneberg Kriterien der Dossiererstellung für das Eurozine-Archiv diskutiert und mögliche Anwendungsfälle erarbeitet, die im Folgenden vorgestellt werden. Die interdisziplinäre Auseinandersetzung mit dem Thema – teilautomatisierte Dossiererstellung – diente dem Wissens- und Informationsaustausch zwischen Fachexpertin und Informatiker. Entsprechend der unterschiedlichen Hintergründe sind Anwendungsfälle unterschiedlich verortet: Während Schöneberg die Möglichkeiten teilautomatisierter Dossiererstellung aus Informatikersicht darstellt, werden in diesem Kapitel aus fachlicher Sicht die Erschließungsmöglichkeiten des Eurozine-Archivs betrachtet. Die Grundlagen von Dossiers und die spezifischen Interessen von Eurozine wurden durch die Vorarbeiten (vgl. Kap. 2 und 3) präzisiert und mit Schöneberg abgesprochen, wer welche Verfahren auf der Basis von Bag-of-Words-Ansätzen testen würde. Die Ergebnisse der jeweiligen Untersuchungen wurden wiederum gemeinsam ausgewertet.

Das vorliegende Kapitel teilt sich in vier Abschnitte: Zunächst wird am Beispiel eines bestehenden Focal Points aus dem Eurozine-Archiv eine voralgorithmische Analyse durchgeführt. Die Ergebnisse werden dann mit den Treffern einer Volltextsuche verglichen. Im zweiten Abschnitt wird eine Tag-Cloud zur Analyse von Artikeln bzw. Focal Points eingesetzt. Im dritten Teil werden die Arbeit bzw. die Untersuchungsergebnisse von Schöneberg vorgestellt. Im vierten Abschnitt erfolgt eine kurze Skizzierung der Möglichkeiten von Overview, einem Mining-Tool, mit dem große Textmengen geclustert werden können. Alle Verfahren gehen der Frage nach, ob es über Bag-of-Words-Ansätze möglich ist, Vorschlagsysteme für die Erstellung von Focal Points zu konzipieren, deren Qualität der der bereits bestehenden, redaktionell erstellten Focal Points entspricht.

5.1 Voralgorithmische Analyse – händisches Auszählen

Ausgangspunkt der voralgorithmischen Analyse war die Annahme, dass das Vorkommen eines oder mehrerer Substantive aus dem Titel eines Focal Points in den zugehörigen Artikeln überdurchschnittlich hoch sein könnte. Grundlage dieser Annahme war, dass die Titel der Focal Points häufig „sprechende Titel“ sind, die sich aus mehreren Wörtern zusammensetzen.

Die Untersuchung wurde exemplarisch für den Focal Point „The ends of democracy“ vorgenommen, der aus 29 Artikeln aus den Jahren 2001 bis 2013 besteht.[1] Die Artikel thematisieren Demokratiefragen inner- und außerhalb der EU. Die Themen der einzelnen Artikel reichen vom Klimawandel über die Gezi-Proteste in Istanbul und Überwachung bis hin zu Perspektiven demokratischer Staaten.

Die Länge der Artikel variiert stark: Der kürzeste Artikel besteht aus 964 Wörtern, der längste aus 6.911. Die durchschnittliche Artikellänge beträgt 4.004 Wörter.[2] Die unterschiedliche Länge der Artikel weist bereits darauf hin, dass die Ermittlung des relativen Vorkommens des gesuchten Wortes relevanter ist als die absoluten Häufigkeiten. Aus fachlicher Perspektive war offensichtlich, dass das Substantiv „democracy“ der zentrale Begriff aus dem Titel des Focal Points ist.[3] Die Analyse bestand darin, die Häufigkeit des Vorkommens von „democra*“, dem Stamm des Wortes „democracy“, für alle Artikel des Focal Points zu ermitteln.[4]

Untersucht wurde, ob eine Korrelation zwischen dem Vorkommen eines einzelnen Wortes in einem Artikel und seiner redaktionell bestimmten Zugehörigkeit zu einem Focal Point existiert. Sollte das Vorkommen des zentralen Begriffs in den Artikeln des Focal Points überdurchschnittlich hoch sein, könnte es sich lohnen, das komplette Archiv in dieser Richtung genauer zu untersuchen. In diesem Fall würde ein einziges Wort das diskriminierende Kriterium für die Zusammenstellung von Dossiers darstellen.

Um eine Vergleichbarkeit der Artikel zu schaffen, wurde für diese erste Untersuchung ein eindimensionaler Feature-Vektor als Metabeschreibung des Texts verwendet, bei dem sich die Distanz der Artikel zueinander an der ermittelten Häufigkeit des Vorkommens von „democra*“ bemisst.[5]

Untersucht wurde, wie hoch das Vorkommen des Wortes „democra*“

- in den Titeln/Untertiteln und Abstracts der Artikel des ausgewählten Focal Points – also an prominenten Stellen – und
- in den eigentlichen Texten der einzelnen Artikeln ist.

[1] http://www.eurozine.com/comp/focalpoints/democracy.html. Eine Liste der Artikel findet sich im Anhang in Kap. 7.3

[2] Die ermittelte Dateilänge entspricht jeweils der Gesamtwortzahl der XML-Dateien (und nicht nur der Anzahl der im body-Tag der XML-Datei enthaltenen Wörter.

[3] Wenngleich diese Entscheidung für den vorliegenden Focal Point zweifelsfrei getroffen werden konnte, muss festgehalten werden, dass sie sich nicht ohne Weiteres automatisieren lässt.

[4] Einbezogen waren also die Wörter „democracy“, „democratic“, „democratization, „non-democratic“ etc.

[5] Diese Arbeit mit dem Zweck einer Voruntersuchung beschränkt sich zunächst bewusst auf einen aus einem Wort bestehenden Bag-of-Words. Schöneberg hat parallel mit dem gleichen Korpus mit einem n-dimensionalen Feature-Vektor vorgenommen, bei dem zunächst die Häufigkeitsverteilung aller vorkommenden Wörter ermittelt wurde. Schöneberg bilanziert, dass der Feature-Raum dabei unübersichtlich wird und eine Reduktion des Feature-Vektoren hilfreich ist (vgl. Schönebergs Arbeit, S. 45).

Für beides wurde zudem anschließend unter Hinzuziehung von fünf zufällig ausgewählten Artikeln eine Gegenstichprobe gemacht, um das Ergebnis zu überprüfen. Gezeigt werden soll, ob es sich bei den ermittelten Worthäufigkeiten um signifikante Ergebnisse handelt.

5.1.1 Häufigkeit des Vorkommens von „democra*“ in Titeln und Abstracts der Artikel des Focal Points

Für jeden Artikel des Focal Points wurde ermittelt, wie häufig das gesuchte Wort im Titel/Untertitel der Artikel vorkommt – wobei die Tags „title“ und „subtitle“ als Einheit betrachtet wurden – und wie oft im Abstract („blurb“).

Das Auszählen der absoluten Häufigkeiten ergab (vgl. Abb. 5.1), dass

- 14 von 29 Titel den Wortstamm „democra*“ enthalten (48%) und
- 21 von 29 Abstracts den Wortstamm „democra*“ enthalten (72%).

Das bedeutet, dass bei 25 von 29 Artikeln – also einem Anteil von 86 % – das Wort „democra*“ entweder im Titel oder im Abstract enthalten ist. Nur vier der 29 Artikel (knapp 14 %) enthalten „democra*“ weder als Teil des Titels noch des Abstracts.[6]

Das Ergebnis dieser Analyse stützt die Hypothese, dass Titel bzw. Abstracts von Artikeln Wörter enthalten, die auf das Thema des Artikels hinweisen und somit Titel und Abstracts ggfs. zur Ermittlung hinsichtlich diskriminatorischer Wörter / Merkmale für eine Zuordenbarkeit von Artikeln zu einem Focal Point verwendet werden können. Ob diese Korrelation jedoch weitergehende Aussagen erlaubt, muss durch eine Gegenstichprobe überprüft werden.

5.1.2 Gegenstichprobe – Häufigkeit des Vorkommens von „democra*“ in Titeln und Abstracts zufällig ausgewählter Artikel

Für die Gegenstichprobe, mit der die Aussagekraft der vorangegangenen Ergebnisse überprüft werden sollen, wurden fünf randomisiert ausgewählte Artikel des englischsprachigen Teil des Artikelarchivs auf das Vorkommen von „democra*“ in Titeln und Abstracts analysiert.[7] Das Auszählen der absoluten Häufigkeiten ergab, dass

- „democra*“ in keinem der fünf Artikel im Titel enthalten ist und

[6] Die Artikel, deren Titel und Abstracts das gesuchte Wort nicht enthalten, weisen allerdings Wörter auf, die dem thematischen Feld Demokratie aus fachlicher Sicht zugeordnet werden können. Es handelt sich um Wörter wie z. B. state, movement, global justice, participatory, transnationality, citizenship und communities. Für einen Ansatz, der sich mit dem Einsatz von Ontologien zur teilautomatisierten Dossiererstellung befassen würde, wäre dieser Aspekt von Bedeutung.

[7] Eine Liste der Artikel findet sich im Anhang, s. Kap. 7.4.

- nur in einem Artikel im Abstract vorkommt.[8]

Die Gegenstichprobe bestätigt also die Vermutung, dass das zentrale Wort des Themas des analysierten Focal Points signifikant häufigker in den Artikeln des Focal Points vorkommt als in zufällig ausgewählten anderen Artikeln des Archivs. Dies zeigt, dass bereits ein eindimensionaler Feature-Vektor, der sich auf ein einziges Wort bzw. sein Vorkommen in Titeln und Abstracts beschränkt, eine Möglichkeit bietet, die Distanz von Artikeln zueinander zu ermitteln. Auf diese Weise kann ggfs. eine valide Vorsortierung von Artikeln für eine Weiterverarbeitung zu einem Dossier vorgenommen werden.

5.1.3 Häufigkeit des Vorkommens von „democra*“ in den vollständigen Dateiinhalten des Focal Points

Im nächsten Schritt wurde das Vorkommen von „democra*“ in den vollständigen Dateien der 29 Artikel des Focal Points ermittelt[9] – unter anderem um zu sehen, ob das Wort in den Artikeln, in denen es nicht in Titel und Abstract vorkam, enthalten ist. Dies könnte zur Bewertung führen, dass eine Analyse von sehr kurzen, wenngleich prägnanten Textstellen wie Titel und Abstract nicht unbedingt ausreicht, um diskriminatorische Merkmale von Texten für die Erstellung von Dossiers zu ermitteln. Die Analyse der Texte ergab, dass

- in jedem der 29 Artikel des Focal Points das Wort „democra*“ enthalten ist. Die Häufigkeit des Vorkommens variiert zwischen sechs- und 121-mal pro Artikel. Das entspricht einer durchschnittlichen Häufigkeit von 38 Erwähnungen pro Artikel.

- in den Artikeln, in denen „democra*“ im Titel oder im Abstract enthalten war, überdurchschnittlich oft im Text verwendet wurde (zwischen 31- und 121-mal pro Artikel).

- in den vier Artikeln, in denen „democra*“ nicht im Titel oder Abstract enthalten war, das Wort mit einer Häufigkeit zwischen neun- und 23-mal pro Artikel deutlich seltener auftaucht als in den übrigen Artikeln – sowohl bezogen auf die absolute als auch auf die prozentuale Häufigkeit (vgl. Abb. 5.1).[10]

[8] Aufgefallen ist bei dieser Auszählung zudem, dass ein Artikel kein Abstract hat (2006-11-26-sluga-en). Diese Beobachtung stellt einmal mehr heraus, welche Schwierigkeiten die uneinheitliche Struktur eines Archivs und ein nicht valides XML mit sich bringen. Sofern mit dem gesamten Korpus gearbeitet werden würde, müsste entschieden werden, wie in einem solchen Fall verfahren werden soll, damit sichergestellt werden kann, dass der Artikel nicht bei der Analyse ignoriert wird.

[9] Analysiert wurden dabei die kompletten XML-Datensätze der Artikel

[10] Insgesamt 23-mal wird das Wort bei Fraser verwendet. Diese absolute Häufigkeit entspricht bei einer Dateilänge von 6.355 Wörtern einem Anteil von 0,36% Vorkommen des Wortes im Verhältnis zur Gesamtwortzahl des Artikels. 11 Verwendungen bei Leggewie. Das entspricht bei einer Dateilänge von 6.911 Wörtern einem Anteil

F.P. Democracy Suchwort: democra*	Vork im Titel	Vork im Abstr.	Vork im ges. Art.	Wortzahl des Art.	prozent. Anteil
2001-11-27-rosenberg	1	2	67	5389	1,24
2008-05-02-wennerha	0	2	55	4027	1,37
2008-11-21-leggewie\	1	0	41	3002	1,37
2009-04-21-fraser-en	0	0	23	6355	0,36
2009-07-14-biscione-	1	0	31	6089	0,51
2009-09-09-kavaliaus	0	1	47	4379	1,07
2010-09-14-ditchev-e	1	0	23	3147	0,73
2011-07-11-bluhdorn-	2	2	121	4968	2,44
2011-11-02-G1000-er	0	1	55	3965	1,39
2011-11-10-sierakows	0	3	23	3831	0,60
2011-12-19-amirpur-e	1	1	79	4008	1,97
2012-01-25-halmai-er	1	2	24	5618	0,43
2012-02-08-elsenhan	1	1	15	3748	0,40
2012-09-05-jahanbeg	0	0	9	2342	0,38
2012-11-21-holmes-e	0	2	91	4975	1,83
2013-02-01-krastev-e	0	2	56	5421	1,03
2013-02-08-wallerstei	0	1	6	2797	0,21
2013-02-19-leggewie-	0	0	11	6911	0,16
2013-02-26-james-en	0	2	15	2217	0,68
2013-05-03-muller-en	1	2	54	4168	1,30
2013-06-14-pomeranl	0	1	16	4983	0,32
2013-07-29-gole-en	1	2	41	4757	0,86
2013-08-13-krastev-e	1	1	38	2039	1,86
2013-08-20-leggewie	1	0	32	4992	0,64
2013-09-11-deniztekir	0	1	9	964	0,93
2013-11-08-vidanava-	0	1	9	2197	0,41
2013-11-22-offe-en	1	2	43	3090	1,39
2013-12-12-margetts-	0	0	9	3897	0,23
2013-12-12-pogonyi-e	1	3	56	1854	3,02
Summe	**15**	**35**	**1099**	**116130**	
Mittelwert	0,51724	1,2069	38	4004	**0,95**

Abbildung 5.1: Häufigkeit des Wortvorkommens im Titel, im Abstract und in den gesamten Dateiinhalten

von 0,16%. Jeweils neun Verwendungen bei Jahanbegloo und bei Margetts. Das entspricht bei einer Dateilänge von 2.342 bzw. 3.897 Wörtern einem Anteil von 0,38% bzw. 0,23%. Alle vier Artikel, die das gesuchte Wort nicht in Titel oder Abstract aufwiesen, weisen also auch eine unterdurchschnittlich häufige Verwendung des Wortes in den Artikeln auf. Die durchschnittliche Häufigkeit des Vorkommens des gesuchten Wortes in den Artikeln des Focal Points liegt bei 0,95 %.

Abbildung 5.1 zeigt die absolute und die prozentuale Häufigkeit des Vorkommens. Die beiden anschließenden Diagramme zeigen das absolute (5.2) und das relative (5.3) Vorkommen des Wortes in den Artikeln des Focal Points. Das erste Diagramm stellt die Häufigkeitsverteilung auf Basis der absoluten Zahlen, also dem absoluten Vorkommen des gesuchten Wortes, dar. Die unterschiedliche Länge der Texte wird in dieser Berechnung ignoriert, was ggfs. zu falschen Bewertungen der Artikel hinsichtlich des Vorkommens von „democra*“ führen könnte.

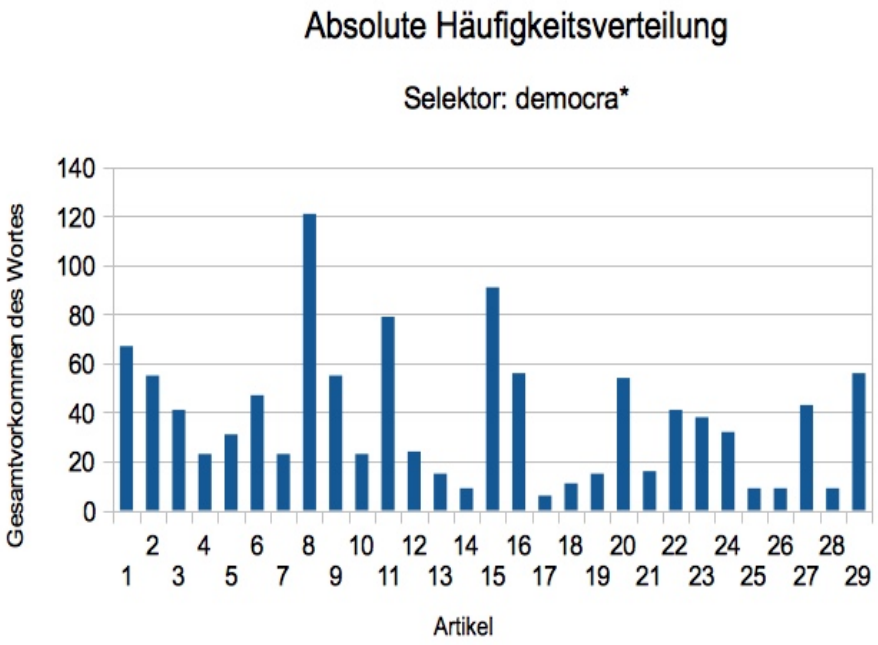

Abbildung 5.2: Absolute Häufigkeitsverteilung des Wortes democra* in den Artikeln des Focal Points

Der kürzeste Text des Focal Points hat eine Länge von 964 Wörtern (deniztekin), der längste besteht aus 6.911 Wörtern (leggewie). Aufgrund dieser gravierenden Unterschiede der Artikellängen schien es erforderlich, ergänzend zur absoluten die relative Verteilung des Vorkommens von „democra*grqq zu ermitteln – und die Distanz der Texte mittels eines Feature-Vektoren zu berechnen, der der relativen Häufigkeit des Vorkommens des Wortes in einem Artikel entspricht. Die Berechnung und Darstellung der relativen Häufigkeit im zweiten Diagramm bietet eine validere Basis für einen aussagekräftigeren Vergleich der Artikel miteinander.

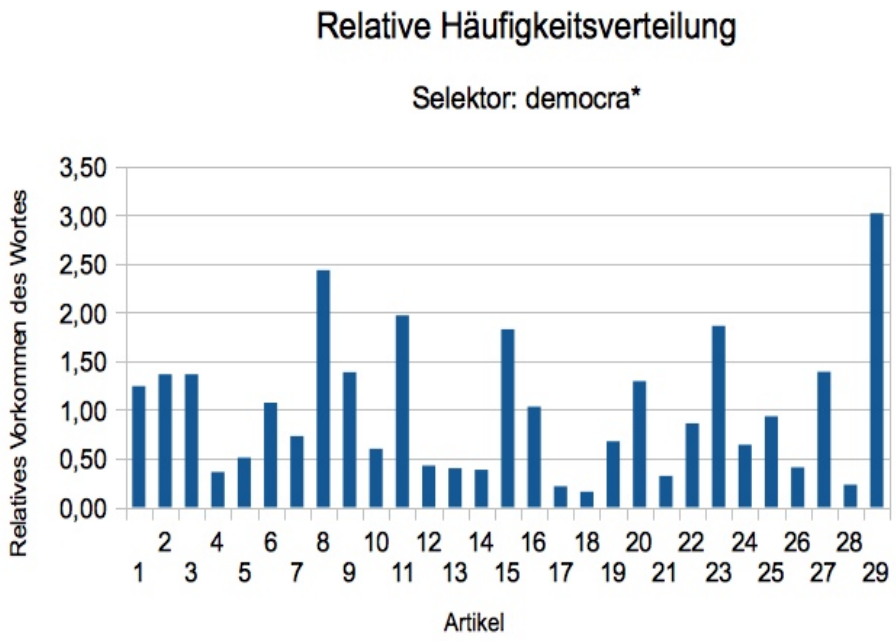

Abbildung 5.3: Relative Häufigkeitsverteilung des Wortes democra* in den Artikeln des Focal Points

Auffällig sind – aufgrund der stark variierenden Länge der Artikel – die Unterschiede zwischen den beiden Diagrammen. Während bspw. bei den Artikeln elf, 15 und 20 die relative Häufigkeit in etwa der relativen entspricht, ist bei Artikel 29 ein gravierender Unterschied in der Bewertung relativ versus absolut zu sehen.

Der Vergleich der beiden Diagramme zeigt deutliche Unterschiede: Während bei der Berechnung absoluter Häufigkeiten bezogen auf das Wort „democra*“ die Artikel acht, elf und 15 die höchsten Werte aufweisen, sind es bei der relativen Häufigkeitsverteilung die Artikel scht, elf und 29.

5.1.4 Gegenstichprobe – Häufigkeit des Vorkommens von „democra*“ in den vollständigen Dateiinhalten zufällig ausgewählter Artikel

Das Auszählen des absoluten Vorkommens von „democra*grqq in den Artikeln der Gegenstichprobe ergab, dass

- nur drei von fünf Artikeln das Wort beinhalten,

- in einem der drei Artikel das Wort einmal vorkommt (gellner), in einem anderen fünfmal (santos) und im letzten Artikel zweimal (ferrard).

Setzt man diese absoluten Zahlen ins Verhältnis zur Länge der Artikel, entspricht das Vorkommen einer Häufigkeit von 0,07% pro Artikel. Damit ist festgestellt, dass die zufällig ausgewählten Artikel das Wort „democra*“ siginifikant seltener enthalten als die Artikel des Focal Points, die ein Vorkommen von durchschnittlich 0,95% aufwiesen.

5.1.5 Bewertung der voralgorithmischen Analyse

Die voralgorithmische Analyse eines Focal Points hat gezeigt, dass eine Korrelation zwischen dem zentralen Begriff des Themas eines Focal Points und der Häufigkeit der Verwendung des gleichen Begriffs in den einem Focal Point zugeordneten Artikeln zu bestehen scheint. Das Wort, das als Selektor verwendet wurde, wird überdurchschnittlich häufig in den Artikeln des Focal Points verwendet, während es in zufällig ausgewählten Artikeln äußerst selten auftaucht. Dies lässt die Annahme zu, dass Ähnliches für weitere Focal Points festgestellt werden könnte. Sollte dem so sein, wäre eine Analyse auf der Basis eines kleinen Bag-of-Words-Ansatzes für den gesamten Artikelkorpus evtl. eine effiziente Möglichkeit, das Archiv zu erschließen und ein Vorschlagsystem für die Erstellung von Dossiers zu entwickeln.

Bei einer erweiterten Analyse sollte jedoch u.a. Folgendes berücksichtigt werden:

- Die bisherige Analyse bezog sich auf das Vorkommen eines einzigen Wortes. Um herauszufinden, welche anderen Wörter siginfikant häufig in den Artikeln verendet werden und bei der thematischen Erschließung des Archivs unterstützen könnten, muss mit einem größeren Feature-Vektor gearbeitet bzw. der verwendete Bag-of-Words vergößert werden. Ggfs. ist es auch sinnvoll, innerhalb eines komplexeren Verfahrens Wortkombinationen zu berücksichtigen. Die Artikel im Focal Point „The ends of democracy“ thematisieren bspw. die politischen Entwicklungen von Protestbewegungen in der Türkei und die Regierungspolitik in Ungarn, präsentieren Interviews mit Intellektuellen etc. Es wäre sinnvoll, das Wort democracy mit anderen Wörtern, wie z. B. Gezi, Protest, Hungary, Islam, Nation etc., zu kombinieren, um zu aussagekräftigeren Ergebnissen hinsichtlich der Distanz von Artikeln zueinander zu kommen.

- Die (absolute bzw. relative) Worthäufigkeit ist zwar im vorangegangenen Beispiel als Kriterium zur Analyse von Texten herangezogen worden, kann aber nicht ohne Weiteres als Kriterium zur qualitativen Beurteilung eines Textes hinsichtlich seiner Zugehörigkeit zu einem Focal Point verwendet werden.

- Ein Wordcounting sollte nicht nur über die Titel und Abstracts laufen gelassen werden, sondern über die gesamten Dateiinhalte. Auf diese Weise könnten auch thematische Cluster entstehen, mit denen weitergearbeitet werden könnte

- Ein Fokussieren auf die der Häufigkeit des Vorkommens einzelner Wörter ignoriert, dass insbesondere auch selten vorkommende Wörter in Texten diskriminatorische Bedeutung haben und für einen Focal Point von besonderer Relevanz sein können. Ein typischer Fehler besteht darin, dies nicht zu berücksichtigen. Während es bei Vorschlagsystemen schnell möglich ist, fälschlicherweise in die Auswahl gerutschte Artikel (sog. „false positives“) auszusortieren, ist es nahezu unmöglich, mit der vorgeschlagenen Methode sogenannte „false negatives“ zu finden – Texte also, die eigentlich in die Vorauswahl gehören, aber aufgrund der Kriterien durch eine automatische Suche nicht gefunden werden konnten. Sie setzen die genaue Kenntnis des gesamten Textkorpus voraus, aus dem ausgewählt wurde.

5.1.6 Einsatz der Volltextsuche zum Vergleich der Ergebnisse der voralgorithmischen Analyse

Nachdem die Analyse der Artikel des Focal Points „The ends of democracy“ mit einem eindimensionalen Feature-Vektor als sinnvolle Erschließungsmethode erkannt worden ist, wurde die auf der Website vorhandene Volltextsuche für einen Vergleich herangezogen, um einen Eindruck davon zu bekommen, welche Ergebnisse mit dieser Methode erzielt würden. Die Volltextsuche nach „democra“ (mit und ohne Asterisk) ergab keinen einzigen Treffer. Offensichtlich erkennt die Volltextsuche von Eurozine keine Wortteile, sondern nur vollständige Wörter. Sucht man nach dem ganzen Wort – „democracy“ – bekommt man 1.545 Treffer.[11] Die Treffer setzen sich aus Artikel, Reviews und Heftzusammenfassungen zusammen. Eine Eingrenzung der Trefferanzahl ist derzeit nicht möglich. Es gibt zwar die Möglichkeit, bei den zur Auswahl stehenden Suchoptionen <author>, <language> und <date> in Letzterem als Veröffentlichungsdatum „2015“ anzugeben. Damit reduziert sich die Anzahl der Treffer auf 191. Allerdings muss dann festgestellt werden, dass die Treffer keine für das gesuchte Wort, sondern die im angegebenen Jahr veröffentlichten Artikel sind.

Die Implementierung einer funktionsfähigeren Volltextsuche, die nicht nur vollständige Wörter, sondern auch Wortteile erkennt, ebenso wie eine funktionierende Eingrenzung der Suche könnte sowohl für die LeserInnen von Eurozine als auch für die Redaktion eine Un-

[11] Stand: 16. April 2015

terstützung bedeuten. Wünschenswert wäre, eine Suche zu implementieren, in die ein Wort eingegeben kann, jedoch nach dem Vorkommen des Wortstamms gesucht wird.

5.2 Inhaltliche Erschließung des Eurozine-Archivs mittels einer Tag-Cloud

In einem nächsten Schritt wurden die Erschließungsmöglichkeiten des Archivs mittels eines Programms vorgenommen, das die Tag-Cloud-Technologie benutzt.[12] Dabei wird ein Set vorab festgelegter Artikel auf die Häufigkeiten der darin vorkommenden Wörter analysiert und visualisiert: Die in den Artikeln am häufigsten verwendeten Wörter werden in unterschiedlicher Größe als Tag-Cloud dargestellt, wobei die am häufigsten vorkommenden Wörter größer als die seltener vorkommenden angezeigt werden.

Während bei dem vorherigen Verfahren, der voralgorithmischen Analyse, die Artikel nur hinsichtlich des Vorkommens eines einzelnen Wortes untersucht wurden, werden in diesem Verfahren die Häufigkeiten aller in den Artikeln vorkommenden Wörter ermittelt. Die Untersuchung wird dadurch komplexer: Die Berechnung von Feature-Vektoren für die einzelnen Artikel, die vorher eine einzige Dimension (für ein einziges Wort) beinhalteten, wird bei der Tag-Cloud zugunsten einer Mehrdimensionalität abgelöst: Alle Wörter werden gezählt und in die Berechnung des Feature-Vektoren, der für den Vergleich der Artikel herangezogen wird, einbezogen. Der deutlich größere, n-dimensionale Feature-Vektor verschafft einen viel differenzierteren Eindruck der Artikel und damit eine deutlich differenziertere Grundlage, um die Artikel hinsichtlich ihrer Distanz miteinander zu vergleichen.

Herausgefunden werden sollte mit dieser Methode zum einen, inwieweit Artikel anhand der Häufigkeit des Vorkommens verschiedener Wörtern charakterisiert werden können. Zum anderen sollte untersucht werden, ob dies eine gute Methode zur Bewertung von Distanzen zwischen Artikeln darstellt und eine Vorsortierung bzw. Erschließung des Archivs auf Basis der Worthäufigkeiten erfolgen könnte. Angenommen wurde, dass der Einsatz des Tag-Cloud-Programms besser als die vorherige Methode durch die Extraktion häufig vorkommender Begriffe sowohl die Themen von Artikel sichtbar machen kann (und bei einer Analyse des gesamten Archivs auch die prominentesten Themen des Archivs) und diese Methode somit als ein Schritt bei der Erstellung teilautomatiierter Dossiers ermöglichen könnte, das thematische Spektrum verschiedener Dateien oder eines ganzen Korpus darzustellen.

[12] Das Tag-Cloud-Tool, das für diese Arbeit zum Einsatz kam, tagcloud-0.1, geht zurück auf Gespräche mit Tobias Eichler und Ali Rahimi über die Anforderungen an Archive und Dossiers. Vielen Dank an beide und vor allem an Tobias für die Vorbereitung des Korpus sowie das Schreiben des Programms entsprechend der Vorgaben der Verfasserin.

Das hier eingesetzte Programm „tagcloud-0.1“ führt eine Datenvorbereitung durch (Stopwords löschen, Feature-Auswahl etc.).[13] Im Anschluss daran wird das Wordcounting durchgeführt – wahlweise für alle Wörter (abzüglich der Stopwords) oder nur für alle Substantive: Die Häufigkeit des Vorkommens der Wörter wird ermittelt und visualisiert. Die Anzahl der dargestellten Wörter wird vorab entschieden, die Größe der einzelnen Wörter ist abhängig von der Häufigkeit des Vorkommens.

Zur Funktionswiese des Programms:[14] Beim Programmaufruf über die Konsole wird für jede zu erstellende Tag-Cloud entschieden,

- ob alle Wörter gezählt werden sollen oder nur Substantive,
- ob die Wörter gestemmt (auf ihren Wortstamm reduziert) oder die vollständigen (deklinierten) Wörter analysiert werden sollen.[15]
- aus wie vielen Wörtern die Tag-Cloud (html-Datei) bestehen soll. Bei den durchgeführten drei Durchgängen wurden 20 Begriffe angezeigt.
- Die Tag-Cloud wird nach Ablauf des Programms als HTML-Datei gespeichert und im Browser aufgerufen.
- Beim Mouse-over über die einzelnen Begriffe der Tag-Cloud wird angezeigt, wie häufig der Begriff im untersuchten Datenset vorkommt.

Die Erstellung einer Tag-Cloud wurde zunächst für den Focal Point „The ends of democracy“ durchgeführt, der bereits für die voralgorithmische Analyse verwendet wurde. Im Anschluss daran wurde der gesamte englischsprachige Korpus von Eurozine als Tagcloud visualisiert. Letzteres verfolgte zwei unterschiedliche Absichten:

- Der Vergleich der Tag-Cloud des gesamten Korpus mit der des Focal Points sollte zeigen, ob die häufigsten Wörter des Focal Points sich siginfikant von denen des gesamten Korpus unterscheiden. Sollte dem so sein, würde es sich lohnen, die Artikel des Focal Points auch einzeln auf ihre Signifikanz hinsichtlich häufig verwendeter Wörter zu untersuchen.

[13] Für das Löschen der Stopwords wird auf eine im Programmablauf hinterlegte Liste englischer Stopwords referenziert.

[14] Vorab werden die zu analysierenden Daten in einen in einen Ordner kopiert, auf den das Programm zugreift.

[15] Der Vorteil des Stemmens ist, dass Singular und Plural eines Wortes als gleich erkannt werden, der Nachteil besteht darin, dass nach dem Stemming die Wörter auch als Wortstämme in der Tag-Cloud erscheinen, was die Lesbarkeit u.U ungünstig einschränkt.

- Auch könnten die Ergebnisse des Vergleichs Überlegungen befördern, häufig verwendete Wörter zur Bildung von Kategorien einzusetzen und so eine nachträgliche automatisierte Verschlagwortung des Archivs zu bewerkstelligen.

5.2.1 Die Tag-Cloud des Focal Point „The ends of democracy“

Die Artikel des Focal Points „The ends of democracy“ wurden als eine Tag-Cloud aus 20 gestemmten Substantiven dargestellt. Die überschaubare Anzahl an Begriffen diente der Übersichtlichkeit und Fokussierung auf die häufigsten Wörter. Selbiges galt für die Eingrenzung auf Substantive: Die Annahme war, dass Substantive Themen wiedergeben (wenngleich Facetten oder Positionen eines Artikels auf diese Weise nur schlecht analyisert werden können). Das Reduzieren der Wörter auf den jeweiligen Wortstamm bezweckte die Vermeidung von Dopplungen in der Tag-Cloud, die durch eine häufige Verwendung von Singular und Plural, beispielsweise des Wortes democracy, hätten auftauchen können.[16]

Untersucht wurde, ob a) die Tag-Cloud das Ergebnis der voralgorithmischen Analyse (die häufige Verwendung des Wortes „democracy“) bestätigt und b) welche weiteren Wörter in den Artikeln des Focal Points häufig verwendet werden.

Abbildung 5.4: Die Tag-Cloud des Focal Points „The ends of democracy“ – 20 häufigste Wörter, gestemmt, nur Substantive

[16] Ergänzend wurde eine zweite Tag-Cloud mit ungestemmten Wörtern erstellt, auf die in dieser Arbeit nicht näher eingegangen wird. Sie war zwar einfacher zu lesen, gab aber, wie zu vermuten war, ein ungenaueres Ergebnis wieder.

Die Tag-Cloud des Focal Points „The ends of democracy“ zeigte:

- Das bei Weitem am häufigsten verwendete Wort in den Artikeln des Focal Points ist das Wort „democraci“. Es wird insgesamt 487-mal verwendet.[17]
- Weniger als halb so oft finden sich die Worte „govern“ (218-mal), „state“ (206-mal), „societi“ (180-mal) und „movement“ (163-mal) in den Artikeln.
- Die Annahme, dass thematisch bedeutende Wörter in den Artikeln des Focal Points häufig benutzt werden, bestätigt sich durch die Tag-Cloud. Die 20 häufigsten verwendeten Wörter geben aus fachlicher Sicht die Idee dessen wieder, was bereits in der Charakterisierung der Artikel beschrieben worden ist, und vermitteln in diesem Sinne einen thematischen Eindruck des Focal Points.

Ein Problem beim Einsatz dieser Methode ist, dass keine Zurodnung der Wörter zu den einzelnen Artikeln möglich ist. Würde man weitergehend mit der Tag-Cloud arbeiten wollen, wäre die Ergänzung einer Funktion, mittels der angezeigt werden könnte, wie häufig welches Wort in welchem Artikel zu finden ist, sinnvoll.

Ergänzend sollte über eine Tag-Cloud zur Visualisierung selten verwendeter Begriffe zur Erschließung eines Archivs nachgedacht werden, die einen ähnlichen Zweck verfolgen würde wie der Einsatz von TF-IDF: Die Berechnung der term frequency (TF) wird häufig ergänzt um die Berechnung der inverse term frequency (IDF), um nicht nur die häufigsten, sondern auch explizit selten auftauchende Begriffe zu berücksichtigen. Wie die Funktionsweise einer solchen Tag-Cloud genau aussehen müsste, übersteigt den Umfang dieser Arbeit und ist hier nur als Hinweis gedacht, der dazu beitragen soll, sogenannte „weak signals“ nicht zu übersehen.

5.2.2 Die Tag-Cloud des englischsprachigen Artikelarchivs von Eurozine

In einem nächsten Schritt wurde eine Analyse des gesamten englischsprachigen Artikelarchivs von Eurozine und eine Visualisierung der am häufigsten verwendeten Wörter als Tag-Cloud vorgenommen. Analog zur Gegenstichprobe bei der voralgorithmischen Analyse sollte in diesem Fall die Qualität der Ergebnisse der Focal-Point-Tag-Cloud durch eine Tag-Cloud des gesamten Archivs beurteilt werden können. Die Tag-Cloud des Archivs wurde aus den häufigsten 100 (gestemmten) Substantiven erstellt. Gezeigt werden sollte, ob es Übereinstimmungen zwischen den Wörtern der Archiv-Tag-Cloud und der Tag-Cloud des Focal Points gibt

[17] Hinterfragt werden müsste bei einer tiefergehenden Auseinandersetzung mit der Tag-Cloud, wodurch die Differenz zwischen dieser Zahl und dem Vorkommen des gleichen Wortes in der voralgorithmischen Analyse zustande kommt. Dort war democra* insgesamt 1.099-mal in den Artiken des Focal Points gefunden worden.

bzw. ob und wodurch sich die Tag-Clouds voneinander unterscheiden, Zudem sollte betrachtet werden, ob die Tag-Cloud der 100 meistbenutzten Wörter des Archivs Aussagen hinsichtlich des thematischen Spektrums des Archivs zulassen würde, die wiederum für eine weitre Erschließung des Archivs hilfreich wären.

Abbildung 5.5: Die Tag-Cloud des englischsprachigen Artikel-Archivs – mit den 100 häufigsten Wörtern, gestemmt, nur Substantive

Die Tag-Cloud, deren Erstellung aufgrund des Umfangs der Daten deutlich länger dauerte (der Rechner benötigte ca. fünf Minuten, um die 2.697 Dateien zu analysieren.), zeigte:

- Am häufigsten verwendet werden die Wörter „time“ (10.581-mal), „world“ (9.297-mal), „state“ (8.985), „wai“ (8.488-mal), „articl“ (7.452-mal) und „societi“ (6.651-mal). Es folgen „democraci“ (4.043-mal) und „univers“ 2.076-mal.
- Alle 100 am häufigsten im Archiv vorkommenden Substantive sind jeweils häufiger als 1.700mal in den Artikeln des Archivs vorzufinden.
- Im Unterschied zur Analyse des Focal Points kann hier festgestellt werden, dass die prominentesten Wörter des Archivs inhaltlich weitaus weniger prägnant für die thematische Bestimmung des Archivs sind, als es beim Focal Point festgestellt werden konnte. Es ist quasi nicht möglich, mithilfe der Tag-Cloud auf das thematische Spektrum des Archivs zu schließen. Allerdings lässt sich aus der Betrachtung der gesamten Tag-Cloud sehr wohl ein Spektrum von Themen an, das für Eurozinevspezifisch ist, ableiten.

Beim Vergleich der beiden Tag-Clouds zeigt sich, dass die diskriminatorische Relevanz der Tag-Cloud für die inhaltliche Erschließung des Focal Points gering ist: Die zahlreichen Übereinstimmungen von Wörtern in beiden Tag-Clouds (abgesehen von „democraci“ betrifft dies die Wörter „state“, „life“, „countri“, „cultur“, „system“, „articl“, „movement“, „govern“ etc.) deuten darauf hin, dass die Ermittlung von Worthäufigkeiten nur bedingt eine Unterstützung bei der Charakterisierung von Texten oder Textsammlungen ist. Ggfs. würde auch hier das Einbeziehen von weak signals im Sinne der Berücksichtigung selten auftauchender Begriffe bessere Ergebnisse und eine klarere Charakterisierung von Texten ermöglichen.

5.2.3 Bewertung des Einsatzes einer Tag-Cloud

Die Untersuchung hat einen ersten Eindruck davon vermittelt, was die Vorteile eines Word-Counting in n Dimensionen gegenüber einem Word-Counting auf Basis eines einzigen Wortes sind. Mithilfe eines Programms, das Tag-Clouds der häufigsten Wörter eines Artikels, eines Focal Points oder eines ganzen Archivs visualisiert, wird eine maschinelle Erschließung von Dateien ermöglicht. Dies wiederum ist eine gute Basis für einen Vergleich von Dateien miteinander: Je stärker sich die entstandenen Tag-Clouds ähneln, desto ähnlicher bzw. „näher“ sind sich die miteinander verglichenen Artikel bzw. Dateien, und je stärker sich die Tag-Clouds voneinander unterscheiden, desto eher kann davon ausgegangen werden, dass die Dateien der einzelnen Tag-Clouds eigene thematische Cluster bilden. Würde diese Untersuchung fortgeführt, wäre ein wünschenswertes Ergebnis, dass der Vergleich der Tag-Clouds der restlichen Focal Points anders als der bereits analysierte kaum Übereinstimmungen mit der Tag-Cloud des gesamten Archivs aufweist. Auf diese Weise könnten Begriffe automatisch extrahiert werden, die als Zuordnung von Kategorien für die Strukturierung des Archivs nutzbar wären. In einem weitern Schritt könnten Tag-Clouds der einzelnen Artikel erstellt werden, um auf Artikelbasis Vergleiche hinsichtlich der Distanz von Artikeln anstellen zu können.

Für die Arbeit mit der Tag-Cloud-Technologie werden im Folgenden ein paar Anmerkungen zusammengefasst, die bei einer weitergehenden Auseinandersetzung mit diesem Verfahren Berücksichtigung finden sollten:

- Aus fachlicher Sicht sind der Arbeit mit der eingesetzten Tag-Cloud-Methode qualitative Grenzen gesetzt, da die alleinige Auswertung von Worthäufigkeiten die Charakteristika kleiner Gruppen bzw. selten verwendeter Wörter nicht berücksichtigt. Eine Lösung könnte sein, ergänzend mit TF-IDF zu arbeiten.
- Die dargestellte Untersuchung basiert auf der Ermittlung der absoluten Worthäufigkeiten. Da die Länge der Eurozine-Artikel, wie bereits erwähnt wurde, jedoch sehr stark

variiert, müsste eine weitergehende Arbeit mit diesem Verfahren die Häufigkeit von Wörtern relativ zur Textlänge berechnen. Dies würde eine bessere Vergleichsgrundlage schaffen. Bisher ist davon auszugehen, dass lange Artikel im Verhältnis zu kurzen (davon ausgehend, dass in langen Artikeln Wörter häufiger wiederholt werden als in kurzen) überdurchschnittlich hoch bewertet werden.

- Wünschenswert wäre zudem, wenn eine Zuordnung der visualisierten Wörter zu den entsprechenden Artikeln möglich wäre – nicht zuletzt, um zu sehen, in welchem Verhältnis die Wörter aus der Tag-Cloud in den einzelnen Artikeln zu finden sind.

- Des Weiteren wäre es für die Erschließung und ein perspektivisches Vorschlagsystem für die Erstellung von Dossiers von Interesse, individuell bestimmen zu können, welche Teile der Dateien untersucht werden. Auf diese Weise könnte es möglich sein, nur die Titel oder Abstracts zu analysieren und hinsichltich ihrer Distanzen miteinander zu vergleichen und zu clustern.

- Auch wäre eine Erweiterung um eine Drag-and-Drop-Funktion, mithilfe derer einzelne Tags und die zugehörigen Artikel isoliert werden könnten, ggfs. eine sinnvolle Unterstützung in Richtung einer teilautomatisierten bzw. eines Vorschlagsystems für die Erstellung von Dossiers.

- Die beiden erstellten Tag-Clouds zeigen, dass zu den am häufigsten verwendeten Wörtern des Archivs auch Wörter gehören, deren Aussagekraft für eine inhaltliche Erschließung des Archivs unbedeutend ist. Dazu gehören Wörter wie „articl“ und „author“ ebenso wie Begriffe wie „problem“ oder „term“. Für eine Erweiterung des Programms sollte die Liste der Stopwords um solche Wörter ergänzt werden, damit gewährleistet wird, dass ausschließlich Wörter, die einer Diskrimination dienlich sind, visualisiert werden. Welche Wörter jedoch inhaltlich irrelevant sind, muss von redaktioneller Seite entschieden werden.

- Ein letzter an dieser Stelle anzumerkender Aspekt betrifft die Reduzierung des Feature-Vektoren: Das Ziel beim Einsatz der Tag-Cloud ist, Artikel miteinander vergleichen zu können. Um eine übersichtliche Vergleichsbasis zu erlangen, könnte es ratsam sein, den Feature-Vektoren zu reduzieren. So müsste das Programm in der Lage sein, „United States of America“, „U.S.“ und „America“ als ein Feature zu erkennen, statt drei verschiedene Wörter zu zählen.

5.3 Analyse des Artikelarchivs mithilfe von RapidMiner

Rapidminer [2015] ist ein von der TU Dortmund entwickeltes Open-Source-Framework für Textmining. In einer Art „Baukastensystem“ können mehr als 100 unterschiedliche Funktionalitäten (Operatoren) per Drag and Drop in einer übersichtlich gestalteten Nutzerumgebung miteinander kombiniert und auf diese Weise zu einem komplex arbeitenden Mining-Prozess „zusammengesteckt“ werden. Individuelle Mining-Abläufe werden so konstruierbar, und Dokumentensammlungen können entsprechend ihrer spezifischen Anforderungen analysiert werden. Die zur Verfügung stehenden Operatoren lassen sich durch Hinzufügen von eigenem Code ergänzen. Land u. Fischer [2012] fassen die Möglichkeiten von RapidMiner folgendermaßen zusammen:

> [RapidMiner] provides a wide range of methods from simple statistical evaluations such as correlation analysis to regression, classification and clustering procedures as well as dimension reduction and parameter optimization. [...] All these analyses can be fully automated and their results visualised in various ways. (S. V)

Schöneberg [2015a] hat in seiner Arbeit mithilfe von RapidMiner den Artikel-Korpus von Eurozine hinsichtlich der Möglichkeiten, teilautomatisierte Dossiers zu erstellen, untersucht und dazu unterschiedliche Versuche für Textmining-Verfahren aufgesetzt.[18] Herausgefunden werden sollte, ob die Zusammenstellung eines bestehenden Focal Points automatisiert reproduziert werden kann. Entsprechend den bereits vorgestellten Ansätzen basierte auch bei Schöneberg die Analyse auf Wordcounting bzw. auf einem Bag-of-Words-Ansatz. Die Analyse sollte zeigen, wodurch sich die Distanz der Artikel eines Focal Points (in diesem Fall von „The ends of democracy“ als einem durch die Eurozine-Redaktion sinnhaft zusammengestellten Dossier) im Verhältnis zu zufällig ausgewählten Artikeln bemisst.

Ausgangspunkt der Untersuchung war in diesem Fall ein Artikel aus dem Focal Point, der als Leitartikel gesetzt wurde.[19] Die Berechnung der Distanzen der Artikel des Focal Points sah nicht vor, jeden Artikel mit jedem zu vergleichen, sondern jeden Artikel hinsichtlich seiner Distanz zum Leitartikel. Das zu untersuchende Datenset beinhaltete 58 Artikel – die (verblie-

[18] Die vorliegende Arbeit gibt nur eine kurze Zusammenfassung der wesentlichen Erkenntnisse der Arbeit von Schöneberg wieder. Die vollständige Arbeit kann hier eingesehen werden: https://users.informatik.haw-hamburg.de/ ubicomp/arbeiten/master/schoeneberg.pdf

[19] Als Leitartikel wurde der erste auf der Website gelistete Artikel des Focal Points ausgewählt. Die Annahme bei diesem Vorgehen war, dass Redaktionen häufig vor der Aufgabe stehen, einem gerade erschienenen Artikel ähnliche Artikel im Sinne von related content aus dem Archiv zuzuordnen und auf diese Weise Dossiers zu erstellen.

benen) 28 Artikel des Focal Points „The ends of democracy“ und 28 weitere, zufällig ausgewählte Artikel.

Auch diese Analyse beinhaltete, wie das Programm zur Erzeugung einer Tag-Cloud, eine Datenvorbereitung (entsprechend dem klassischen KDD-Prozess), zu der u.a. das Löschen der Stopwords, die Umwandlung des Datensets in vollständige Kleinschreibung, Stemming und ggfs. die Featureauswahl (nur Substantive, nur häufigste Wörter etc.) gehören.

Die Distanz der ausgewählten 58 Artikel zum Leitartikel wurde zunächst mit euklidischer und im Anschluss daran mit Kosinusdistanz berechnet (Schöneberg [2015b], S. 20ff). Letztere ermöglichte, die stark variierende Länge der unterschiedlichen Artikel zu berücksichtigen und relative Worthäufigkeiten (Term Frequencies) zu ermitteln, die eine validere Grundlage für den Vergleich der Artikel darstellen.

In einem weiteren Teil der Untersuchung nahm Schöneberg, der Annahme folgend, dass Titel und Abstract (und Zwischenüberschriften) signifikante Abschnitte eines Artikels darstellen und zentrale Wörter enthalten (vgl. S. 43), eine Gewichtung zur Ermittung der Distanzen vor: Wörter, die in den als signifikant deklarierten Teilen der Artikel vorkamen, wurden stärker gewichtet als die Wörter, die im Body des Artikels vorkamen. Erwartet wurde durch diese Herangehensweise das Erzielen eines besseren Ergebnisses in Form einer höheren Übereinstimmung der Treffer mit den Artikeln des bestehenden Focal Points.

Schöneberg kam in seiner Arbeit zu folgenden Ergebnissen (vgl. S. 58ff.): Eine Gewichtung von als siginifikant erachteten Teilen der Artikel verbessert die Qualität der Treffer. Der Focal Point konnte durch die Berechnung gewichteter Distanzen auf der Basis von Wordcounting automatisch nahezu fehlerfrei reproduziert werden.

Der Versuch, das Archiv mit Unterstützung von WordNet [2015] zu erschließen, hat wider Erwarten keine verwertbaren Ergebnisse hervorgebracht. Im Gegenteil hat das Einbinden des Lexikons dazu geführt, dass Verbindungen zwischen Artikeln ermittelt wurden, die aus fachlicher Sicht falsch sind bzw. nicht existieren. Die Überlegung, Lexika einzubinden oder auf eine speziell auf die Themen des Eurozine-Archivs ausgerichtete Ontologie zu referenzieren, bleibt jedoch auch Schöneberg zufolge eine sinnvolle Option (S. 59).

Die zentrale Erkenntnis ist auch in der Arbeit von Schöneberg, ähnlich den bereits getätigten Anmerkungen dieser Arbeit, dass der Feature-Raum eines Dokuments reduziert werden muss, um eine sinnvolle Vergleichsbasis zu schaffen – jedoch ohne dabei die wichtigen „Kerninformationen zu verlieren“ (S. 59).

5.4 Inhaltliche Erschließung des Eurozine-Archivs mit Overview

Im folgenden Abschnitt wird mit Overview [2014] ein Open-Source-Verfahren für die Erschließung des Eurozine-Korpus eingesetzt, mit dem große Datenmengen in Echtzeit analysiert und in Form hierarchischer Cluster visualisiert werden. Es ist ein Verfahren, das speziell für journalistische Bedarfe erstellt worden ist.[20] Das Clustering-Tool bietet für die vorliegende Arbeit, die untersucht, welche Verfahren für eine teilautomtisierte Dossiererstellung nützlich sind, eine gute Ergänzung zu den bisher vorgestellten Ansätzen. Overview ermöglicht die Zuordnung von (Such-)Wörtern zu einzelnen Artikeln – und behebt damit das Manko, das z. B. beim Einsatz des Tag-Cloud-Programms formuliert worden war.

Zunächst wird das Verfahren und seine Funktionsweise vorgestellt. Im Anschluss daran wird die Erschließung des Eurozine-Archivs mit Overview beschrieben und bewertet.

5.4.1 Die Funktionsweise von Overview

Overview läuft im Browser. Die zu analysierenden Daten werden als Datenset hochgeladen (z. B. als csv-Datei), analysiert und in Form von hierarchischen Ordnern visualisiert. Das Ergebnis sind Cluster der Artikel, aus denen ersichtlich wird, welche Wörter im Datenset – und in den einzelnen Clustern – häufig vorkommen. Overview ermöglicht,

- Dokumentensammlungen sehr schnell und übersichtlich anhand häufig vorkommender Wörter zu clustern,
- die Ergebnisse in unterschiedlichen Arten zu visualisieren und
- Dokumentengruppen mit Tags zu versehen.

Diese Funktionalitäten schaffen für JournalistInnen die Möglichkeit, aktiv mit dem Datenset zu arbeiten. Brehmer u. a. [2014] merken an, dass standardisierte Volltextsuchen diesen Anforderungen oft nicht genügen würden: Selbst wenn klar sei, nach was gesucht würde, bestünde die Schwierigkeit bei herkömmlichen Suchen oft darin, eine eindeutige Suchanfrage zu formulieren.

Brehmer u. a. [2014] weisen in ihrer Arbeit darauf hin, dass die Entwicklung von Overview in verschiedenen Versionen erfolgt ist, in deren Verlauf (und Evaluation) sich die Notwendigkeit herausstellte, sich in der Funktionalität des Programms auf wenige Aufgaben zu be-

[20] Bekannt geworden ist Overview durch den Einsatz bei der Analyse der sogenannten Snowden-Files.

schränken. Die zentralen Aspekte des Programms, die nachfolgend kurz beschrieben werden, sind die Topic Tree Visualization, die Document List und der Document Viewer.

Die Topic Tree Visualization bezeichnet die Darstellung gefundener „Topics“, gewissermaßen wiederkehrenden Wortkombinationen, in Form einer Ordnerstruktur. Die Ordner enthalten Annotationen zu Wörtern, die häufig, manchmal oder in allen Artikeln eines Ordners vorkommen („most“, „some“ und „all“).

Sehr anschaulich erklärt Brett [2012], was genau Topic Modeling ausmacht und worin es dem reinen Wordcounting überlegen ist:

> As you read through the article, you use a different color for the key words of themes within the paper as you come across them. When you were done, you could copy out the words as grouped by the color you assigned them. That list of words is a topic, and each color represents a different topic.

Topic Models dienen also der Erhebung und Darstellung wiederkehrender Muster von Wörtern (vgl. dazu auch: Blei u. Lafferty [2009]).

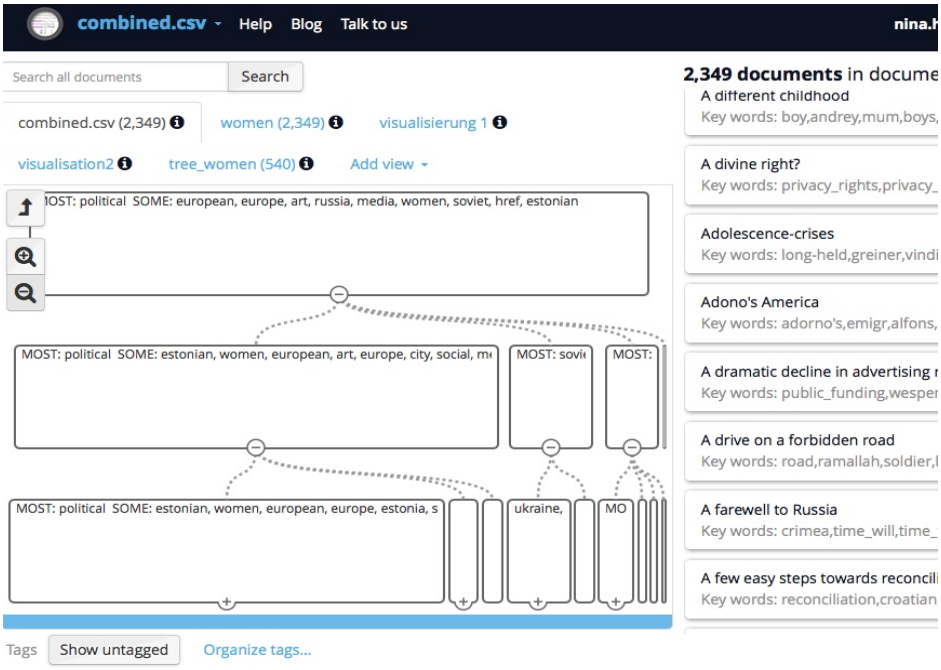

Abbildung 5.6: Overview – links die hierarchische Ordnerstruktur, rechts daneben die Liste der Artikel

Während die Topic Tree Visualization den größten Teil des Bildschirms einnimmt, sind die Document List nud der Document Viewer auf der rechten Seite des Bildschirms verortet. Anders als die Ordnerstruktur geben sie keinen Überblick über das gesamte Datenset, sondern einen genaueren Einblick in ausgewählte Dokumente. So listet die Document List die Titel und Keywords von vorab ausgewählten Dokumenten und ermöglicht das Hineinzoomen in einzelne Dokumente durch den Document Viewer, mit dem vollständige Dokumente angezeigt und gelesen werden können.

Auch bei Overview ist die Datenvorbereitung ein Teil des Programms: Stopwords werden herausgefiltert, zudem werden POS-Tagging und Wordcounting durchgeführt. Grundlage der Analyse ist eine Volltextsuche, mittels der die Worthäufigkeiten in den einzelnen Dateien des Datensets ermittelt werden. Inwieweit weitere Regeln als die Häufigkeit die Extraktion der später visualisierten Topics beeinflusst, weiß man als NutzerIn von Overview nicht.

Dateien, die ähnliche Worthäufigkeiten aufweisen, werden geclustert dargestellt. Die Ordner tragen Annotationen in Form von Labeln, die darüber Auskunft geben, welche Wörter in den Artikeln, die in einem Ordner gebündelt wurden, häufig, manchmal oder in allen im Ordner befindlichen Dokumenten vorkommen. Die Label sind: „Most“, „Some“ und „All“. Die Ordner können vom User weiter unterteilt werden, sodass sich eine differenzierte Baumstruktur entfaltet.

Es ist möglich, Begriffe zu taggen und farblich zu markieren, sodass auch in der Dokumentensammlung erkennbar bleibt, welche Begriffe sich wo wiederfinden. Es gibt zudem unterschiedliche Möglichkeiten, innerhalb der Dokumente zu suchen – unter anderem über eine Volltextsuche – und die Ergebnisse farblich zu markieren.

5.4.2 Analyse des Eurozine-Archivs mit Overview

Für die Analyse des Eurozine-Korpus mit Overview stand ein Datensatz aus 2.349 Artikeln zur Verfügung, der zu einer CSV-Datei konvertiert und ins Programm geladen wurde.[21] Die Analyse und Tree Visualization der Eurozine-Daten gibt als häufig vorkommendes Keyword in den größten Clustern „political“ aus (Abb. 5.6). Weitere häufig verwendete Wörter in den drei größten Clustern sind u.a. „european“, „europe“, „art“, „women“ und „estonian“. Betrachtet man die kleineren Ordner, differenzieren sich die Themen aus. So ist z. B. ein Cluster mit Artikeln bestückt, in denen ein häufig vorkommendes Keyword „soviet“ ist.

[21] Der Datensatz war kleiner als der für die Tag-Cloud genutzte Datensatz aus 2.697 Artikeln. Der Grund liegt darin, dass bei der Konvertierung der XML-Daten in eine CSV-Datei knapp 300 Artikel aufgrund ihrer invaliden XML-Struktur aus der Dokumentensammlung genommen werden mussten.

Im Vergleich zur Analyse des Eurozine-Archivs mit dem Tag-Cloud-Programm fällt auf, dass Overview „european“ und „europe“ als zwei unterschiedliche Keywords nennt, statt sie dem gleichen Wortstamm zuzuordnen. Offensichtlich wird kein Stemming vorgenommen. Anzumerken ist zudem, dass zu den häufig verwendeten Wörtern auch „href“ gehört, das in den Eurozine-Artikeln für die Einbettung von Link-Verweisen steht. Daran zeigt sich, dass, sollte man weiterhin mit Overview arbeiten wollen, eine Ergänzung der Stopwords-Liste sinnvoll wäre. Denn ähnlich der Untersuchung mit dem Tag-Cloud-Verfahren sollten Tags oder, wie in diesem Fall, Keywords inhaltlich möglichst sinnhafte Wörter sein – Wörter also, die bezogen auf das jeweilige Archiv diskriminatorische Relevanz haben.

Um auch bei diesem Verfahren Aussagen über die Erschließungsmöglichkeit des Archivs ausgehend vom Focal Point „The ends of democracy“ treffen zu können, wurde in einem nächsten Schritt nach dem Wortteil „democr*“ im gesamten Datenset gesucht. Overview zeigte in den entsprechenden Clustern insgesamt 855 Dokumente, die den Wortteil enthalten (Abb. 5.7).

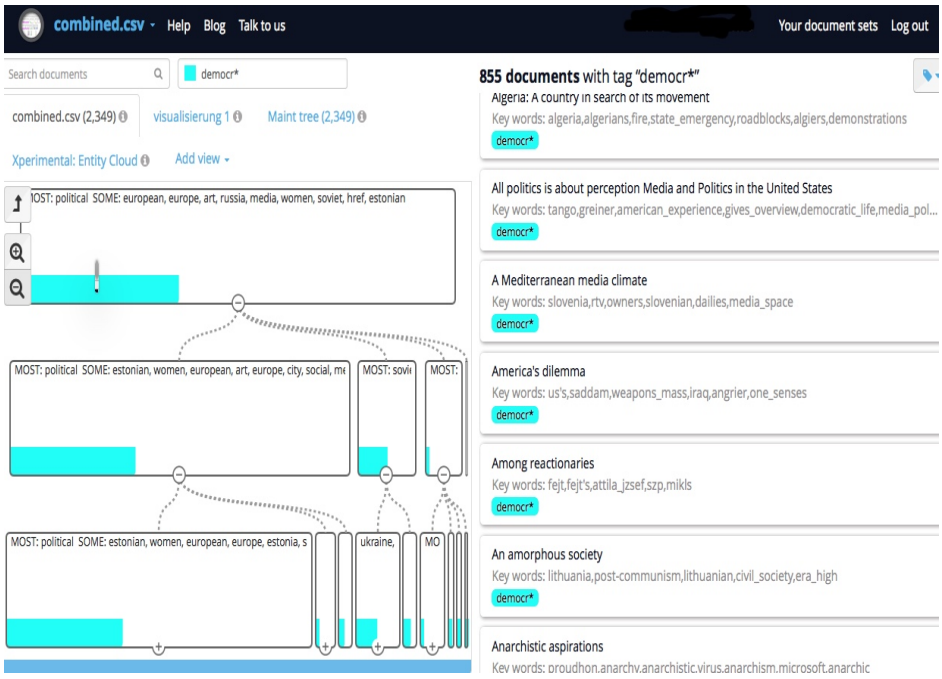

Abbildung 5.7: Overview – Auswahl der Artikel, die „democr*“ enthalten

Das zusätzliche Taggen des Suchwortes ermöglichte die Visualisierung des Suchworts in den verschiedenen Clustern in Form eines Balken, der die Menge abbilden sollte (Abb. 5.7). Wenngleich die Treffer dieser Untersuchung zahlenmäßig nicht mit den Ergebnissen der Tag-Cloud verglichen werden können, kann festgehalten werden, dass die Methode, Wörter farbig

sichtbar zu machen (durch die Zuweisung eines farbigen Tags) und die gleichzeitige Sortierung der Artikel in thematisch sortierte Cluster ein gangbarer Weg zu sein scheint, um ein Archiv zu erschließen bzw. einen Überblick über die Themen und ihre Verteilung zu bekommen.

Ein weiterer Schritt bestand darin, Artikel zu suchen, die die nächsthäufigsten Wörter des Focal Points „The ends of democracy“ enthalten (bezogen auf die Ergebnisse der Tag-Cloud des Focal Points). 1.290 Artikel des Datensets enthielten das Wort „govern*“, 1.445 Artikel „societ*“ und 1.470 Artikel „state“. Das Vorkommen aller vier Begriffe in jeweils mehr als der Hälfte der dem Datenset zugehörigen Artikel zeigt, dass die verwendeten Wörter kaum diskriminatorische Relevanz haben. Anders ausgedrückt: Es ist kaum möglich, einen Artikel über einen einzelnen dieser Begriffe zu charakterisieren, da es sich um Merkmale handelt, die auf sehr viele Artikel des Datensets zutreffen. Anders sieht es jedoch aus, wenn man eine Suchanfrage nach dem gleichzeitigen Vorkommen aller vier Begriffe in Dokumenten stellt. Die Anfrage ergibt, dass 655 Artikel – also gut ein Viertel der Artikel des Datensets (27,88 Prozent) alle vier Begriffe enthalten. Die Kombination von Suchbegriffen bzw. Keywords hat demnach eine weitaus bessere diskriminatorische Relevanz als die Suche nach einzelnen Begriffen.

Zum Einsatz von Overview als Erschließungsmethode für ein Artikelarchiv wie das von Eurozine kann zusammenfassend festgehalten werden:

- Das Clustern der Artikel eines Datensets ermöglicht einen schnellen Eindruck über die Themen eines Archivs. Die Anzeige der Keywords – und eine grobe Angabe über die Häufigkeit ihres Vorkommens – erleichtert es zudem, thematische Cluster zu erfassen. Dies könnte eine zweckdienliche Methode sein, um eine Debatte zu einem Thema oder auch Facetten eines Themas teilautomatisiert als Dossiers bündeln zu können.

- Zweckdienlicher als bei dem Tag-Cloud-Programm ist auch die Zuordnung von Keywords zu den einzelnen Artikeln. Auf diese Weise wird es möglich, nicht nur die Häufigkeit des Vorkommens von Wörtern im gesamten Archiv zu ermitteln, sondern ihr Vorkommen in einer bestimmten Anzahl an Dokumenten zu erkennen.

- Auch bietet die hierarchische Struktur der Ordner für eine prospektive Dossiererstellung Vorteile: Da es möglich ist, sich innerhalb der Ordnerstruktur zu immer kleineren Ordnern fortzubewegen, landet man letztlich bei Ordnern, die eine überschaubare Anzahl an Artikeln beinhalten. Ob diese Artikelzusammenstellung sich als Vorschlagsgrundlage für Dosiers eignen könnte, die redaktionell weiterbearbeitet werden, muss Thema einer anderen Arbeit sein.

- Gleichwohl muss festgestellt werden, dass die Arbeitsweise von Overview für ein dynamisches Archiv, wie es das Eurozine-Archiv ist, deutlich weniger praktikabel ist als für einen statischen Korpus. Da das Archiv regelmäßig um neue Artikel erweitert wird, müssten die neuen Artikel jedes Mal in die für die Untersuchung erstellte CSV-Datei integriert werden.

5.5 Zusammenfassung des Anwendungsfalls

Die im vorangegangenen Kapitel eingesetzten Verfahren haben gezeigt, dass das Ermitteln von Worthäufigkeiten eine mögliche Methode zur Erschließung des Eurozine-Archivs darstellt. Gezeigt werden konnte, dass die Signifikanz von Artikeln sowohl über die Analyse vollständiger Artikel als auch über die isolierte Analyse von Textabschnitten festgestellt werden kann. Auf diese Weise lassen sich Distanzen zwischen Artikeln ermitteln – eine Voraussetzung für die darauf aufbauende Erstellung von Dossiers.[22]

Präziser als der Vergleich von Texten auf der Basis eines einzigen Wortes ist – wie bereits vermutet worden war – der Vergleich von Dateien durch die Ermittlung aller Worthäufigkeiten in den zu analysierenden Dateien – unter Berücksichtigung eines n-dimensionalen Feature-Vektors. Zur qualifizierteren Bewertung dieser Ergebnisse schlug Marcel Schöneberg in seiner Arbeit daraufhin eine Reduzierung des Feature-Raumes vor – um so der Unübersichtlichkeit eines n-dimensionalen Feature-Vektors zu begegnen. Dies ebnet den Weg hin zur Überlegung, dass in einem nächsten Schritt, der allerdings jenseits dieser Arbeit erfolgen muss, die Ermittlung von für das Eurozine-Archiv inhaltlich wichtigen Wörtern erfolgen sollte: Der Aufbau eines Lexikons bzw. die Entwicklung einer Ontologie, die die Einteilung der im Archiv vorkommenden Wörter in wichtig/unwichtig und hinsichtlich einer Zuordnung zu Kategorien bzw. Rubriken ermöglichen würde, wäre für eine inhaltsbezogenere, speziellere Analyse des Archivs vielversprechend und für die Erstellung teilautomatisierter Dossiers nutzbar.

Was man bei einem solchen Vorgehen jedoch noch nicht zu greifen bekommt, ist die teilautomatisierte Erfassung eines breiten Spektrums an Postionen zu einem Thema. Die bisher genutzten Distanzfunktionen, die auf der Ermittlung identischer oder parallel vorkommender Wörter Nähe und Ähnlichkeit von Artikeln berechnen lassen, sind insofern nur ein erster Schritt hin zu einer teilautomatisierten Erstellung von Dossiers, die speziellere Anliegen

[22] Aufgrund der bereits erwähnten stark unterschiedlichen Länge der Dateien sollte darauf geachtet werden, dass für einen Vergleich die relativen Worthäufigkeiten zur Gesamtlänge der Dateien ermittelt werden – nicht die absoluten.

verfolgen. Wie sich Artikel hinsichtlich einer für eine solche Arbeit zu ermittelnde Distanz charakterisieren lassen könnten, übersteigt den Rahmen dieser Arbeit.

Ein letzter Aspekt, der in diesem Zusammenhang genannt werden soll, ist, dass die bisherige Arbeit sich vor allem mit der Reproduktion bereits erschienender Focal Points auseinandergesetzt hat. Für eine Voruntersuchung wie die vorliegende Arbeit war das ein sinnvolles Vorgehen. Da das Ziel allerdings die Erstellung neuer Focal Points aus dem vorhandenen Archiv ist, muss auf der Basis der getätigten Versuche untersucht werden, ob auf diese Weise auch vollständig neue Dossiers erstellt werden können. Die Qualität von auf diese Weise erstellten Dossiers sollte dann von RedakteurInnen geprüft werden. Selbige könnten, im Fall des Gelingens, aus den automatisiert zusammengestellten Dossiervorschlägen nach einer inhaltlichen Prüfung die endgültigen neuen Dossiers zusammenstellen.

6 Zusammenfassung und Ausblick

Die zentrale Frage, die sich heutige und zukünftige Redaktionen von Zeitungen und Zeitschriften stellen, formulierte der Eurozine-Mitgründer Tjark L. im Interview folgendermaßen: „Wie können wir das ganze fantastische Material auf die eine oder andere Art sichtbar machen?“ (Tjark L., Teil 3, 6:35) Mit „das ganze fantastische Material“ ist das Eurozine-Archiv bzw. die darin enthaltenen Artikel gemeint, die aufgrund ihrer mittlerweile unübersichtlichen Menge nicht mehr mit analogen Methoden erschlossen werden können. Die Antwort auf seine Frage muss – das zeigt diese Arbeit – lauten: durch den Einsatz von Textmining-Methoden zugunsten einer teilautomatisierten Zusammenstellung von Dossiers. Ob die Dossiers dann eine Recherchegrundlage für die Arbeit an neuen Artikeln bieten, der Dokumentation der eigenen Arbeit dienen oder als neues Produkt an LeserInnen bzw. an kooperierende Zeitungen weitergegeben werden – immer sind sie der Weg, entlang bestimmter Kriterien aus vorhandenen Daten neues Wissen zu generieren.

Natürlich gibt es schon jetzt Lösungen dafür, und sie werden auch genutzt: ArchivarInnen, BibilothekarInnen und RedakteurInnen verschlagworten und vergeben Signaturen, sichten, wählen aus und stellen Material neu zusammen. Allerdings gehen dabei einerseits viele möglicherweise wertvolle Informationen in den Untiefen der Archive verloren, andererseits sind diese Prozesse langwierig und voraussetzungsvoll. Die Herstellung einer Wiederauffindbarkeit von Informationen ist eine Wissenschaft für sich – die der Medienarchivare.

Davon ausgehend, dass Textmining ein wertvolles Werkzeug für die zukünftige Erstellung von Dossiers sein könnte, untersuchte diese Arbeit die Voraussetzungen für ein Gelingen dieses Prozesses. Dafür wurde zunächst der Frage nachgegangen, was genau ein Dossier ausmacht. Es stellte sich heraus, dass es unterschiedliche Vorstellungen von und Praxen mit Dossiers gibt. Abhängig vom beruflichen Kontext der ExpertInnen wurden sowohl verschiedene Erstellungs- als auch Verwendungsarten von Dossiers genannt. Es wurde klar, wie wichtig es ist, zur Erstellung eines Dossiers folgende grundlegende – strukturelle und kontextbezogene – Fragen zu beantworten:

- Um was für ein Archiv handelt es sich? Woraus besteht es?

- Wie sieht die Archiv-, also die Datenbankstruktur, und wie sieht die Artikelstruktur aus? Auf welche der vorhandenen Strukturen aufbauend können Textminingverfahren ansetzen?
- Für welchen Zweck und mit welchem Umfang sollen Dossiers erstellt werden?
- Was ist die Fragestellung, aus der sich der Inhalt der Dossiers generieren soll?
- Wer ist die prospektive LeserInnenschaft?

Die Untersuchung der und das Wissen um die genannten Aspekte ist auch für die Erstellung teilautomatisierter Dossiers eine wichtige Voraussetzung, denn abhängig von den verschiedenen Faktoren kann das Set an Methoden stark variieren.

Ein großes Problem des Eurozine-Archivs ist – neben seiner Mehrsprachigkeit und dem invaliden XML – die schiere Masse an Text: Die durchschnittliche Länge eines Artikels liegt bei 3.500 Wörtern. Dazu kommt das Problem der Artikelüberschriften: Zwar sind weniger Buzzwords enthalten, als es mittlerweile oft bei Tageszeitungen der Fall ist, allerdings muss festgestellt werden, dass gerade bei Eurozine die Überschriften der Artikel nicht immer Aufschluss über ihren Inhalt geben. Wenngleich beim testweisen Auszählen der häufigsten Wörter der Überschriften eines Focal Points die vorliegende Untersuchung zwar zu dem Ergebnis kam, dass der Focal Point reproduzierbar wäre, lässt sich auf diese Weise nur schwerlich ein ganzes Archiv erschließen. Buzzwords, wie im beschriebenen Fall „democracy“, verführen zu falschen Schlüssen und laden dazu ein, die Facetten eines Themas zu ignorieren. Für eine differenziertere Erstellung teilautomatisierter Dossiers ist es aus diesem Grund wichtig, den Bag-of-Words-Ansatz zu erweitern.

Die vorliegende Arbeit hat gezeigt, dass Mining-Verfahren jedweder Art die Strukturen der Trägersysteme von den Inhalten unterscheiden können müssen. Ein Beispiel dafür, was geschieht, wenn diese Vorbedingung nicht erfüllt ist, konnte in dieser Arbeit gesehen werden: Overview analysierte das HTML-Tag „href“ nicht als Strukturelement, sondern als ein häufig in den Artikeln vorkommendes Inhaltselement. Im vorliegenden Fall wurden entsprechend Artikel anhand des Vorkommens des HTML-Tags geclustert, woraus sich ein Cluster ohne inhaltlich-thematische Relevanz ergab. Eine einfache Lösung für einen solchen Fall (vorausgesetzt, der Fehler wird bemerkt) könnte darin bestehen, die Stopword-Liste um das HTML-Tag zu erweitern und es auf diese Weise aus der inhaltlichen Erschließung des Archivs auszuklammern. Für jedes Archiv, das mit Mining-Verfahren erschlossen werden soll, müssen Faktoren wie diese gesondert untersucht werden.

Zusammengefasst werden kann, dass angesichts der in dieser Arbeit gewonnenen Erkenntnisse die weitergehende Arbeit zu Vorschlagsystemen bzw. Empfehlungsstrukturen vielversprechend erscheint. Die in der vorliegenden Arbeit verfolgten Ansätze sind im Kontext mit anderen Arbeiten zu betrachten, die derzeit an der HAW verfolgt werden und als Co-Experimente den Einsatz von Algorithmen aus utilitaristischer Sicht untersuchen. Außer der Arbeit von Marcel Schöneberg, der feststellen konnte, dass die Analyse der Abstracts qualitativ hochwertigere Ergebnisse mit sich bringt als die Analyse der Überschriften eines Focal Points, gehören dazu auch z.B. die Vorarbeiten von Jan Paul Assendorp. Assendorp untersucht, ebenfalls auf der Basis des Eurozine-Archivs, Möglichkeiten zur semi-automatisierten Erstellung journalistischer Dossiers: „Anhand eines Klassifikators in Form eines künstlichen neuronalen Netzes soll zu vorgegebenen Focal Points eine Menge an geeigneten Dokumenten identifiziert und dem Anwender vorgeschlagen werden.“ (S. 1) Assendorp konzentriert sich dabei auf die Arbeit mit neuronalen Netzen. Er erwartet, auf diese Weise bessere Ergebnisse zu erzielen, als durch den Einsatz von Klassifikationsmethoden (vgl. Assendorp, S. 8).

Die Untersuchung hat außerdem gezeigt, dass für eine inhaltliche Erschließung des Eurozine-Archivs sowohl Potenzial in der Analyse von Teilabschnitten als auch bei der Ermittlung von Wortkombinationen liegt: Im Anwendungsfall konnte dargestellt werden, dass die isolierte Analyse von Teilabschnitten eines Artikels ebenso wie die Gewichtung von als relevant erachteten Abschnitten eines Artikels (in diesem Fall der Abstracts) gute Ansätze für eine Erschließung des Archivs bietet, die an die Qualität der redaktionell erstellten Focal Points herankommen könnte. Ob es sich tatsächlich um valide Ergebnisse handelt, müsste in weitergehenden Untersuchungen – anhand weiterer Focal Points – überprüft werden. Nachgedacht werden sollte, ausgehend von Wortkombinationen, über komplexere Verfahren, die die Distanz von bzw. Zusammenhänge zwischen Artikeln genauer ermitteln können. Ein zweckdienlicher nächster Schritt wäre aus diesem Grund die Erarbeitung einer archivspezifischen Ontologie (mit kultur- und geisteswissenschaftlichen Begriffen), um diese zur Erschließung und Strukturierung des Korpus und zur Erstellung von Dossiervorschlägen einzusetzen.

7 Anhang

7.1 Gesprächsleitfaden Interviewpartner

Leitfaden „Dossiers“

- Worüber müssen wir Ihrer Meinung nach sprechen, wenn wir über Dossiers sprechen wollen?
- Was macht ein gutes Dossier aus?
- Können Sie mir in ein paar Sätzen sagen, wie Sie vorgehen, wenn Sie ein Dossier erstellen?
- Welche drei Kriterien sind für die Erstellung eines guten Dossiers am wichtigsten?
- Wie unterscheidet sich ein gutes Dossier von einem schlechten?
- Wofür werden Dossiers genutzt?
- Für welche Zwecke erstellen Sie in Ihrem beruflichen Kontext Dossiers?
- Wer liest Ihre Dossiers?
- Welches Feedback bekommen Sie?
- Gibt es noch etwas, das wir noch nicht besprochen haben, das aber beim Thema Dossiers unbedingt erwähnt werden sollte?

7.2 Programmaufruf Tagcloud

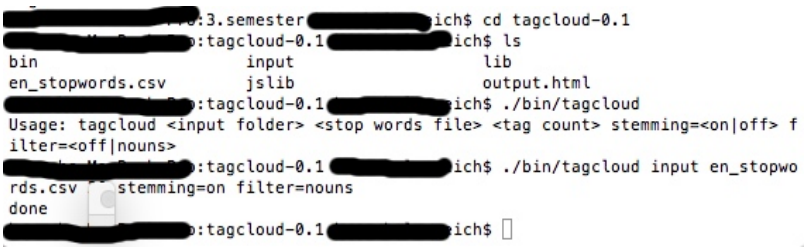

Abbildung 7.1: Der Programmaufruf Tag-Cloud über die Konsole

7.3 Die Liste der Artikel des Focal Points „The ends of democracy“

Die urls der Artikel setzen sich aus der Adresse http://www.eurozine.com/articles/ und jeweils einem der folgenden Dateinamen zusammen.

- 2001-11-27-rosenberg-en
- 2008-05-02-wennerhag-en
- 2008-11-21-leggewiewelzer-en
- 2009-04-21-fraser-en
- 2009-07-14-biscione-en
- 2009-09-09-kavaliauskas-en
- 2010-09-14-ditchev-en
- 2011-07-11-bluhdorn-en
- 2011-11-02-G1000-en
- 2011-11-10-sierakowski-en
- 2011-12-19-amirpur-en
- 2012-01-25-halmai-en

- 2012-02-08-elsenhans-en
- 2012-09-05-jahanbegloo-en
- 2012-11-21-holmes-en
- 2013-02-01-krastev-en
- 2013-02-08-wallerstein-en
- 2013-02-19-leggewie-en
- 2013-02-26-james-en
- 2013-05-03-muller-en
- 2013-06-14-pomerantsev-en
- 2013-07-29-gole-en
- 2013-08-13-krastev-en
- 2013-08-20-leggewienanz-en
- 2013-09-11-deniztekin-en
- 2013-11-08-vidanava-en
- 2013-11-22-offe-en
- 2013-12-12-margetts-en
- 2013-12-12-pogonyi-en

7.4 Die Liste der Artikel der Gegenstichprobe

Die urls der Artikel setzen sich aus der Adresse http://www.eurozine.com/articles/ und jeweils einem der folgenden Dateinamen zusammen.

- 2000-08-28-gellner-en
- 2003-03-26-santos-en
- 2006-11-26-sluga-en

- 2009-01-20-ferrard-en
- 2012-05-02-sturmmartin-en

Literaturverzeichnis

[Agarwal u. Liu 2008] Agarwal, Nitin ; Liu, Huan: Blogosphere: Research Issues, Tools, and Applications. In: *SIGKDD Explor. Newsl.* 10 (2008), Mai, Nr. 1, 18–31. http://dx.doi.org/10.1145/1412734.1412737. – DOI 10.1145/1412734.1412737. – ISSN 1931–0145

[Aggarwal 2015] Aggarwal, Charu C.: *Data Mining. The Textbook.* Springer, 2015

[Anderson u. a. 2012] Anderson, C.W. ; Bell, Emily ; Shirky, Clay: *Post-Industrial Journalism. Adapting to the Present.* http://towcenter.org/wp-content/uploads/2012/11/TOWCenter-Post_Industrial_Journalism.pdf, 2012. – [Online; accessed 15-June-2015]

[Assendorp] Assendorp, Jan P.: Digital Journalism. Automatisierte Dossier-Erstellung mittels Textmining. https://users.informatik.haw-hamburg.de/~ubicomp/projekte/master2015-proj/assendorp.pdf

[Blätter 2015] Blätter: *Dossiers.* https://www.blaetter.de/aktuell/dossiers, 2015. – [Online; accessed 7-July-2015]

[Blei u. Lafferty 2009] Blei, David M. ; Lafferty, John D.: *Topic Models.* http://www.cs.princeton.edu/~blei/papers/BleiLafferty2009.pdf. Version: 2009. – [Online; accessed 12-November-2015]

[Boersch] Boersch, Ingo: Data Mining., Vortrag, gehalten am 21.5.2015 bei der Ringvorlesung Next Media @ HAW Hamburg

[Bogner u. a. 2005] Bogner, Alexander ; Littig, Beate ; Menz, Wolfgang: *Das Experteninterview.* 2. durchgesehene Aufl. 2005. Opladen : VS Verlag für Sozialwissenschaften, 2005

[Brehmer u. a. 2014] Brehmer, M. ; Ingram, S. ; Stray, J. ; Munzner, T.: Overview: The design, adoption, and analysis of a visual document mining tool for investigative journalists. In: *IEEE Trans. Visualization and Computer Graphics (TVCG / Proc. InfoVis)* 20 (2014), Nr.

12, S. 2271–2280. http://dx.doi.org/10.1109/TVCG.2014.2346431. – DOI 10.1109/TVCG.2014.2346431

[Brett 2012] Brett, Megan R.: *Topic Modeling. A Basic Introduction.* http://journalofdigitalhumanities.org/2-1/topic-modeling-a-basic-introduction-by-megan-r-brett/. Version: 2012. – [Online; accessed 12-November-2015]

[Cios u. a. 2007] Cios, W. K.J. an P. K.J. an Pedrycz ; Swiniarski, R.W. ; Kurgan, L.: *Data Mining. A Knowledge Discovery Approach.* Heidelberg : Springer, 2007

[Cleve u. Lämmel 2014] Cleve, Jürgen ; Lämmel, Uwe: *Data Mining.* Oldenbourg Wissenschaftsverlag GmbH, 2014

[Duden 2013] Duden: *Dossier, das.* http://www.duden.de/rechtschreibung/Dossier, 2013. – [Online; accessed 25-October-2014]

[Ebitsch 2014] Ebitsch, Sabrina: *Dahin gehen, wo es wehtut. Dossier zum Thema Toleranz.* http://www.sueddeutsche.de/leben/dossier-zum-thema-toleranz-dahin-gehen-wo-es-wehtut-1.2207571, 2014. – [Online; accessed 26-January-2015]

[Eurozine 2014a] Eurozine: *About.* http://www.eurozine.com/about_Eurozine.html, 2014. – [Online; accessed 25-October-2014]

[Eurozine 2014b] Eurozine: *Newsletter.* http://www.eurozine.com/newsletter.html, 2014. – [Online; accessed 25-October-2014]

[Fayyad u. a. 1996] Fayyad, Usama ; Piatetsky-Shapiro, Gregory ; Smyth, Padhraic: The KDD Process for Extracting Useful Knowledge from Volumes of Data. In: *Commun. ACM* 39 (1996), November, Nr. 11, 27–34. http://dx.doi.org/10.1145/240455.240464. – DOI 10.1145/240455.240464. – ISSN 0001–0782

[Feldman u. Sanger 2007] Feldman, Ronen ; Sanger, James: *The Text Mining Handbook - Advanced Approaches in Analyzing Unstructured Data.* Cambridge : Cambridge University Press, 2007. – ISBN 978–0–521–83657–9

[Fernández u. a. 2012] Fernández, Norberto ; Arias Fisteus, Jesús ; Sánchez, Luis ; López, Gonzalo: IdentityRank. In: *Expert Syst. Appl.* 39 (2012), August, Nr. 10, 9207–9221. http://dx.doi.org/10.1016/j.eswa.2012.02.084. – DOI 10.1016/j.eswa.2012.02.084. – ISSN 0957–4174

[Gruber 2007] Gruber, Tom: *Ontology*. http://tomgruber.org/writing/ontology-definition-2007.htm, 2007. – [Online; accessed 26-November-2015]

[Hälker 2014] Hälker, Nina: Textmining für Newssites. (2014). http://users.informatik.haw-hamburg.de/~ubicomp/projekte/master-nm-2014/haelker/bericht.pdf

[Hearst 1999] Hearst, Marti A.: Untangling Text Data Mining. In: *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*. Stroudsburg, PA, USA : Association for Computational Linguistics, 1999 (ACL '99). – ISBN 1–55860–609–3, 3–10

[Hippner u. Rentzmann 2006] Hippner, Hajo ; Rentzmann, René: Text Mining. (2006). https://www.gi.de/service/informatiklexikon/detailansicht/article/text-mining.html. – [Online; accessed 12-November-2015]

[Jain u. a. 1999] Jain, A. K. ; Murty, M. N. ; Flynn, P. J.: Data Clustering: A Review. In: *ACM Comput. Surv.* 31 (1999), September, Nr. 3, 264–323. http://dx.doi.org/10.1145/331499.331504. – DOI 10.1145/331499.331504. – ISSN 0360–0300

[Kroeze u. a. 2003] Kroeze, Jan H. ; Matthee, Machdel C. ; Bothma, Theo J. D.: Differentiating Data- and Text-mining Terminology. In: *Proceedings of the 2003 Annual Research Conference of the South African Institute of Computer Scientists and Information Technologists on Enablement Through Technology*. Republic of South Africa : South African Institute for Computer Scientists and Information Technologists, 2003 (SAICSIT '03). – ISBN 1–58113–774–5, 93–101

[Land u. Fischer 2012] Land, Sebastian ; Fischer, Simon: RapidMiner 5. RapidMiner in academic use. (2012). https://rapidminer.com/wp-content/uploads/2013/10/RapidMiner_RapidMinerInAcademicUse_en.pdf. – [Online; accessed 12-November-2015]

[Lehnert 2012] Lehnert, Petra: *Die zeitgeschichtliche Entwicklung der Zeitschrift ? von der Höhlenmalerei zum Magazin*. http://www.fachzeitungen.de/zeitschriften-zeitgeschichte, 2012. – [Online; accessed 15-June-2015]

[Lewandowski 2005] Lewandowski, Dirk: Web Information Retrieval. Technologien zur Informationssuche im Internet. (2005). http://www.durchdenken.de/

lewandowski/web-ir/download/Web-IR-Buch.pdf. – [Online; accessed 12-November-2015]

[LMd 2015a] LMd, Le Monde d.: *Dossiers*. http://monde-diplomatique.de/dossiers, 2015. – [Online; accessed 7-July-2015]

[LMd 2015b] LMd, Le Monde d.: *Edition*. http://www.monde-diplomatique.de/pm/.edition/edition, 2015. – [Online; accessed 7-July-2015]

[Manning u. a. 2009] Manning, Christopher D. ; Raghavan, Prabhakar ; Schütze, Hinrich: *An Introduction to Information Retrieval*. Cambridge University Press, 2009

[Marr 2015] Marr, Bernard: *Big Data: 20 Mind-Boggling Facts Everyone Must Read*. http://www.forbes.com/sites/bernardmarr/2015/09/30/big-data-20-mind-boggling-facts-everyone-must-read/, 2015. – [Online; accessed 12-November-2015]

[Marti 2015] Marti, Michael: „Zeitungen sind zäh. Sie sterben langsam“. Interview: Besteht der Journalismus der Zukunft aus Katzenvideos? Braucht es noch Reporter? Oder reichen Programmierer? Gibt es ein Leben jenseits von Klickzahlen? Medienprofessorin Emily Bell kennt die Antworten. (2015). http://mobile2.12app.ch/articles/22368360. – [Online; accessed 12-November-2015]

[Mayring 2010] Mayring, Philipp: *Qualitative Inhaltsanalyse - Grundlagen und Techniken*. Neuausgabe, 11. vollständig überarbeitete Aufl. Langensalza : Beltz, 2010. – ISBN 978–3–407–25533–4

[Neofonie 2013] Neofonie: *Zeitmaschine*. http://labs.neofonie.de/zeitmaschine/, 2013. – [Online; accessed 12-November-2015]

[Noelle-Neumann u. a. 1994] Noelle-Neumann, Elisabeth ; Schulz, Winfried ; Wilke, Jürgen: *Fischer Lexikon Publizistik Massenkommunikation*. Fischer, 1994

[Overview 2014] Overview: *Overviewproject*. https://www.overviewproject.org/, 2014. – [Online; accessed 25-October-2014]

[Rapidminer 2015] Rapidminer: *Home*. https://rapidminer.com, 2015. – [Online; accessed 5-November-2015]

[Ritter u. a. 2012] Ritter, Alan ; Mausam ; Etzioni, Oren ; Clark, Sam: Open Domain Event Extraction from Twitter. In: *Proceedings of the 18th ACM SIGKDD International*

Conference on Knowledge Discovery and Data Mining. New York, NY, USA : ACM, 2012 (KDD '12). – ISBN 978–1–4503–1462–6, 1104–1112

[Runkler 2015] Runkler, Thomas A.: *Data Mining. Modelle und Algorithmen intelligenter Datenanalyse*. Springer, 2015

[Schek 2005] Schek, Markus: *Automatische Klassifizierung und Visualisierung im Archiv der Süddeutschen Zeitung*. http://www.gfft-akademie.eu/files/DIZ.pdf, 2005. – [Online; accessed 26-November-2015]

[Schöneberg 2015a] Schöneberg, Marcel: Automatisierte Erstellung von Pressedossiers durch Textmining. Projektbericht. (2015). http://users.informatik.haw-hamburg.de/~ubicomp/projekte/master2015-proj/schoeneberg.pdf. – [Online; accessed 12-November-2015]

[Schöneberg 2015b] Schöneberg, Marcel: *Konzepte zur semiautomatisierten Erstellung von Pressedossiers*. https://users.informatik.haw-hamburg.de/~ubicomp/arbeiten/master/schoeneberg.pdf. Version: 2015. – [Online; accessed 12-November-2015]

[Seidel] Seidel, Ludwig M.: *Text Mining als Methode zur Wissensexploration. Konzepte, Vorgehensmodelle, Anwendungsmöglichkeiten*. http://www.wi.hs-wismar.de/~cleve/vorl/projects/da/13-Master-Seidel.pdf

[Sjurts 2012] Sjurts, Insa: *Stichwort: Zeitschrift. In: Gabler Wirtschaftslexikon*. http://wirtschaftslexikon.gabler.de/Archiv/569826/zeitschrift-v2.html, 2012. – [Online; accessed 15-June-2015]

[Wikipedia 2014] Wikipedia: *Dossier*. http://de.wikipedia.org/wiki/Dossier, 2014. – [Online; accessed 25-October-2014]

[Witte u. Mülle 2006] Witte, Rene ; Mülle, Jutta: *Text Mining. Wissensgewinnung aus natürlichsprachigen Dokumenten*. 2006 (Interner Bericht 2006-5. Fakultät für Informatik, Universität Karlsruhe)

[WordNet 2015] WordNet: *Home*. https://wordnet.princeton.edu/, 2015. – [Online; accessed 5-November-2015]

Hiermit versichere ich, dass ich die vorliegende Arbeit ohne fremde Hilfe selbständig verfasst und nur die angegebenen Hilfsmittel benutzt habe.

Hamburg, 27.11.2015 Nina Hälker