Hochschule für Angewandte Wissenschaften Hamburg
Hamburg University of Applied Sciences

Masterthesis

Quirine Philipsen

Die Rolle von Daten- und Textminingverfahren für die journalistische Arbeit

Fakultät Technik und Informatik
Department Informatik

Faculty of Engineering and Computer Science
Department of Computer Science

Quirine Philipsen

Die Rolle von Daten- und Textminingverfahren für die journalistische Arbeit

Abschlussarbeit zum Erlangen des akademischen Grades Master of Arts
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: Prof. Dr. Susanne Draheim

Eingereicht am 06.03.2017

Quirine Philipsen

Thema der Masterthesis

Die Rolle von Daten- und Textminingverfahren für die journalistische Arbeit

Stichworte

Datenjournalismus, Datadriven Journalism, Computer Assisted Reporting, Data Mining, Text Mining, Datenanalyse, Textanalyse, Kontextdaten, Big Data, Data Storytelling, Datenvisualisierung, KDD-Prozess, Algorithmen, Open Data, Open Source

Zusammenfassung

Die Arbeit von Journalisten hat sich durch die Digitalisierung in den letzten Jahren stark verändert. Zusammen mit dem technologischen Fortschritt wächst das Potenzial neuer Erzählformen, Recherchemöglichkeiten und automatisierter Arbeitsabläufe. Neben der Entstehung neuer Medienkanäle, beispielsweise des Social Web, bilden Massendaten aus Big Data zunehmend die Arbeitsgrundlage journalistischer Tätigkeit und haben den Fachbereich Datenjournalismus entstehen lassen.

Diese Arbeit handelt von der heutigen IT-gestützen Tätigkeit als Journalist, dem Umgang mit Massendaten sowie deren Analyse. In der technisch digitalisierten Welt wird es zukünftig zunehmend Aufgabe der Medien sein, relevante Informationen aus der Datenflut zu generieren, sie zu kuratieren und verständlich aufzubereiten. Es konnte zum aktuellen Zeitpunkt festgestellt werden, dass sich viele Arbeitsprozesse bis zu einer bestimmten Arbeitsphase automatisieren lassen und daher die Arbeit ergänzen. Aber Journalisten sind weder durch Computer noch durch Informatiker ersetzbar. Im Gegenteil – die traditionellen Fähigkeiten eines Journalisten, Geschichten aufzuspüren, zu recherchieren, Informationen zu filtern und sie verständlich wiederzugeben, sind in Zeiten von Big Data noch vielfältiger und wichtiger geworden. Sie stehen nicht im Gegensatz zur Anwendung technischer Verfahren und moderner Softwarelösungen. Die Möglichkeiten zur Ausübung des Berufes haben sich lediglich erweitert.

Quirine Philipsen

Title of the paper

The role of data- and textmining procedures for journalist work

Keywords

Data Journalism, Data driven Journalism, Computer Assisted Reporting, Data Mining, Text Mining, Data Analysis, Text Analysis, Context Data, Big Data, Data Storytelling, Data Visualization, KDD Process, Algorithms, Open Data, Open Source

Abstract

Due to digitalization, the work field of journalism has changed dramatically over the last years. Technological progress creates potential for new forms of narration, research and automated workflows. Besides emerging new media channels like the social web, information coming from big data are increasingly building the fundament of today's journalism. This development has led to a whole new special field called data journalism.

This paper explains the importance of journalists working with the most advanced IT technology to date and big data. In the increasingly digitalized world, one of the key elements of journalists´ duties will be to make meaningful conclusions based on large scale data analysis.

At present, numerous processes can already be automated which could significantly help journalists in their work. While some traditional journalists' skills, e.g. an ability to track down a story, will remain almost impossible to replace by machines or technology, some types of research and information analysis will be increasingly done by computers. Computation power becomes especially important in the time of big data. Thus, the technology is to open new horizons for journalists' work.

Danksagung

Ich möchte mich an dieser Stelle bei allen bedanken, die mich in dieser Arbeit unterstützt haben.

Hervorzuheben ist meine Familie, die mir mit viel Verständnis immer den Rücken freigehalten hat und in dieser sehr intensiven Zeit für mich zurückgesteckt hat. Nur durch ihre Unterstützung war die Vereinbarkeit von Beruf, Studium und Familie überhaupt möglich.

Ein weiterer wirklich besonderer Dank geht an Prof. Dr. Kai von Luck und allen, die an diesem Studiengang beteiligt waren. Sie haben mich während der gesamten Zeit begleitet, beraten, motiviert und nachhaltig inspiriert.

Ich möchte allen Freunden für ihr Verständnis, ihren Zuspruch und bedingungslosen Rückhalt in Form von Motivation, Kinderbetreuung, Lektorat oder einfach nur Ablenkung danken. In diesem Zusammenhang seien insbesondere Kathrin Baitinger und Andy Herzberg erwähnt, die meine Studienzeit sehr bereichert haben.

Abschließend gedenke ich noch meinem engsten Freund und Vertrauten, der zum Erfolg dieser Arbeit zu großen Teilen beigetragen hat. Ich bedauere es zutiefst, dass er den Abschluss dieser Arbeit nicht mehr mit mir teilen kann.

.

Inhaltsverzeichnis

1 Einleitung

Der Journalismus erlebt seit einigen Jahren unruhige Zeiten. Wie in vielen Branchen hat die Digitalisierung und Automatisierung auch die Medien stark verändert. Es herrscht weitverbreitet eine große Unsicherheit über das Wegfallen alter Geschäftsmodelle und die technologischen Entwicklungen.

Innovationen erfordern es, neue Wege zu gehen, alte Gewohnheiten aufzugeben, sich weiterzubilden, neugierig zu sein, zu experimentieren, zu investieren und auch Fehler zu machen. Lange existierten alte Denkmuster wie „never change a running system“, die bereits dem Foto-Pionier Kodak zum Verhängnis wurden: Das amerikanische Traditionsunternehmen, das zu seiner Zeit Vorreiter in der analogen Fotografie war, versäumte das Digitalzeitalter und war nach über 110 Jahren insolvent.

Auch im Generationenkonflikt zwischen der alten, hierarchischen und neuen Arbeitswelt finden langsam Veränderungen statt. Wenn es die Wirtschaftlichkeit zulässt, geht der Trend zur Arbeit in interdisziplinären Teams. Es wurde erkannt, dass viele neue Arbeitsfelder ohne die Zusammenarbeit zwischen IT-Wissenschaftlern, Betriebswirtschaftlern aber auch Kreativen nicht funktioniert.

Prof. Dr. Patricia Wolf, Leiterin des Zukunftslabors CreaLab und Forschungskoordinatorin des Instituts für Betriebs- und Regionalökonomie an der Hochschule Luzern [vgl. 2015] weist anhand einer Studie daraufhin, dass interdisziplinäre Teams dabei bisher ihr Potenzial längst noch nicht ausschöpfen. Auffällig seien die gleichbleibend starren Führungsordnungen und Hierarchien sowie zu geringe Alters-, Geschlechts- und Nationalitätsunterschiede der Teammitglieder.

Obwohl es scheint, als würde dieses Jahrhundert Traditionsunternehmen, Arbeitsprozesse und Arbeitsplätze durch den technologischen Fortschritt und ein verändertes Mediennutzungsverhalten nur so verschlingen, schenkt es allen Branchen auch eine neue, gemeinsame und bedeutungsvolle Ressource – Daten. Ob zur Prozess- und Warenlageroptimierung, für Prognose- oder Empfehlungssysteme – Big Data heißt die neue Chance vieler Branchen, die jeden Tag immer mehr mit Daten aus unseren Online-Aktivitäten, Nutzerverhalten und intelligenten Geräten gefüttert werden. Zusammen mit dem technologischen Fortschritt wächst dadurch das Potenzial neuer Erzählformen, Recherchemöglichkeiten und automatisierter Arbeitsabläufe auch im Journalismus. Neben dem traditionellen journalistischen Handwerk, wie beispielsweise der investigativen Recherche, wird es folglich in der rasant wachsenden, technisch digitalisierten Welt auch verstärkt Aufgabe der Medien sein, relevante Informationen aus der Datenflut zu filtern, sie zu kuratieren und verständlich aufzubereiten.

In dieser Arbeit steht daher die journalisische Arbeit mit Daten im Mittelpunkt und welche Bedeutung dabei modernste Datenanalysemethoden spielen. Zur Bewältigung der Massendaten werden die Analysetechniken des Data- und Text Mining im Hinblick auf die journalistische Recherche, Datensammlung, Datenerhebung, Auswertung von Informationen und Repräsentation aus dem aktuellen Stand der Forschung betrachtet sowie aktuell verwendetet Softwarelösungen vorgestellt. Es wird aufgezeigt, welche Bedeutung derartige Automatisierungsprozesse für die journalistische Tätigkeit haben und wie der Umgang mit strukturierten und unstrukturierten Daten erfolgt.

Ziel dieser Arbeit ist es, die zukünftigen Arbeitsabläufe, Aufgaben, Herausforderungen und Chancen als Journalist im digitalen Datenzeitalter zu definieren und zu beurteilen. Es wird beschrieben, welche Grundlagen, Arbeitswerkzeuge und Fachkenntnisse zukünftig für den Beruf hilfreich sein werden. Anhand eines generalisierbaren Beispiels für ein datenjournalistisches Projekt wird der Arbeitsablauf in der Praxis beschrieben und die Rolle journalistischer Arbeit in Zusammenhang mit technischen Verfahren aufgezeigt. Im Rahmen der Zusammenfassung kann eine Gesamtbeurteilung über die Bedeutung moderner Datenverarbeitung im Journalismus gegeben werden.

Es wird deutlich, dass trotz zunehmender Veränderung des Medienberufes und der gesellschaftlichen, politischen und wirtschaftlichen Rahmenbedingungen der Journalismus gleichbleibend bedeutsam ist. Neben zwingend notwendigen Fortbildungen und Interesse am technologischen Fortschritt wird hier ein generelles Umdenken von der analogen in die digitale Welt erforderlich sein. Prinzipiell haben sich die Grundlagen journalistischen Handwerks aber nicht verändert: Die Fachkenntnisse, die Fähigkeit Narrationen zu entwickeln und detektivische Recherchearbeit zu leisten, Informationen zu filtern, sie verständlich und interessant aufzubereiten, sind selbst im eigentlichen Data- und Textminingverfahren unabkömmlich. Die hier erfolgte Ausarbeitung kann aufzeigen, dass sich lediglich das Spektrum neuer Arbeitsmethoden und Möglichkeiten für die Journlisten erweitert hat, die klassischen Aufgaben des Journalismus aber erhalten bleiben.

Aufbau der Arbeit

Diese Arbeit beschäftigt sich mit der Rolle von Daten- und Textminingverfahren im Journalismus. Ziel ist die Herausarbeitung und Bewertung der Möglichkeiten moderner Computerrecherche und Daten-Narrationen. Es soll festgestellt werden, wie sich der Journalismus durch immer fortschrittlichere Computeranwendungen verändert. Es soll herausgestellt werden, welchen Einfluss die Digitalisierung auf traditionelle journalistische Arbeitsprozesse hat und welche Veränderungen und Chancen sich hier ergeben.

Diese Arbeit gibt zunächst eine Einführung in die heutige journalistische Arbeit. Dabei wird besonders auf die aktuelle Berufs- und Arbeitssituation, den Wandel durch die Digitalisierung, die neuen Herausforderungen und Chancen für Journalisten durch Big Data sowie die rechtlichen Rahmenbedingungen eingegangen. Das anschließende vierte Kapitel beschreibt die Arbeit mit strukturierten und unstrukturierten Daten. Als allgemeine Grundlage werden einzelne Arbeitsprozesse der Datensammlung und Analyse beschrieben. Ein besonderer Fokus liegt anschließend auf der automatischen Verarbeitung von Text und den Herausforderungen der Computerlinguistik. Das folgende Kapitel fünf veranschaulicht die vorher beschriebenen Themenfelder anhand eines beispielhaften Anwendungsfalls aus der Praxis. Eine abschließende Zusammenfassung und ein Ausblick in Kapitel sechs geben Aufschluss über zu empfehlender zukünftiger Arbeitsweisen infolge der neuen technologischen Chancen.

2 Daten im Journalismus

2.1 Definitionen

Im Folgenden werden zunächst Begrifflichkeiten eingeführt, die als Grundlage zum Verständnis dieser Arbeit dienen.

Big Data

Big Data lässt sich über vier Merkmale definieren: Datenmenge (*engl. volume*), Datenvielfalt (*engl. variety*), Geschwindigkeit (*engl. velocity*) sowie die Analyse und Wertschaffung aus diesen Daten (*engl. value*) [vgl. Oracle: 2016]. Wenn hinsichtlich dieser Kategorien herkömmliche Methoden zur Speicherung und Analyse der Werte aufgrund des Datenvolumens nicht mehr ausreichen, wird die beschriebene Größenordnung unter dem allgemeinen Begriff Big Data zusammengefasst [vgl. Bagnoli, Marten, Wagner: 2012]. Es handelt sich ursprünglich um eine neue Wortkreation aus den Begriffen BI (Business Intelligence) und Data Warehouse.

Der Bericht von IDC (vgl. ebd.) bestätigt den vorherrschend schwierigen Umgang mit den Massendaten: Nur 22% der Daten können bisher analysiert werden und nur 5% davon werden bisher analysiert. Das Wissen und die benötigten Technologien sind bisher noch am Anfang. Täglich werden weltweit riesige Datenmengen aus unterschiedlichsten Bereichen wie beispielsweise Internet, Mobilfunk, Finanzindustrie, Energiewirtschaft, Gesundheitswesen und Verkehr erzeugt. Laut dem Cisco Visual Network Index (VNI) wurden 2015 mehr als 20.000 GB Daten pro Sekunde ausgetauscht, die sich bereits bis 2020 verdreifachen sollen [vgl. Cisco: 2016].

> ***„Today, making sense if big data, particularly unstructured data, will be a central goal for data scientists around the world, whether they work in newsrooms, Wall Street, or Silicon Valley.“***
>
> *Alex Howard, O'Reilly Media [Gray et al.: 2012, S. 17]*

Datenjournalismus

Unter Datenjournalismus (*engl. data driven journalism*) wird allgemein der journalistische Prozess verstanden, der auf Datensammlung, Analyse und Filtern von großen Datensätzen basiert, um neue Geschichten zu generieren [vgl. Jakubetz: 2013]. Dies ist zunächst unabhängig von Ressorts. Der Journalistik Professor Mark Coddington [vgl. Coddington: 2014] unterscheidet genauer zwischen:

- **Computer Assisted Reporting**
  (Der Umgang mit Daten ist erkenntnisorientiert und unter Anwendung sozialwissenschaftlichen Methoden. Der Journalismus basiert auf öffentlichem Interesse.)
- **Datajournalism**
  (Im Mittelpunkt steht die öffentliche Beteiligung und Arbeit mit offenen Daten.)
- **Computational Journalism**
  (Hier stehen Nachrichten-Prozesse, die Automatisierung und Abstraktion im Vordergrund.)

Diese drei Unterscheidungen teilt Coddington weiter in vier Dimensionen zwischen Transparenz vs. Verschlossenheit, professionelle Fachkenntnisse vs. gutes Netzwerk, Big Data (große Datenmengen) vs. gezielten Stichproben sowie aktiver oder passiver Öffentlichkeit ein.

Aufgrund keiner allgemein geltenden Definition des Begriffes „Datenjournalismus“ wird im Rahmen dieser Arbeit von der weiten Definition ausgegangen und der Oberbegriff „Datenjournalismus“ verwendet, der alle verwandten Unterscheidungen wie beispielsweise Programmer-Journalist inkludiert.

2.2 Entwicklung des Datenjournalismus

Datenjournalismus ist kein neues Genre im Journalismus. Narrationen auf Grundlage von Daten und Zahlen sowie statistische Methoden sind beispielsweise aus Sportberichten, Wetterprognosen, Wirtschafts- sowie Börsenberichten bekannt. Die ersten unabhängigen Finanzberichte haben den Ursprung im 17. Jahrhundert, in dem die ersten Tageszeitungen in Europa aufkamen [vgl. Matzat: 2010a]. Als Pionier in grafischen Darstellungsformen von Statistiken gilt der Ingenieur und Volkswirt William Playfair. Er erfand Ende des 18. Jahrhunderts das Balken- und Kreisdiagramm [vgl. Howard: 2014, S. 9].

Mitte des 20. Jahrhunderts erleichterten Computer die systematische und organisierte Arbeit. CAR (*engl. Computer Assisted Reporting*) bezeichnet die Unterstützung von Computertechnik zur Datensammlung und Analyse, um Nachrichten zu verbessern. Die Mathematikerin und Informatikerin Grace Hopper entwarf im Team beispielsweise statistische Modelle für den ersten kommerziellen Computer der USA (Univac). Mit einer Stichprobe von 5% und nur 1% Abweichung vom Endergebnis sagte sie die Präsidentschaftswahl zugunsten Dwight D. Eisenhower voraus [vgl. ebd., S. 10 ff.].

Journalist Philip Meyer prägte in den siebziger Jahren den Begriff des Präzisions-Journalismus (*engl. precision journalism*). Er kritisierte die geringe Anwendung von Informatik, unzuverlässige Quellen und geringe Qualität im Journalismus. Er forderte mehr sozial- und verhaltenswissenschaftliche Forschungsmethoden wie Umfragen und öffentliche Protokolle in der journalistischen Praxis [vgl. Gray et al.: 2012, S. 25].

Im 20. Jahrhundert wurde die journalistische Arbeit durch verbesserte Recherchemöglichkeiten und Datenintegration geprägt. Mit dem Internet, das als Medium zunehmend aufwendige Recherchen, Visualisierungen und Nutzerinteraktionen zuließ, entstand schließlich die Grundlage für den Datenjournalismus.

Die Entwicklung und Verbreitung des Internets sowie der rasante Fortschritt neuer Technologien verändert auch das allgemeine Mediennutzungsverhalten.[1] Die vermehrte Nutzung mobiler Endgeräte, die Optimierung von Inhalten auf unterschiedliche Anwender und Displaygrößen, die zunehmende Bedeutung von Benutzerkontext sowie die Entwicklung der sozialen Netzwerke bieten neue Chancen, aber stellen die Medien auch vor große Herausforderungen. Sie müssen sich im digitalen Zeitalter neu erfinden, denn traditionelle Medienkanäle und Geschäftsmodelle verändern sich. Die Pressearbeit, die über lange Zeit zwischen wenigen großen Medienhäusern fest aufgeteilt war, bekam neue Konkurrenz. Die Werbeumsätze wurden neu verteilt und die Qualität durch Konsumenten stärker geprüft. Zeitlich vordiktierte Fernsehausstrahlungen oder gedruckte Tageszeitungen mit veralteten Informationen werden abgelöst, denn das Internet kann Informationen sofort in Echtzeit liefern. [vgl. Gray et al.: 2012, S.13].

2006 beschäftigte sich der amerikanische Journalist und Programmierer Adrian Holvaty mit der Notwendigkeit einer neuen Organisationsstruktur, die später Datenbank-Journalismus genannt wurde. Er kritisiert in „a fundamental way newspaper sites need to change“, dass journalistische Fakten nicht einheitlich in strukturierter Form gespeichert werden, obwohl sie oft gleich sind und sich an den berühmten W-Fragen[2] orientieren [vgl. Holovaty: 2006].

Holovaty wurde unter anderem durch die Veröffentlichung einer Stadtkarte von Chicago bekannt, auf der alle polizeilich registrierten Kriminalitätsvorfälle visualisiert wurden (*chicagocrime map*).

> ***„ […] At the same time, as the Web pumps out more and more data, readers from around the world are more interested in the raw facts behind the news than ever before. When we launched the Datablog, we thought the audiences would be developers building applications. In fact, it's people wanting to know more about carbon emissions, Eastern European immigration, the breakdown of deaths in Afghanistan, or even the number of times the Beatles used the word "love" in their songs (613).“***
>
> *Simon Rodgers, leitender Datenjournalist des Guardian [op. cit., S.39]*

[1] Aktuelle Studien zum Mediennutzungsverhalten sind nachzulesen in der jährlich erscheinenden ARD-ZDF-Onlinestudie zur Mediennutzung, beim Statistik Portal Statista.de oder Institut für Demoskopie Allensbach.

[2] W-Fragen bezeichnet Fragen, die mit dem Anfangsbuchstaben W beginnen. Im Journalismus dient die Beantwortung der Fragen nach was, wer, wo, wann, wie, warum und woher (Quelle) als Grundlage jeder Recherche [vgl. von la Roche: 2008].

Als erste Tageszeitung gründete der englische Guardian im Jahr 2009 einen Datablog mit dazugehöriger einsehbarer und vollständiger Datenbank. Ein Jahr später erschien dort eine detaillierte und interaktive Visualisierung zu den auf der Internetplattform „WikiLeaks“[3] veröffentlichten Tagebüchern des Afghanistan-Krieges (*engl. Afghan War Logs*). Es wurden insgesamt 92.201 Rohdaten analysiert und der Öffentlichkeit durch eine verständliche, journalistische Aufbereitung zugänglich gemacht [vgl. ebd., S. 38 ff., 72ff.].

Datenjournalismus beschreibt daraufhin eine eigene Erzählform (*engl. new storytelling*) im Online-Journalismus, dessen Entstehung eine logische Konsequenz aus Big Data ist. Es muss die gewaltige Datenexplosion und Informationsflut in Bereichen der Industrie, Social Media, Regierungen und der Gesellschaft von Redaktionen aber auch Konsumenten bewältigt werden. Ein in diesem Zusammenhang vielfach benutztes Zitat von Rudolf Augstein, Journalist und Herausgeber von „Der Spiegel“ lautet: *„Die Zahl derer, die durch zu viele Informationen nicht mehr informiert sind, wächst.“* Das Kuratieren von Medien und Nachrichten gewinnt folglich zunehmend an Bedeutung, um aus den Daten alle relevanten Fakten für die Öffentlichkeit zugänglich zu machen und verständlich zu präsentieren. Darüber hinaus ist das Potenzial, das aus Massendatenanalysen neues Wissen generiert und durch erweiterte technologische Recherchemöglichkeiten Geschichten findet, nahezu unerschöpflich. Daten werden zu einer neuartigen und immer wichtigeren Quelle.

2.3 Aktuelle Situation der (Daten-) Journalisten

Traditionelle Medienhäuser, die lange durch lukrative Werbeeinnahmen finanziell abgesichert waren, stehen mittlerweile unter starkem finanziellem Druck. Sie müssen Lösungen für stetig sinkende Auflagen, fehlende Werbeeinnahmen und Konkurrenz durch neue Informationskanäle wie dem Social Web finden, die sich durch ihre Schnelligkeit und Reichweite auszeichnen.

> ***„Das Prinzip ist klar: Wir dürfen nicht mehr vom Publizieren herdenken, sondern überlegen, an welchen Stellen Nutzer unsere Geschichten konsumieren können. Journalisten müssen Kommunikationsversteher und Kommunikationsstifter werden, sie müssen Infrastruktur verstehen und erforschen, verstehen, auf welchen Kanälen wie kommuniziert wird.“***
>
> *Marco Maas, Geschäftsführer und Datenjournalist Datenfreunde GmbH [2015, S. 47]*

[3] WikiLeaks setzt sich aus dem hawaiischen Wort *wiki* (schnell) und englischen Wort *leaks* (Lecks, Löcher, undichte Stellen) zusammen. Es handelt sich dabei um eine Enthüllungsplattform, auf der geheime und vertrauensvolle Dokumente beispielsweise zu Korruption, Insiderhandel, Menschenrechtsverletzungen, Datenmissbrauch anonym veröffentlicht werden (*Whistleblowing*). WikiLeaks setzt dabei ein grundsätzliches öffentliches Interesse an den Informationen voraus [vgl. Wikileaks: 2016].

Von Beginn an sind Nutzer daran gewöhnt, dass Inhalte im Internet kostenlos zur Verfügung gestellt werden. Die Einstellung der Nutzer ist laut der Studie vom Institut für Demoskopie Allensbach [vgl. Schneller: 2016] über Jahre gleichbleibend. Knapp 78% der Nutzer haben kein Interesse an bezahlpflichtigen digitalen Ausgaben von Tageszeitungen und Zeitschriften. Die konkurrierenden Informationskanäle haben auf die Entwicklungen in den letzten Jahren unterschiedlich reagiert. Einige bleiben aus Angst vor schwindenden Nutzern weiterhin kostenlos, andere wiederum bieten ihre Inhalte trotz der geringen Bezahlbereitschaft mit unterschiedlichen Abrechnungsmodellen (engl. *Paywalls*) wie Abonnements, Freemium-Modellen (ausgewählte kostenpflichtige Inhalte), Metered Modellen (kostenfreies Kontingent) oder gar freiwilligen Spenden an [vgl. BDZV: 2016b]. Aufgrund der Wirtschaftlichkeit und zur Sicherung der journalistischen Qualität entschließen sich immer mehr Medienhäuser, ihre Inhalte kostenpflichtig zur Verfügung zu stellen. Und trotzdem liegt der Fokus oftmals weiterhin auf Print, nicht auf Pixeln.

Es hat sich aus der wirtschaftlichen Not heraus eine Anspruchshaltung der Medienhäuser gegenüber ihren Mitarbeitern entwickelt. Don Heider, Dekan an der Journalistenschule der Loyola Universität, beschreibt den heutigen Qualifikationsanspruch an Journalisten wie folgt:

> ***„Writing, reporting, copy editing, photography, video shooting and editing, gathering and synthesizing information, verifying facts, communicating ethically, using social media to find and disseminate stories, coding, Web design, page layout, headline writing, search engine optimization. That's a start."***
>
> *[Lynch: 2015, S. 15]*

Die Mitarbeiterzahl in Redaktionen schrumpft und es ist zu beobachten, dass zunehmend Kameramänner gleichzeitig die Aufgaben eines Redakteurs und Tonassistenten übernehmen oder Grafik-Designer als Layouter und Programmierer arbeiten. Eine Fach-Qualifizierung allein reicht nicht mehr aus. Nach einer Umfrage unter amerikanischen Medienunternehmen [Stencel et al.: 2014, S. 10 ff.] geht ein Großteil der Führungskräfte davon aus, dass Web-Producer oder Web-Editoren gleichzeitig neben dem alltäglichen Schreiben, Posten und „Copy Editing" auch die Kunst des Layouters, Designers, Datenanalysten und Programmierers beherrschen.

Obwohl die Wichtigkeit der IT längst bekannt ist, verhindert weiterhin oft die alte hierarchische Denkweise der Medienhäuser den Zusammenschluss der Abteilung mit den Redaktionen zu einem gemeinsamen Newsroom. Viele Führungskräfte ordnen digitale Werkzeuge und Softwareanwendungen als Erweiterung der traditionellen Medien ein, nicht aber als eigenständige Erzählformen. Technische Fachkenntnisse sind kaum in Newsrooms zu finden, allenfalls für die Optimierung der Websites aus werblichen Gründen, und auch die technische Infrastruktur, um den Austausch der Inhalte zwischen den Abteilungen zu gewährleisten, ist oft zu unflexibel [vgl. ebd., S.5-6]. Eine Begründung für die Prioritätensetzung innerhalb der Organisationen liefert das Poynter Institute [Finberg; Klinfer: 2014] in einem Bericht zu journalistischen Fähigkeiten der Zukunft: Bei einer

Befragung bewerteten 55% der journalistischen Führungskräfte das Generieren von Geschichten als wichtigste Aufgabe von Journalisten und die Interpretation großer Datensätze als weniger wichtig. 73% der Lehrer und Professoren sahen das genau umgekehrt. Darüber hinaus hielten 80% der Lehrer und Studenten, aber nur 59% der Journalisten mit Entscheidungskompetenzen das Auswerten und Analysieren von Statistiken, Daten und Graphen für wichtige journalistische Fachkenntnisse der Zukunft.

Nach einem Bericht des Pew Research Center [vgl. Barthel: 2016] sank die Anzahl der angestellten Journalisten in amerikanischen Redaktionen um 10% zum Vorjahr. Viele von Arbeitslosigkeit bedrohte Journalisten wechseln mittlerweile zur Öffentlichkeitsarbeit größerer Unternehmen. Die Arbeitsbedingungen zeichnen sich mittlerweile durch sicherere Arbeitsplätze, mehr Gehalt, bessere Arbeitszeiten und Aufstiegschancen aus [vgl. Amerland: 2013]. Diese Entwicklung ist kritisch zu betrachten, denn sie könnte langfristig zur Schwächung der vierten Gewalt[4], dem unabhängigen Journalismus, führen. Und genauer betrachtet schließt der zukünftige Arbeitsmarkt Journalisten nicht aus. Der niederländische Journalist Jerry Vermanne fasst zusammen:

> ***„In a time when sources are going digital, journalists can and have to be closer to those sources. The Internet opened up possibilities beyond our current understanding. Data journalism is just the beginning of evolving our past practices to adapt to the online."***
>
> *[Gray et al.: 2012, S. 16]*

Wie in vielen Berufen, die von der Digitalisierung und Automatisierung betroffen sind, verändern sich dadurch die Rahmenbedingungen. Dazu stellt der Berliner Datenjournalist Lorenz Matzat [2015] fest: „Das Internet mit seinen massiven Auswirkungen auf nahezu alle Bereiche der Gesellschaft ist grundlegend datengetrieben. Es ist frappierend, dass diese umwälzende Technologie nicht viel mehr Thema im Journalismus ist." Er hält Technik- und Digitalisierungsthemen längst nicht mehr für „special interest" Themen. Sie müssen im Gegenteil mit gleicher Priorität wie Politik- und Wirtschaftsthemen behandelt werden. In allen Branchen wächst derzeit Bedarf an Datenwissenschaftlern, die Data Scientist, Data Analysts, Business Intelligence oder eben Datenjournalisten genannt werden. Daten gelten als wichtigste Ressource des 21. Jahrhundert. Wer sie beherrscht und analysieren kann, ist derzeit sehr gefragt. Bereits 2011 prognostizierte daher die Unternehmensberatung McKinsey & Company, dass 2018 allein in Amerika ein Mangel von 140.000 bis 190.000 Data Scientists herrschen wird [vgl. Manyika et al.: 2011]. Doch der alleinige Fokus auf technisches Wissen würde zunehmend die Besten bei der Suche nach qualifiziertem Personal ausschließen. Laut dem Journalisten Mike

[4] Neben der Gewaltenteilung Exekutive (ausführend), Legislative (gesetzgebend) und Judikative (rechtsprechend) werden die Medien in Demokratien oft als vierte Gewalt bezeichnet. Sie ist den klassischen Staatsgewalten allerdings nicht gleichzusetzen. Ihre Aufgabe ist es zu informieren, kritisieren, diskutieren und aufklären. Sie haben keine direkte Möglichkeit politische Änderungen zu vollziehen oder Missbrauch zu ahnden [vgl. bpb: 2016].

Loukides [vgl. 2012] sollten die Firmen dem oft beklagten Fachkräftemangel mit erweiterten Qualifizierungsmöglichkeiten entgegenwirken.

> ***„New digital technologies bring new ways of producing and disseminating knowledge in society. Data journalism can be understood as the media's attempt to adapt and respond to the changes in our information environment, including more interactive, multidimensional storytelling enabling readers to explore the sources underlying the news and encouraging them to participate in the process of creating and evaluating stories."***
>
> *César Viana, University of Goiás [vgl. Gray et al.: 2012, S. 18]*

Datenjournalismus ist sehr zeit- und ressourcenaufwendig. Die Medienhäuser müssen daher bereit sein, zu investieren und zu experimentieren. Laut der Einschätzung von Matzat [vgl. 2014] sind mindestens 50 % der Arbeit mit Datenerhebung und Aufbereitung verbunden. Gleichzeitig besteht jederzeit das Risiko, abbrechen zu müssen, wenn die Daten keine Informationen hergeben oder es zu aufwendig ist, weiter zu arbeiten (siehe Kapitel 2.9: Erzählformen mit Daten). Nach der anfänglichen Euphorie zum Datenjournalismus setzte in den Redaktionen schnell wieder das routinierte Arbeiten ein. Überall gibt es Konferenzen, Lehrgänge, Vorträge und Literatur, nur nicht in der Praxis [vgl. Lewis: 2015, S. 322 ff). Dass sich die Arbeit aber lohnen kann, zeigt die Veröffentlichung der „Panama Papers" im Frühjahr 2016. Die Süddeutsche Zeitung arbeitete in Kooperation mit dem ICIJ (International Consortium for Investigative Journalists) und 400 Journalisten aus 80 Ländern an der Aufdeckung von weltweit 214.000 Briefkastenfirmen[5], die durch die panamaische Kanzlei Mossack Fonseca verkauft worden waren. Dazu mussten insgesamt 2,6 Terrabyte unterschiedliche Datentypen in Form von PDFs, Emails und Fotos analysiert werden. Mittels einer automatischen Texterkennung, OCR-Prozess genannt, konnten in zwölf Monaten 11,5 Millionen Dokumente maschinenlesbar aufgearbeitet und analysiert werden (*engl. Optical Character Research*, siehe Kapitel 3.11: Textmining). Es gilt als umfangreichstes und einzigartiges Beispiel für investigativen Datenjournalismus [vgl. Obermayer et al.: 2016].

Einer ebenfalls umfangreichen und ressourcenaufwendigen Recherche in unstrukturierten Daten stellte sich 2009 der Guardian nach dem „UK MP Expenses Scandal". Nach einer investigativen Recherche und Aufdeckung eines Spesenbetrugsskandals im Parlament stieg der Druck durch die Bevölkerung, so dass etwa 400.000 Dokumente von der Regierung offengelegt wurden. Dabei handelte es sich zunächst um unbrauchbare Daten, die in vielen unterschiedlichen Formaten abgelegt wurden. Diese Art von Offenlegung und Transparenz wurde auch als „BlackOutGate" bezeichnet, da eine Aufklärung unmöglich schien. Doch der Guardian entwickelte eine bis dahin einmalige, beispiellose

[5] Eine Briefkastenfirma ist eine der Steuerersparnis dienende Firma, die mit ihrem Sitz im Ausland an Geschäftsausstattung über kaum mehr als einen Briefkasten verfügt [www.duden.de].

Crowdsourcing-Kampagne, indem er sich der Hilfe und Schwarmintelligenz[6] der Leser bediente. Alle Dokumente wurden in ihrer Größe vereinheitlicht und in einem für jedermann öffentlich zugänglichen Content Management System veröffentlicht. Integrierte spieltypische Elemente sollten zusätzlich den Anreiz zur Teilnahme an der Recherche schaffen, die einer vermeintlich langweiligen Buchhaltertätigkeit glich. So motivierten ein Ranking für den besten Fund, die Detektiv-Arbeit, Neugier und der Wunsch nach Gerechtigkeit kurzerhand 20.000 Briten freiwillig zur Aufklärung und Sichtung des Materials. Das Ergebnis nach kürzester Zeit: viele Strafverfahren und sogar eine Gesetzesänderung [vgl. Daniel; Flew, 2010]. Den Erfolg einer Datenredaktion kann Simon Rodgers vom Guardian Datablog bestätigen: Die Verweildauer der User beträgt bei datenjournalistischen Geschichten des Guardian durchschnittlich sechs Minuten gegenüber allen restlichen Geschichten von nur einer Minute [vgl. Gray et al.: 2012, S. 148]. Für Lorenz Matzat [2015] ist die bisherige Ausübung professionellen Datenjournalismus zu wenig: „So gibt es Themen, die tagesaktuell Bedeutung haben. Wo war etwa das Datenstück zu Griechenland? [..] hat im deutschsprachigen Raum sich wirklich jemand den europäischen Finanzdaten, dem ganzen Schuldenuniversum der EU gewidmet und das verständlich – gestützt durch Visualisierungen – runtergebrochen? Dass dies nicht geschehen ist, ist fahrlässig und für das Genre Datenjournalismus eigentlich beschämend. “

2.4 Die Arbeit als Datenjournalist

> ***„You need the ideas of the reporters, to mesh with the ideas of developers.“***
> *Marty Kaiser, Milwaukee Journal Sentinel [Stencel et al.: 2014, S.11]*

Datenjournalisten erzählen Geschichten mit Daten und Zahlen oder finden sie in ihnen. Doch die Arbeit als Daten-Journalist besteht nicht nur aus der Recherche, der Wiedergabe von Statistiken oder aufwendigen Visualisierungen. Während CAR die Daten als reine Quelle für Analysen und Beweismittel nutzt, besteht im Datenjournalismus neben der Präsentation der Anspruch, die Daten zusätzlich zu veröffentlichen. Es soll Transparenz und Wiederverwendbarkeit der Datensätze nach dem Open Source Gedanken (siehe auch Kapitel 3.13.1: Arbeitswerkzeuge für Journalisten, Open Source) sowie Anwenderfreundlichkeit und individuelle Interaktion geschaffen werden [vgl. Howard: 2014, S. 16 ff.]. Daten sammeln, strukturieren, organisieren, analysieren und visualisieren wird zusätzlich mit der traditionellen Arbeit des Journalisten wie Hintergrundrecherchen, Expertenbefragungen und Interviews kombiniert. Es erfordert neben dem Gespür für Geschichten und Zusammenhänge sowie der gesunden Skepsis gegenüber Fakten auch interdisziplinäre Fähigkeiten aus Wissensbereichen der Informatik, Statistik, Soziologie und Verständnis für große Datenbanken [vgl. ebd., S.15].

[6] Schwarmintelligenz meint die Macht der Masse, die aus gebündelten Fähigkeiten von Individuen entsteht.

Anforderungen und Kompetenzen

Die Anforderungen an Journalisten und ihr Aufgabenspektrum haben sich durch die Digitalisierung geändert. Heute sind Kenntnisse in der Programmierung nicht zwingend nötig, aber Journalisten sollten mindestens digital affin sein.

> ***„Data journalism is a new set of skills for searching, understanding, and visualizing digital sources in a time when basic skills from traditional journalism just aren't enough."***
>
> *Jerry Vermannen, NU.nl [Gray et al.: 2012, S. 16]*

Besonders Führungskräfte und Entscheidungsträger müssen künftig über ein Grundwissen verfügen, das technische Aspekte der Informatik wie künstliche Intelligenz für Voraussagemodelle, Empfehlungssysteme und Personalisierung sowie Datamining für Content-Analysen umfasst**.** Der Arbeitsalltag wird zunehmend durch Analysemöglichkeiten bestimmt, die klare Aussagen über Zielgruppe, Inhalte, Bewertung, Zeitpunkt, Verweildauer, Medium und kontextuale Zusammenhänge treffen. Das Nutzungsverhalten wird interpretierbar und das gewonnene Wissen ist die Grundlage für erfolgreiche Geschäftsmodelle (siehe Kapitel 2.9: Erzählformen mit Daten). Anstatt die Nachrichten wie früher auf verschiedene Tage zu verteilen, geben heute Daten-Analysten vor, wann und wo die beste Zeit ist, Inhalte zu publizieren. Sie verstehen die Zusammenhänge.

> ***„In kaum einer anderen Branche wird über technologische Wissensdefizite so leichtfertig hinweggesehen wie im Journalismus – aber als ausgewiesene und ausgebildete Kommunikationsexperten sollte es uns peinlich sein, moderne Kulturtechniken wie Verschlüsselung oder das Internet der Dinge nicht einordnend vermitteln zu können."***
>
> *Marco Maas, Geschäftsführer und Datenjournalist von Datenfreunde GmbH [2015, S. 47]*

Aus diesem Grund werden zukünftige Datenjournalisten oftmals nicht auf Journalistenschulen ausgebildet. Sie studieren zum Beispiel eher Informatik oder Naturwissenschaften [vgl. Howard: 2014, S. 47 ff]. Obwohl an datenjournalistischen Projekten interdisziplinäre Teams aus Programmierern, Analysten, Visualisierern und Redakteuren arbeiten, gilt ein Grundverständnis für Mathe, Statistik und Informatik als wichtige Voraussetzungen.

> ***„Für eine wirklich relevante Darstellung muss der Journalist technisch einsteigen, Systeme entwickeln, mit denen er an Informationen gelangt, die das Erzählen einer Geschichte ermöglichen – oder zumindest mit den Entwicklern um die besten Lösungen ringen."***
>
> *Marco Maas, Geschäftsführer und Datenjournalist von Datenfreunde GmbH [2015, S. 47]*

2.5 Datenherkunft

Die Daten, die als Grundlage journalistischer Publikationen dienen, können eigens erhoben sowie von öffentlichen oder privat unternehmerischen Quellen recherchiert sein (Siehe auch Kapitel 3.5: Datenrecherche). Solange die Daten personenbezogen sind, können Betroffene ihre Daten bei öffentlichen sowie privaten Stellen einfordern. Auf die rechtlichen Rahmenbedingungen und Grundlagen wird im Kapitel 2.6 eingegangen. Die Aushändigung der Daten in Zusammenarbeit mit Firmen wird Corporate Publishing[7] genannt. Diese Kooperation gehen die Unternehmen nur ein, wenn sie Vorteile von der Datenanalyse haben wie beispielsweise bei Werbezwecken. Ansonsten schützen sie ihre sensiblen Daten und Geschäftsgeheimnisse vor Betriebsexternen. Es wird im Datenjournalismus überwiegend investigativ und unabhängig gearbeitet, also in der ursprünglichsten Form des Journalismus. Folgend werden verschiedene grundlegende Datenquellen sowie datenjournalistische Erzählformen vorgestellt.

Offene Daten

Mit dem Medium Internet ist auch die Richtigkeit von Informationen leichter überprüfbar geworden. In Blogs, Foren und sozialen Netzwerken werden Meinungen und Informationen schnell verbreitet und ausgetauscht, verglichen, kontrolliert, kommentiert, weitergeleitet und diskutiert. Laut Simon Rodgers sind Journalisten nicht länger die Gatekeeper für Informationen [vgl. Gray et al.: 2012, S. 39]. Der Informationsaustausch in den unterschiedlichen Medienkanälen ist schnell und vielfältig. Besonders in den sozialen Netzwerken wird sofort auf neue Publikationen reagiert, Fehl- und Falschmeldungen entdeckt und Ungenauigkeiten gemeldet (siehe auch Kapitel 2.8: Sorgfaltspflicht von Journalisten). Das schafft große Transparenz und daher steigt auch das öffentliche Interesse an öffentlichen Daten und deren Zugänglichkeit. Für Datenjournalisten sind offene Daten eine wichtige Grundlage ihrer Arbeit. Sie generieren neue Geschichten, gewährleisten eine Reproduzierbarkeit und journalistische Quellen werden nachvollziehbar. Simon Rodgers sieht großes Potenzial im „Open Data Journalism“. Er empfiehlt für eine erfolgreiche Praxis die gleichzeitige Offenlegung freier Roh-Daten sowie der anwenderfreundlichen, maschinenlesbaren und modifizierten Daten. Weiter nennt er die Personalisierungsmöglichkeit durch Interaktion und Visualisierung als wichtigen Bestandteil [vgl. 2013].

[7] Das Content Marketing Forum, ehemals Verband Forum Corporate Publishing (FCP), definiert Corporate Publishing wie folgt: „Corporate Publishing bezeichnet die einheitliche interne und externe, journalistisch aufbereitete Informationsübermittlung eines Unternehmens über alle erdenklichen Kommunikationskanäle (offline, online, mobile), durch welche ein Unternehmen mit seinen verschiedenen Zielgruppen permanent/periodisch kommuniziert.“ Dabei soll ein Informationsmehrwert geschafft werden und sich von Werbung klar abgrenzen. (http://huffmann-business.de/was-macht-corporate-publishing-aus/)

Definition Open Data

Nach der Definition der Open Knowledge Foundation Deutschland[8] [vgl. 2016] bezeichnet Open Data alle Daten, die nicht personenbezogen sind sowie bei Nennung des Urhebers von jedermann frei zugänglich und nutzbar sind. Zu den wichtigsten Eigenschaften zählen Vollzähligkeit und Maschinenlesbarkeit. Nach Daniel Dietrich, offizieller Repräsentant der Open Knowledge Foundation Deutschland und Mitarbeiter der Bundeszentrale für politische Bildung [Dietrich: 2011a], handelt es sich um Open Data *„[..]* wenn es keine rechtlichen, technischen oder sonstigen Kontrollmechanismen gibt, die den Zugang, die Weiterverarbeitung und die Weiterverbreitung dieser Daten einschränken. Der Zugang, die Weiterverarbeitung und die Weiterverbreitung soll jedermann und zu jeglichem Zweck, auch kommerziellem, ohne Einschränkungen und Diskriminierung und ohne Zahlung von Gebühren möglich sein.“ Es handelt sich um Daten, die zum einen mit öffentlichen Geldern und Steuern erhoben wurden. Aber auch privatwirtschaftliche Unternehmen, Bildungseinrichtungen, Medien sowie Non-Profit-Organisationen generieren Daten, die vom öffentlichen Interesse sind [von Lucke: 2001, S.5]. Das können zum Beispiel geolokalisierte Daten von Gebäuden und Straßen zur Erstellung von Karten, Datenerhebungen aus Dienstleistungen, Klimadaten, Finanzdaten über Ein- und Ausgaben, Umwelt- und Verkehrsdaten sowie Daten aus wissenschaftlichen Publikationen sein. Laut Thomas Thurner [vgl. 2011], Herausgeber von Open Government Data Weißbuch Österreich, fördert eine gute digitale Infrastruktur eines Landes die Zusammenarbeit vieler Bereiche. Er sieht große Vorteile in der Veröffentlichung von Regierungsdaten bei der Wiederverwendung durch andere sowie der einheitlichen internen und externen Integration verschiedener Datenbestände öffentlicher Einrichtungen.

Definition Open Government

Im Zusammenhang von offenen, zugänglichen Behörden- und Regierungsdaten sowie Verwaltungshandeln wird zusätzlich der Begriff Open Government verwendet.

> ***„Offene Verwaltungsdaten sind jene Datenbestände des öffentlichen Sektors, die von Staat und Verwaltung im Interesse der Allgemeinheit ohne jedwede Einschränkung zur freien Nutzung, zur Weiterverbreitung und zur freien Weiterverwendung frei zugänglich gemacht werden.“***
>
> *Professor Dr. Jörn von Lucke und Christian Geiger [2010, S. 6]*

Die Bundesregierung stellt frei nutzbare Datensätze zu Themeninhalten beispielsweise aus Politik, Wirtschaft, Geographie, Sozialem, Verkehr, Gesundheit, Haushalt auf dem Internetportal https://www.govdata.de/ zur Verfügung. Bei Open Government Data geht es zusätzlich zur reinen Informationsauskunft auch um die barrierefreie Bereitstellung einheitlich, strukturierter, vollständiger

[8] Die Open Knowledge Foundation Deutschland e.V. ist ein gemeinnütziger Verein und Teil des weltweit aktiven Netzwerkes Open Knowledge Foundation, das sich für offenenes Wissen, offene Daten und Transparenz einsetzt. Die deutsche Initiative wurde 2011 in Berlin gegründet (http://okfn.de/).

Datensätze. Damit wird auf die Forderungen der Open-Data-Bewegung nach vollständiger Transparenz reagiert. Christian Heise von der Open Knowledge Foundation beschreibt die Relevanz von Open Government wie folgt [Kuzev: 2015]: „Der Zugang zu diesem Wissen stellt in einer gut funktionierenden, demokratischen Gesellschaft für Bürgerinnen und Bürger eine wichtige Voraussetzung zur gesellschaftlichen Teilhabe dar." Besonders der jüngere Teil der Bevölkerung kann die vergangene Verschleierung politischen Handelns nicht verstehen, da viele mit dem Internet und dem dazugehörigen Transparenzgedanken aufgewachsen sind [vgl. Hackmack: 2014, S. 12]. Open-Data-Aktivisten und später auch die Sunlight Foundation[9] haben sich auf zehn Prinzipien für offene Verwaltungsdaten geeinigt [vgl. op. cit.: 2016]:

1. **Vollständigkeit** (alle Datensätze enthalten Formeln und Erklärungen)
2. **Rohdaten** (nicht modifiziert oder in gefilterter Form)
3. **Aktualität**
4. **Zugänglichkeit** (ohne physische und technische Barrieren, setzt eine Bereitstellung von Programmierschnittstellen und eine generelle Auffindbarkeit voraus)
5. **Maschinenlesbarkeit** (Ermöglichung der Weiterverwendbarkeit durch Formate wie .txt, .csv, .json, .xml, .rss.)
6. **Diskriminierungsfreiheit** (Gleichberechtigung bei Zugriff auf Daten, unabhängig von Zeitpunkt und ohne Angabe von Gründen)
7. **Offene Standards** (Gewährleistung der Verfügbarkeit ohne spezielle Software und eventuelle verbundenen Lizenzkosten)
8. **Lizenzierung** (Zugänglichkeit ohne Restriktionen oder Nutzungsbedingungen)
9. **Dauerhaftigkeit** (Dokumentation der Versionen über einen langen Zeittraum)
10. **Nutzungskosten** (ohne Nutzungsgebühren)

Tim Breners Lee, Begründer des World Wide Web und Direktor des W3C-Konsortiums für Standardisierung der WWW-Techniken, hat für die Umsetzung einer zukünftigen vernetzten Regierungsdateninfrastruktur Richtlinien nach einem 5-Sterne-Modell[10] entwickelt. Das Modell basiert auf internationalen, offenen W3C-Standards (Linked Open Data) und ist wie folgt hierarchisch aufgebaut und gekennzeichnet:

[9] Die Sunlight Stiftung ist eine nationale, unparteiliche und gemeinnützige Organisation mit Sitz in Washington, USA. Sie setzt sich für transparente Regierungsdaten ein und verfolgt schwerpunktmäßig die Offenlegung aller in politischen Geldtransfers (https://sunlightfoundation.com/about/).
[10] https://www.w3.org/2011/gld/wiki/5_Star_Linked_Data

Kein Stern – Daten im Web (Format egal), ohne offene Lizenz WD

Daten im Web (Format egal) mit offener Lizenz

Daten in strukturiertem Format (z.B. Excel)

Daten in strukturiertem, nicht proprietärem Format (z.B. CSV statt Excel)

Verwendung von eindeutigen URLs, so dass Datensätze verlinkt werden können

Verlinkung der eigenen Daten mit anderen Daten, um Kontext herzustellen

Abbildung 1: 5-Sterne-Modell von Tim Breners-Lee *[vgl. 2011]*

Die Recherche in offen Daten ist oft noch sehr mühsam, da die Bereitstellung nicht in allen Instanzen einheitlich geregelt ist. Laut Thomas Tursis von der Konrad Adenauer Stiftung [vgl. Kuzev: 2015] erschwert die föderale Struktur in Bund, Ländern, Kreisen und Kommunen Deutschlands die übergreifende Datennutzung teils erheblich. Einige Daten werden zentral verwaltet, andere wiederum individuell in den Gemeinden. Ebenso ist es laut Dietrich [vgl. 2011b] mit den unterschiedlichen, teilweise inkompatiblen Lizenzen, die jegliche Weiterverwendung erschweren. Er spricht sich weiter für eine Vereinheitlichung „...der verwendeten Vokabulare und Klassifikationen zur semantischen Beschreibung der Daten“ aus, damit zukünftig ein global einheitlicher Standard einzuführen ist. Die offenen Standards, Formate und Lizenzen von Open Data sollten dazu dezentral, föderiert organisiert und festgelegt werden.

Im Umgang mit relevanten Daten wird oft versucht, den Zugang und die Herausgabe mit unrechten Mitteln zu erschweren oder ganz vorzuenthalten. Journalisten stoßen in ihrer Recherche oft auf übertriebene Forderungen von hohen Gebühren, lang andauernde Bearbeitungszeiten, fehlende Ansprechpartner oder die schlichte Weigerung durch die Berufung auf beispielsweise Urheber- oder Datenschutzrechte. Das Informationsfreiheitsgesetz bietet bisher noch zu viele Ausnahmeregelungen, so dass beispielsweise nur aufgrund des Wettbewerbs relevante Informationen zurückgehalten werden können. Bereits die Recherche nach Daten erfordert eine hohe Ausdauer, Beharrlichkeit und eventuell sogar finanzielle Mittel, um den nötigen Rechtstreit eingehen zu können. Generell gilt Deutschland nicht als Vorreiter für die Offenlegung von Daten. Das hat mitunter auch kulturelle Gründe [vgl. Krabina: S. 279, 2012]. Nach Einschätzung des neuseeländischen Journalisten Charles Andresen [vgl. Howard: 2012, S. 68] besteht beispielsweise ein großer Unterschied zwischen USA und Deutschland. Deutschland hat auf der einen Seiten alle relevanten Gesetze, die Transparenz fördern. Doch im Gegensatz dazu wird hier nicht die Kultur des Teilens gepflegt, die eine der wichtigsten Bestandteile aller Open-Bewegungen darstellt. Open Data wird zukünftig zur Grundlage eines demokratischen Rechtstaates frei nach der Redensart: Wer sich nichts zu Schulden kommen lässt, der

muss auch nichts befürchten. Wenn Regierungen wissen, dass sie kontrollierbar sind, zwingt es sie zur Einhaltung von Gesetzen und Regularien und verhindert Korruption.

Datenjournalisten können unter *http://dataportals.org* weltweit nach offiziellen Open-Data-Portalen suchen. Dort sind auch bekannte Portale wie *Guardian World Goverment Data* (Datenblog der britischen Tageszeitung Guardian) und *Data.gov.* (amerikanisches Datenportal) zu finden. Die Bereitstellung und Qualität von offenen Regierungsdaten ist generell jedoch sehr unterschiedlich. Auskünfte gibt auch die Initiative Open Data Research Network – eine Zusammenarbeit von der Web Foundation und dem International Development Research Centre (IDRC) – die im Netzwerk zu Open Data forscht. Als weiteres Beispiel gilt auch das Datenportal der Vereinten Nationen. Hier werden von der Statistikabteilung (UNSD) in 35 Datenbanken 60 Millionen Datensätze unter data.un.org bereitgestellt. Und auch die Initiative UN-Habitat stellt unter urbandata.unhabitat.org offene Daten zu Städten und Ländern weltweit zur Verfügung [vgl. Gray et al.: 2012, S.96-100].

Beispiel Hamburger Transparenzgesetz[11]

Im Oktober 2012 trat das Hamburger Transparenzgesetz (HmbTG) in Kraft, das zuvor auf Grundlage der Volksinitiative „Transparenz schafft Vertrauen“ im Juni von der Hamburger Bürgerschaft einstimmig beschlossen wurde. Das Gesetz ersetzt das bis dahin in Hamburg geltende Informationsfreiheitsgesetz (HmbIFG) mit dem Unterschied, dass zukünftig eine aktive Bereitstellung der öffentlichen Daten gewährleistet werden soll. Ziel ist es, alle „vorhandenen Informationen unter Wahrung des Schutzes personenbezogener Daten unmittelbar der Allgemeinheit zugänglich zu machen und zu verbreiten, um über die bestehenden Informationsmöglichkeiten hinaus die demokratische Meinungs- und Willensbildung zu fördern und eine Kontrolle des staatlichen Handelns zu ermöglichen.“ (§ 1 Abs. 1 HmbTG). Nach § 1 Abs. 2 des HmbTG hat jede Person fortan „Anspruch auf unverzüglichen Zugang zu allen Informationen der auskunftspflichtigen Stellen...“. Das dazugehörige Transparenzportal, ein Informationsregister, hat zur Aufgabe, einen zentralen Zugang zu öffentlichen Verwaltungsdaten zur Verfügung zu stellen und eine einfache Suche nach Inhalten zu beispielsweise Mitteilungen des Senats an die Bürgerschaft, amtlichen Statistiken und Tätigkeitsberichten, Gutachten und Studien, Geodaten oder Subventionen zu ermöglichen.

Am Hamburger Beispiel lässt sich erkennen, wie aufwendig eine strukturierte und einheitliche Datenintegration ist. Neben technischen Schnittstellen zur automatisierten Datenabfrage werden die Daten teils manuell eingepflegt. Überwiegend handelt es sich um PDF-Dokumente, die eine Einhaltung des § 10, Abs. 1 des (HmbTG) gewährleisten: „Alle Dokumente müssen leicht auffindbar, maschinell durchsuchbar und druckbar sein.“ Doch um den großen Verwaltungs- und Behördenapparat nach den Vorgaben der Open-Government- und Open-Data-Bewegungen zu verändern, muss zusätzlich eine Aufbereitung von Daten vollzogen werden, damit sie nach persönlichen Interessen weiterverwendet

[11] http://transparenz.hamburg.de/

werden können. Neben den technischen Hürden gibt es auch rechtliche Konflikte: Auskunftsverweigerungen, Schwärzungen in Veröffentlichungen sowie geschützte Geschäftsgeheimnisse sind umstritten und führen die Diskussionen über die in § 4 Abs. 1 aufgeführten Ausnahmen des Transparenzgesetzes an. Johannes Caspar, Hamburgischer Beauftragter für Datenschutz und Informationsfreiheit, stellt in seinem Tätigkeitsbericht Informationsfreiheit [vgl. 2015] beispielsweise den Umgang mit dem Landesamt für Verfassungsschutz (§ 5 Nr. 3 HmbTG) in Frage: „Eine generelle Ausnahme ist jedoch nicht zeitgemäß und – wie die Beispiele aus anderen Bundesländern zeigen – auch nicht erforderlich." Weiter kritisiert er in seiner Bestandsaufnahme die angedachte freiwillige Beteiligung der mittelbaren Staatsverwaltung. Lediglich zwei Anstalten des öffentlichen Rechts („fördern & wohnen" sowie „Hamburger Friedhöfe") pflegen Daten in das Transsparenzregister. Hamburgs Transparenzgesetz gilt trotz allem als wegweisend in Deutschland.

2.6 Rechtliche Grundlagen datenjournalistischer Arbeit

> ***„[...] Ich glaube, wenn ich nicht fit bin in Presserecht und im Informationsfreiheitsgesetz, dann fehlen mir als Datenjournalist mindestens anderthalb Arme und ich kann ganz, ganz viele Geschichten als Journalist nicht machen. [...]"***
>
> *Datenjournalist Daniel Drepper von CORRECT!V [2016, ab Min. 14:08]*

Es gibt verschiedene Gesetze in Deutschland, die den Umgang mit Informationen und Daten regeln. Die Arbeit von Journalisten wurde gesondert aufgeführt. Folgend werden alle gesetzlichen Rahmenbedingungen vorgestellt, die für die datenjournalistische Arbeit relevant ist.

Grundgesetz Artikel 5

Das in der Bundesrepublik Deutschland geltende Grundgesetz (GG) von 1949 gewährleistet die Basis journalistischer Arbeit. In Artikel 5 ist die allgemeine Pressefreiheit ohne jegliche Zensur festgelegt. Es legitimiert Journalisten in ihrer Recherche und gilt als Grundlage des Presserechts. Die Einschränkungen bilden dabei der besonders zu behandelnde Jugendschutz, die Persönlichkeitsrechte sowie gesondert aufgeführte, beeinflussende Gesetze.

> *(1) Jeder hat das Recht, seine Meinung in Wort, Schrift und Bild zu äußern und zu verbreiten und sich aus allgemein zugänglichen Quellen ungehindert zu unterrichten. Die Pressefreiheit und die Freiheit der Berichterstattung durch Rundfunk und Film werden gewährleistet. Eine Zensur findet nicht statt.*
>
> *(2) Diese Rechte finden ihre Schranken in den Vorschriften der allgemeinen Gesetze, den gesetzlichen Bestimmungen zum Schutze der Jugend und in dem Recht der persönlichen Ehre.*
>
> *(3) Kunst, Wissenschaft, Forschung und Lehre sind frei. Die Freiheit der Lehre entbindet nicht von der Treue zur Verfassung.*[12]

[12] https://www.bundestag.de/grundgesetz

Presserecht und Medienrecht

Auf der Rechtsgrundlage des Grundgesetztes, das die Pressefreiheit in Artikel 5 definiert, ergänzt das Medienrecht die allgemeinen Rahmenbedingungen zur privaten und öffentlichen Anwendung und Anwendbarkeit medialer Inhalte. Es behandelt daher alle Aspekte der Meinungsvielfalt, Medienrezipienten, Daten- und Jugendschutz sowie den Schutz geistigen Eigentums. Nach Art. 30 GG in Verbindung mit Art. 70 Abs. 1 GG obliegt den Ländern grundsätzlich die Gesetzgebekompetenz für Rundfunk und Presse [vgl. Dörr, Schwartmann: 2012, S. 15ff]. In den Landespressegesetzen werden beispielsweise die jeweiligen Auskunftsrechte gegenüber Behörden oder die Impressumspflicht geregelt. Die rasant wachsende Medienentwicklung steht oft trägen Gesetzgebern gegenüber, so dass es trotzdem schnell zu ungeklärten Verhältnissen oder Konflikten kommen kann. Das Medienrecht unterliegt somit gezwungenermaßen oft noch der richterlichen Rechtsprechung (case law).

Rundfunkstaatsvertrag

Der Rundfunkstaatsvertrag ist das Ergebnis der Einigung der Länder, über eine einheitliche Regelung der Aufgaben des öffentlich-rechtlichen Rundfunks. Der Rundfunkstaatsvertrag (RStV) beinhaltet die rechtlichen Rahmenbedingungen zu folgenden Bereichen:[13]

- das duale Rundfunksystem (Koexistenz von öffentlich-rechtlichem und privatem Rundfunk)
- Auftragsdefinition für den öffentlich-rechtlichen Rundfunk
- die Dauer und Form der Rundfunkwerbung (Fernseh- und Radiowerbung)
- das Recht auf Kurzberichterstattung
- die Überwachung der Medienkonzentration
- die Einführung und Nutzung von analogen und digitalen Übertragungsverfahren
- Vorschriften zu inhaltlich geprägten Telemedien[14]
- Einteilung der Sender in die mit Vollprogramm und die mit Spartenprogramm

[13] Seit dem 1. Januar 2016 ist die Achtzehnte Fassung des Staatsvertrags für Rundfunk und Telemedien (Rundfunkstaatsvertrag - RStV) vom 31. August 1991 in Kraft. Online unter: http://www.diemedienanstalten.de (abgerufen am 23. Juli 2016)

[14] Seit 2007 wurde der Rundfunkstaatsvertrag um das Telemediengesetz (TMG) ergänzt. Mit der Entwicklung des Internets wurden multimediale Inhalte seit 1997 zunächst noch getrennt durch das bundesweite Informations- und Kommunikationsdienste-Gesetz (u.a. Teledienstgesetz) und den Mediendienste-Staatsverträgen (MDStV) der Länder geregelt. Als Mediendienste zählen alle Informations- und Kommunikationskanäle, die an die Allgemeinheit gerichtet sind. Teledienste wiederum beziehen sich auf die Individualkommunikation.
Gabler Wirtschaftslexikon, online unter: http://wirtschaftslexikon.gabler.de/Archiv/12699/mediendienste-staatsvertrag-md-stv-v12.html (abgerufen am 23. Juli 2016)

Pressekodex

Der Pressekodex wurde erstmals vom Deutschen Presserat sowie den Presseverbänden 1973 verfasst und definiert die publizistischen Grundsätze. Es wurden selbst auferlegte Maßstäbe zur Sicherstellung der Berufsethik beschlossen, um das Ansehen der Presse zu schützen und die Freiheit der Presse zu bewahren. Der Deutsche Presserat maßregelt Konfliktsituationen selbstständig, unabhängig der juristischen Grundlage.[15] Der Kodex gibt Vorgaben zur Informationsbeschaffung, Informationsverarbeitung und Informationsverbreitung und wird unter folgenden sechzehn Punkten weiter ausgeführt[16]:

- Wahrhaftigkeit und Achtung der Menschenwürde
- Sorgfalt (Überprüfung der Informationen)
- Richtigstellung
- Grenzen der Recherche (keine unlauteren Methoden)
- Berufsgeheimnis (Zeugnisverweigerungsrecht zum Schutz von Quellen)
- Trennung von Tätigkeiten (Gewährleistung der unabhängigen Berichterstattung)
- Trennung von Werbung und Redaktion
- Schutz der Persönlichkeit
- Schutz der Ehre
- Religion, Weltanschauung, Sitte
- Sensationsberichterstattung, Jugendschutz
- Diskriminierungen
- Unschuldsvermutung (objektive Berichterstattung)
- Medizin-Berichterstattung (nicht unangemessen sensationell)
- Vergünstigungen (zur Verhinderung der Bestechlichkeit)
- Rügenveröffentlichung (Veröffentlichungspflicht von Rügen des Presserates)

Informationsfreiheitsgesetz (IFG)

Das Informationsfreiheitsgesetz, in den USA auch „Freedom of Information Act“ (FOIA) genannt, regelt seit 2005 den rechtlichen Anspruch der Bürger auf Einsicht in unveröffentlichte Dokumente der Regierung ohne Begründung oder Voraussetzung. Das Gesetz kann als rechtliche Grundlage und wegweisend für die nachfolgenden Open-Data- und Open-Government-Bewegung angesehen werden [Vgl. Robinson, Yu: 2012]. Bei der Einführung des Gesetzes zeigen sich wiederholt große kulturelle Unterschiede. So gibt es das Gesetz bereits seit 1766 in Schweden und wurde vor mehr als 30 Jahren bereits in den USA eingeführt [vgl. Hackmack: 2014, S. 48 f.]. Die Bundesländer Bayern, Baden-

[15] Springer Gabler Verlag (Herausgeber), Gabler Wirtschaftslexikon, Stichwort: Presserecht, online im Internet: http://wirtschaftslexikon.gabler.de/Archiv/2924/presserecht-v13.html
[16] http://www.presserat.de/pressekodex/pressekodex

Württemberg, Hessen, Niedersachsen und Sachsen haben sich bei der Verpflichtung zur Bereitstellung von öffentlichen Verwaltungs-Dokumenten und Informationen ausgenommen. Laut dem Stern-Journalisten Hans Martin Tillack [vgl. 2013, S. 19] ist der Grund für die Weigerung zur Datenoffenlegung der Schutz der Politiker vor öffentlicher Kritik. Zu viel Transparenz würde das politische Handeln und Ansehen schwächen. Im Sinne der Demokratie fordert Tillack gerade deshalb deutlich erweiterte und allgemein geltende Standards zur Datenoffenlegung.

Ein Paradebeispiel für eine Informationsblockade gibt es aktuell seitens der Bundestagsverwaltung. Seit sieben Jahren bereits versuchte ein BILD-Journalist detaillierte Informationen über Ausgaben von 68 000 Euro zu erhalten, die für Montblanc-Füller im Wert von je 178 Euro angefallen waren. 2014 bestätigte noch das Bundesverwaltungsgericht, dass die persönlichen Daten der Abgeordneten gegenüber der Informationsfreiheit übergeordnet sind [vgl. Zörner: 2016]. Doch am 12. Oktober 2016 entschied nun das Oberverwaltungsgericht Berlin- Brandenburg doch zugunsten der Journalisten auf Auskunftsrecht.

Verschiedene Portale wie FragdenStaat.de, asktheeu.org, alaveteli.org helfen bei der richtigen Anwendung und Fragestellung nach den Regeln des Informationsfreiheitsgesetzes. Bei FOI-Anfragen wurde von den Initiativen *Access Info Europe* und *n-ost* zur Wahrung der Medienfreiheit eine eigene Anleitung für Journalisten entwickelt, die bei Informationsfreiheitsanfragen und rechtmäßiger Ausübung journalistischer Recherche hilft. Das Toolkit für Journalisten ist unter *Leaglleaks.info* abzurufen [vgl. Gray et al.: 2012, S.96-100].

PSI-Richtlinie und Informationsweiterverwendungsgesetz (IWG)

Nach Kuzev [vgl. 2016, S. 3] verpflichtet die PSI-Richtlinie (Richtlinie 2003/98/EG) alle EU-Mitgliedstaaten zur Bereitstellung von zugänglichen behördlichen Informationen und Daten zur Weiterverwendung. Ziel ist es, die Erstellung gemeinschaftlicher Verwaltungsdokumente zu erleichtern, die Nutzung grenzüberschreitend für Unternehmen zu fördern und Wettbewerbsverzerrungen auf den Binnenmärkten entgegen zu wirken. Es handelt sich im Schwerpunkt um eine wirtschaftliche Motivation zur Nutzung von Verwaltungsdaten. Die Entscheidung zur Weitergabe verbleibt bei den einzelnen Staaten und der allgemeine Anspruch auf Zugang von Informationen für Bürger ist dadurch nicht geregelt.

Auf Bundesebene wird die Weitergabe durch das Informationsweiterverwendungsgesetz (IWG) geregelt. Es gilt grundsätzlich, ist aber ebenfalls nicht bindend und verpflichtet nicht gleichzeitig auch zur Bereitstellung. Eine innovative Nutzung und Kombination öffentlicher Verwaltungsdaten soll hauptsächlich Wirtschaftswachstum und soziales Engagement fördern. Es wäre konsequent, wenn standartmäßig alle vom IWG erfassten Daten auch gleichzeitig nach den Grundsätzen von Open Data für alle Bürger zur Verfügung gestellt würden.

Leistungsschutzrecht

Aufgrund der kommerziellen Verbreitung von Presseinhalten durch Suchmaschinen und Internetportale wurde zum 1. August 2013 das Leistungsschutzgesetz beschlossen. Vor allem die Axel Springer SE gilt als ein großer Verfechter dieser Idee [vgl. BDZV: 2016a]. Im Gegensatz zum personenbezogenen Urheberrecht, das bereits das Kopieren ganzer Texte und Zitate regelt, ist das Gesetz unternehmensbezogen und soll dem fairen Wettbewerb dienen. Es wurde zum Schutz der Verlage und Qualitätshüter journalistischer Arbeit im digitalen Zeitalter beschlossen, weil Verleger bisher im Gegensatz zur Film- und Musikindustrie keine explizit rechtlich geregelten Eigentumsrechte besaßen. Auf der einen Seite wird es durch einen großen Interpretationsspielraum stark kritisiert: Die Abgaben sind nicht einheitlich geregelt und einzelne Wörter sowie kleinste Textabschnitte (Snippets) wie beispielsweise die Anzeigentexte der Trefferliste einer Suche von Google sind weiter vom Gesetz ausgeschlossen und ohne eindeutige Definition. Zum anderen ist die Verbreitung von Verweisen auf Inhalte durch Suchmaschinenanbieter eigentlich auch für die Verlage gleichermaßen vorteilhaft. Julia Reda von der Gegen-Initiative „Initiative gegen ein Leistungsschutzrecht“ (IGEL), die unter anderem von Organisationen wie netzpolitik.org, Freischreiber, Computer Chaos Club und Google unterstützt wird: „Eine gesetzliche Einschränkung der freien Verlinkbarkeit führt nicht zu einer besseren Entlohnung von Journalismus, sondern zu Zugangshürden für die Bevölkerung und Verlusten für Verlage, Autorinnen und Autoren.“ [Reda, 2014].

Der Bundesgerichtshof erklärte bereits 2003 im sogenannten Paperboy-Urteil den Internet-Suchdienst für Presseartikel für nicht rechtswidrig [Bundesgerichtshof: 2003]: „Ohne die Inanspruchnahme von Suchdiensten und deren Einsatz von Hyperlinks (gerade in der Form von Deep-Links) wäre die sinnvolle Nutzung der unübersehbaren Informationsfülle im World Wide Web praktisch ausgeschlossen“. Ähnliche Diskussionen und sehr unterschiedliche Lösungsansätze lassen sich auch in anderen Ländern finden: In Frankreich einigte man sich beispielsweise mit der am meisten kritisierten Firma Google, die sich zu der einmaligen Zahlung von 60 Millionen Euro in einen Innovationsfond für Medien verpflichtete. Im Gegensatz zu Spanien, wo sich Google vorerst in Folge einer gesetzlich beschlossenen Kostenübernahme für Kurzmeldungen ganz aus dem Markt zurückgezogen hat. Aktuell wird über eine einheitliche Regelung auf europäischer Ebene diskutiert.

2.7 Datenlizenzen

Im Folgenden werden Datenlizenzen vorgestellt, die für die Arbeit als Datenjournalist wichtig sind. Sie gewährleisten offene Standards für die Bereitstellung offener Daten für Recherche und Weiterverwendung.

Gemeinfreiheit vs. Public Domain (engl. urheberrechtsfrei)

Unter Gemeinfreiheit werden geistige Schöpfungen gestellt, wenn kein Urheberrecht darauf besteht. Ohne Genehmigung und Zahlungsverpflichtung können Werke verwendet werden. Im Gegensatz zur Public Domain (PD) im angelsächsischen Raum, die ohne Einschränkung als „frei von Urheberrecht" gilt, ist die gemeinfreiheitliche Nutzung jedoch nach dem Schutzlandprinzip an die jeweilige nationale Rechtsordnung gebunden [vgl. Peikert: 2012, S. 246 ff., 252]. Trotz Gemeinfreiheit eines Werkes ist es folglich möglich, bei bestimmten Nutzungsanwendungen einzelne Persönlichkeitsrechte zu verletzen. In Europa ist daher das uneingeschränkt geltende Copyright und die Freigabe von geistigen Schöpfungen in die Public Domain umstritten und teilweise nicht anerkannt [vgl. Boyle: 2008]. 2010 schlugen daher die Creative Commons für die Kennzeichnung von Werken die ohne Copyright-Ansprüche in der Public Domain verfügbar sind, das Public Domain Mark (PDM) vor. Das Symbol zeigt ein durchgestrichenes Copyrightzeichen [vgl. Peters: 2010].

Open Data Commons (ODC)[17]

Die ODC-Lizenz gewährleistet Richtlinien zur Nutzung und Bereitstellung von freien Daten insbesondere in Datenbanken.

Datenlizenz Deutschland[18]

Die Lizenz ist der Versuch, auf nationaler Ebene unter einheitlichen Nutzungsbedingungen offene Verwaltungsdaten in Deutschland auf dem Internetdatenportal GovData zur Verfügung zu stellen. Die Lizenz wurde als Empfehlung in freiwilliger Zusammenarbeit mit Bund und Ländern nach der Open Definition entwickelt. Seit Juli 2014 sind zwei Varianten unter Datenlizenz Deutschland „Namensnennung" 2.0 und Datenlizenz Zero 2.0, einschränkungslos als offene Lizenzen vom Open Definition Advisory Council im Juli 2014 anerkannt worden. Christian Heise von der Open Knowledge Foundation Deutschland Heise [Heise: 2014a] sieht jedoch die offenen Versionen der Creative Commons Lizenz aufgrund des internationalen Geltungsbereiches weiterhin als alternativlos an. Er forderte im Rahmen des Bundestagsausschusses „Digitale Agenda": „Der Regelfall in Bezug auf öffentliche Daten sollte Urheberrechtsfreiheit und damit Lizenzfreiheit sein. Lizenzen nutzen in diesem Bereich nicht, sie behindern." [Heise: 2014b]

Copyleft[19]

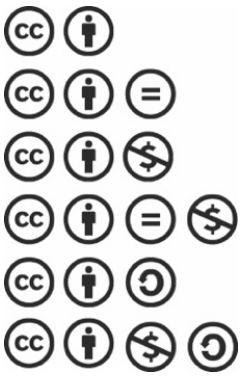
Copyleft gilt als wichtige Arbeitsgrundlage der Open-Source-Bewegung (siehe auch Kapitel 3.13.1). Die Lizenz legt fest, dass bei der Weiterverwendung und Bearbeitung einer Software dieselben Nutzungsbedingungen und Lizenzen des Originalwerkes gelten. So wird ausgeschlossen, dass zum

[17] http://opendatacommons.org/
[18] https://www.govdata.de/lizenzen
[19] Free Software Foundation: http://www.gnu.org/copyleft/copyleft.de.html

Beispiel Unternehmen Open Source als Grundlage für die eigene Firmensoftware nehmen, davon profitieren und nach einer Modifizierung anschließend urheberrechtlich schützen. Ein entscheidender Unterschied zur Gemeinfreiheit: Hier wird zunächst auf das Urheberrecht verzichtet, ebenfalls Veränderungen erlaubt, die dann jedoch als neues Werk wiederum geschützt werden können. Bekannte Beispiele für Softwareanwendungen, die auf Open Source basieren, sind Contentmanagementsysteme wie Typo3, die Bloggersoftware Wordpress, Javaskript-Bibliotheken wie Jquery oder der Web Server Apache. Beispiele für Lizenzen, die das Copyleft-Prinzip mit dem *Share Alike* Standard (engl. für *Weitergabe unter gleichen Bedingungen*) beinhalten:

Creative Commons (CC) [20]

Die Nutzung von Inhalten wie Verbreitung, Vervielfältigung oder Weiterverwendung wurde durch die Digitalisierung vereinfacht und von deutlich mehr Nutzern ohne juristische Fachkenntnisse in Anspruch genommen. Die gemeinnützige Organisation Creative Commons, die 2001 in den USA gegründet wurde, hilft Urhebern bei der Freigabe rechtlich geschützter Inhalte. Es handelt sich meistens um künstlerische und pädagogische Werke. Dafür stellt CC sechs verschiedene Standard-Lizenzverträge mit rechtlichen Rahmenbedingungen für die Verbreitung kreativer Inhalte zur Verfügung. Es lässt sich als eine Erweiterung des Urhebergesetzes einordnen, da die Lizenzen die Möglichkeit zur Abstufung in der Verwendungsart bieten. Die Bestimmungen der jeweiligen Inhalte sind in Form von Meta-Angaben mitgeliefert. Individuelle und nachträgliche Vereinbarungen sind möglich. Bereits der Name des Lizenztypen verrät die wichtigsten Bedingungen:



Namensnennung (by)

Namensnennung, keine Bearbeitung (nd)

Namensnennung, nicht kommerziell (nc)

Namensnennung, keine Bearbeitung, nicht kommerziell

Namensnennung, Weitergabe unter gleichen Bedingungen (sa)

Namensnennung, nicht kommerziell, Weitergabe untergleichen Bedingungen

Abbildung 2: Symbole von Creative Commons *[vgl. 2016]*

Viele verwenden CC-Lizenzen mittlerweile auch als Statement für die Open-Access-Bewegung (*engl. freier Zugang*), für kostenlose wissenschaftliche Literatur.

[20] CC-DE, Creative Commons Deutschland, Europäische EDV-Akademie des Rechts, Merzig, 2016. Online unter: http://de.creativecommons.org/was-ist-cc/

GNU-Lizenzen[21]

Die Free Software Foundation vergibt verschiedene freie Softwarelizenzen mit copyleft wie beispielsweise die GNU General Public License (GPL) oder GNU Free Documentation License (GFDL). Vier Regeln beschreiben seitdem den Umgang mit freier Software [vgl. op cit.: 2007, S.2] und sind Voraussetzung für die Vergabe für GNU-Lizenzen (siehe Datenlizenzen):

- Unbegrenzte Verwendung der Software.
- Uneingeschränkte Möglichkeit zur Untersuchung und Anpassung/Veränderung des Quellcodes.
- Erlaubnis zur Vervielfältigung und Weitergabe des Quellcodes.
- Möglichkeit der Weitergabe eines verbesserten und optimierten Quellcodes.

2.8 Datenschutz

Laut Rechtsanwalt Thomas Schwenke [2013] ist der Journalist im täglichen Zwiespalt zwischen dem „[...]Interesse der Öffentlichkeit an Nachrichten und den Rechten der Betroffenen über die er berichtet. [...] Daher gehört zum journalistischen Handwerk nicht nur ein guter Schreibstil und Gespür für Inhalte, sondern auch die Kenntnis der eigenen Rechte und Pflichten."

Das Bundesdatenschutzgesetz (BDSG) legt die rechtliche Grundlage und Voraussetzung zur Erhebung, Speicherung, Verarbeitung und Nutzung personenbezogener Daten fest. Es regelt in § 19 (1) die Auskunft öffentlicher Stellen und in § 34 (1) die Auskunft nicht öffentlicher Stellen an Betroffene: Demnach hat jede Person Anspruch auf Auskunft über die zur Person gespeicherten Daten sowie deren Herkunft. Weiter müssen bei Aufforderung alle Empfänger der Daten und der Grund der Speicherung offengelegt werden.[22]

Im Zusammenhang mit offenen Daten erklärt Christian Heise [2016]: „Open Data folgt hier der Maxime der Hackerethik: öffentliche Daten nützen, private Daten schützen." Es gibt laut Schwenke [op. cit.: 2013] jedoch ein Medienprivileg: Demnach können „personenbezogene Daten, die ausschließlich zu eigenen journalistisch-redaktionellen oder literarischen Zwecken verwendet werden, von den strengen Datenschutzpflichten im Wesentlichen ausgenommen werden (§ 41 des Bundesdatenschutzgesetzes, in Verbindung mit § 12 der Landespressegesetze und § 57 Rundfunkstaatsvertrag bei Telemedien)."

Unter personenbezogene Daten versteht man alle Angaben wie Name, Adresse, Telefonnummer zu einer Person. Informationen, die keine genaue Identifikation einer Person zulassen, wie Verweildauer

[21] https://www.gnu.org/licenses/license-list#OtherLicenses

[22] Bundesministerium der Justiz und für Verbraucherschutz. Online unter: https://www.gesetze-im-internet.de/bdsg_1990/__34.html (abgerufen am 2. August 2016)

eines Homepage-Besuchers, zählen nicht dazu. Oft gilt die Angabe der persönlichen Daten als Voraussetzung für die Nutzung einer Homepage oder Dienstleistung. Daher spricht man auch von der „Bezahlung mit persönlichen Daten“ anstatt von Geld. Persönliche Daten sind wertvoll geworden und es ist nur eine Frage des Preises oder der Bequemlichkeit, ab wann und wie großzügig der Anwender seine Daten preisgibt.

Und es ist eine deutliche Veränderung im Verhalten mit Selbstauskünften zu erkennen: In den achtziger Jahren wurde noch gegen eine Volkszählung mit Angaben zu Beruf, Schulabschluss und Familienstand demonstriert. Heute geben wir durch die Inanspruchnahme von scheinbar kostenlosen Services beispielsweise von Online-Fitnesstrackern oder Social-Media-Plattformen sehr viel privatere und umfassendere Auskünfte. Firmen wie Acxiom machen sich dieses Verhalten zunutze. Die US-Amerikanische Firma hat Daten von global insgesamt 500 Millionen Menschen, davon 44 Millionen allein in Deutschland. Durch den Abgleich von Offline-Daten (z.B. Register, Statistikämtern, Umzugsservice der Post) und Online-Nutzerdaten verfügt Acxiom nach eigener Auskunft über 1500 Datenpunkte zu jeder einzelnen Person der über 300 Millionen gespeicherten amerikanischen und mittlerweile auch über 44 Millionen deutschen Staatsbürger. Es ist nicht verwunderlich, dass auch Facebook mit Acxiom mittlerweile kooperiert und an Informationen über beispielsweise Alter, Hautfarbe, Vorlieben, Hobbys, Nutzerverhalten, Urlaubswünsche, Krankheiten, Bildungsstand, Einkommen interessiert ist [vgl. McLaughlin: 2013]. Chris Hoofnagle Direktor der Datenschutzabteilung des Center for Law & Technology an der Universität Berkeley in Kalifornien, kritisiert die vorherrschende Praxis und unterschiedliche Auslegungen der Datenschutzrichtlinien in den USA [ebd. 2013]: „Erzähle ich meinem Arzt, dass ich Diabetes habe, bin ich durch das Recht und die Schweigepflicht geschützt. Nehme ich aber an einer Umfrage teil, die indiziert, dass ich krank bin, oder rufe ich eine Gratis-Hotline eines Pharmakonzerns an oder kaufe Medikamente, können diese Daten gesammelt und verkauft werden, ohne geltendes Recht zu verletzen.“ Gerhart Baum, ehemaliger Bundesinnenminister, sieht auch die Notwendigkeit einheitlicher und klarer Datenschutzrichtlinien [2016] und geht auf die Forderung nach einer "Charta digitaler Grundrechte" von Martin Schulz, Präsident des Europäischen Parlaments, und Justizminister Heiko Maas ein: „Wir brauchen eine Weltkonferenz oder ein internationales Algorithmen-Abkommen. In anderen Politikfeldern gibt es solche Vereinbarungen, den Atomwaffensperrvertrag etwa oder das Klimaschutzabkommen. Der massiven Datenaufrüstung muss eine Datenabrüstung folgen.“

Vorratsdatenspeicherung

Im Zusammenhang mit einer anlasslosen Erhebung und Speicherung von personenbezogenen Daten öffentlicher Stellen spricht man auch von Vorratsdatenspeicherung. Diese werden vorrangig mit einer besseren Möglichkeit zur Strafverfolgung oder Prävention begründet und beziehen sich schwerpunktmäßig auf Telekommunikationsdaten. Seit Jahren gibt es sehr unterschiedliche politische Auffassungen und Entscheidungen über die Rechtmäßigkeit. 2009 klagte Malte Spitz, Aktivist und Politiker (Mitglied des Bundesvorstandes von BÜNDNIS 90/DIE GRÜNEN), bei T-Mobile Deutschland

nach BDSG § 34 sein Recht auf Auskunft seiner personenbezogenen Daten und aller dazugehörigen Vorratsdaten ein. Aufgrund der Freigabe konnte ein genaues Persönlichkeits- und Bewegungsprofil angefertigt werden [vgl. Spitz, 2011a]. Zusammen mit Zeit Online wurden die Daten datenjournalistisch aufgearbeitet und visualisiert. Dazu Spitz [2011b]: „Die Veröffentlichung von über 35.000 Datensätzen aus einer sechsmonatigen Vorratsdatenspeicherung meines Mobiltelefons zeigt eindeutig, wie unverhältnismäßig und massiv der Eingriff in die Privatsphäre der Betroffenen ist. Die Unschuldsvermutung wird umgekehrt und jede und jeder unter Generalverdacht gestellt." Dieses Projekt gilt als wegweisend für datenjournalistische Arbeit in Deutschland. Die Einforderung geltenden Rechtes nach § 34 (BDSG) zur Freigabe von gespeicherten Personendaten von einem Privatunternehmen in diesem Umfang gilt als beispielhaft.

Sorgfaltspflicht von Datenjournalisten

Jeder Datenjournalist sollte gegenüber den Quellen skeptisch bleiben, da sie manipuliert oder falsch erhoben worden sein können. Das gilt vor allem bei großen Datensätzen, da die Fehler nicht sofort oder gar nicht erkannt werden. Daher müssen die Quellen immer nach Verlässlichkeit und die Datensätze nach Herkunft, Vollständigkeit und Aktualität geprüft werden. Es gleicht einer Detektivarbeit.

Der Begriff „Dirty Data" beschreibt in diesem Zusammenhang Daten, die falsche Informationen enthalten. Dabei kann es sich um irreführende, doppelte, falsche, ungenaue, nicht integrierte Daten handeln oder Daten, die Geschäftsregeln verletzen, ohne generalisierte Formatierung und unvollständig sind.[23] Sie können auch aufgrund falscher Datenerhebungs- und Verwaltungs- oder Speichermethoden erzeugt werden.

Aber auch die journalistische Datenweiterverarbeitung muss sorgsam erfolgen. Das Sprichwort „Traue keiner Statistik, die du nicht selbst gefälscht hast" hat seinen Grund. Am Beispiel[24] folgender Aussagen wird deutlich, wie unterschiedlich interpretierbar gleiche Daten sein können, obwohl beide wahr sind:

- *Die Wirtschaft schaffte 148.000 neue Jobs, so dass die Arbeitslosenquote auf 7,2% sank!*
- *Nur 148.000 neue Jobs im September! Die Arbeitslosenquote beträgt 7,2%.*

Nicolas Kayser-Bril von Journalism ++ [vgl. Gray et al.: 2012, S. 124] berichtet von einem weiteren Beispiel aus der Praxis. Eine Schlagzeile lautete damals: „Im Durchschnitt ist einer von 15 Europäern Analphabet!". Die Zahl stimmte, denn von 500 Millionen Europäern waren es tatsächlich 36 Millionen,

[23] Definition von Dirty Data von Janalta Interactive Inc. o/a Techopedia.com. Edmonton, Kananda, 2016. Online unter: https://www.techopedia.com/definition/1194/dirty-data
(abgerufen am 8. August 2016)

[24] Weitere Beispiele zu fälschlich interpretierten Statistiken hat die Hamburger Firma Statista unter statista.com/statistik/lexikon/definition/8/luegen_mit_statistiken/ zusammengestellt.

die nicht schreiben und lesen konnten. Es wurde aber nicht berücksichtigt, dass 36 Millionen Europäer unter 7 Jahre waren. Der Journalist Jonathan Stray [2016, S. 36] stellt dazu fest: *[...] „It's been said that data speaks for itself. This is nonsense. [...] the data didn't tell a story, you did. You saw a story that connects the data to the world."* Für Stray ist die Datenanalyse eine Interpretation, die Daten mit Wissen beispielsweise aus bekannten Tatsachen oder kulturellen Gegebenheiten kombiniert. Für sich alleine haben Daten noch keine Bedeutung: *„Imagine a spreadsheet with no column names. It would just be numbers, indecipherable and useless."*

Zusammengefasst bringen Daten Objektivität in Geschichten, die aus der subjektiven journalistischen Aufbereitung und Interpretation entstehen. Nicht alle Daten sollten veröffentlich werden, aber die Offenlegung dient generell der Qualitätssicherung. Es ist vergleichbar mit dem Peer-Review-Verfahren für wissenschaftliches Arbeiten, das die Nachvollziehbarkeit und Kontrolle von Publikationen gewährleistet. Es kann so gegenüber der kritischen Auseinandersetzung mit journalistischen Publikationen, die besonders in sozialen Netzwerken und Expertenblogs stattfindet, mit Transparenz begegnet werden. Gleichzeitig zwingt es die Journalisten zur professionellen und fehlerfreien Berichterstattung.

> ***„Es gibt immer jemanden da draußen in der Welt, der sich bei einem bestimmten Thema noch ein bisschen besser auskennt als der Journalist, der den Artikel geschrieben hat. Und der meldet sich bestimmt."***
>
> *Torsten Beeck, Community Redakteur bei Spiegel Online [2016]*

Datenjournalisten sind verpflichtet, Daten fair und objektiv zu verwenden. Laut Jeff Sondermann [vgl. 2013] sollten das öffentliche Interesse, der Kontext, das Potenzial und die Relevanz stets miteinander abgewogen werden.

2.9 Erzählformen mit Daten

Es gibt unterschiedliche Methoden, wie Journalisten Datensätze verarbeiten können und wie Daten Geschichten erzählen können. Im Folgenden werden einige Anwendungsbereiche und Methoden vorgestellt. Die einzelnen Teildisziplinen zeigen deutlich, dass die vielfältigen Einsatzmöglichkeiten großer Datenbestände auch die interdisziplinäre Teamarbeit zwischen Journalisten, Programmierern, Statistikern, Mathematikern, Designern, Gameentwicklern erfordert.

Kontextabhängiger und personalisierter Datenjournalismus

Eine Vielzahl von Dienstleister- und Produktdaten werden von Privatpersonen durch *Ubiquitous Computing* (dt. „Rechnerallgegenwart“) wie die Nutzung mobiler, internetfähiger Endgeräte generiert [vgl. Dumbill: 2012, S. 3f.]. Überall werden analoge, aber vor allem weitaus mehr digitale Spuren hinterlassen. Beispiele sind Online-Einkäufe, Bonuskarten, Gewinnspiele, Kartenbezahlung, sogenannte Wearables wie Fitnessarmbänder, Social Media, Navigation oder Gaming. Je exakter die Analyse und Auswertung, umso wirtschaftlicher und vorteilhafter kann darauf reagiert werden. Und oft braucht es nicht viele Informationen. Selbst Metadaten, also Informationen zur Beschreibung von Daten, wie beispielsweise Name, Datum, Nummer, Zeitdauer etc. geben bereits genügend Aufschluss über die Hintergründe, Inhalt und Kontext wie Alltagsabläufe, Religion oder Netzwerke. Sie sind sehr leicht zu analysieren und Edward Felten, IT-Professor in Princeton, bezeichnet Metadaten in seiner schriftlichen Stellungnahme zur Observierung durch künstliche Intelligenz daher als „Stellvertreter für Content“. So sind die Meta-Informationen über Datum, Länge, Empfänger und Anschlussverbindungen über verdächtige Gesprächsverbindungen für die NSA interessanter als das Gespräch selbst, das eventuell durch Akzente, verschiedene Sprachen, Codewörter oder eine schlechte Verbindung schwer analysierbar ist [vgl. Felten: 2013]. Diese Daten dienen daher auch den Datenjournalisten als wichtige Grundlage für mögliche Narrationen.

Jede Anwendung kann im Internet nachverfolgt werden. Die Technologie dazu wird Tracking genannt und ermöglicht die Datengenerierung meistens über die IP in Kombination mit Informationen zum genutzten Browser und Plugins. Eine Wiedererkennung des Nutzers erfolgt beispielsweise durch eine temporär gespeicherte Datei, auch Cookie genannt, die bei Aufruf einer Seite auf dem Computer abgelegt wird.

Bei Aufruf einer Seite können die Daten von Verweildauer, Klickverhalten, Uhrzeit, Inhalte und Standort protokolliert und nach statistischen Verfahren ausgewertet werden, um Rückschlüsse auf Interessen, Vorlieben und weitere Eigenschaften zu schließen. Besonders internetfähige Mobiltelefone geben umfangreiche Datenauskünfte, da für die Nutzung der installierten Dienste (Apps) der Zugriff auf beispielsweise Standortdaten, Kontakte, Mikrophon oder Kamera notwendig ist.

Lorenz Matzat [2015] stellt zu der bisherigen Nutzung dieser Daten im Journalismus fest: „Die W-Fragen, die Nachrichten permanent beantworten, werden von keinem Medium systematisch gesammelt, um sie dann auszuwerten und wieder journalistisch zu verwerten. Wenn Daten im Journalismus heute wirklich eine Rolle spielen, sind es Userdaten, die jeden Klick, jede Regung der Maus oder auf dem Touchscreen registrieren.“

Die Analyse von Internetverlaufsprotokollen, reinen Nutzungs- und Geodaten wird aufschlussreicher, wenn es Informationen über die Zusammenhänge des Anwenderverhaltens gibt. Kontextdaten bekommen daher eine wachsende Bedeutung und es folgt die nächste Entwicklung des technologischen Jahrhunderts: Das Internet of Things. Es bezeichnet das Ablösen von großen

stationären Computern durch kleine integrierbare Minicomputer und Sensoren, die unsere Geräte intelligenter machen (*engl. smart thing*).[25] Sie werden in immer kürzeren Zeitspannen kleiner, schneller, leistungsstärker und günstiger. Vergangene Barrieren wie mangelnde und kostenintensive Speicherkapazitäten spielen heute nur noch selten eine Rolle. Geräte kommunizieren untereinander, vernetzen sich und speichern alle möglichen Informationen zu Zeit, Ort, Dauer, Person, persönlichem Befinden und Nutzung ab. Daten, die bisher aus durchschnittlich mehr als 108 Minuten täglicher Internetnutzung stammen (vgl. Ard, Zdf: 2015), können nun durch zahlreiche neue Datenquellen in unterschiedliche Zusammenhänge gebracht werden.

> ***This behavioural data shows us what people really do, and how it differs from what they say."***
>
> *Designer Matt Cooper-Wright [2015]*

Zukünftig ist es denkbar, dass Kaffeemaschinen selbständig Kaffee nachkaufen, wenn der Füllstand leer ist. Autos werden selbst fahren, wenn sie erkannt haben, dass der Fahrer zu müde ist, und Zahnbürsten erinnern den Benutzer an den nächsten Zahnarzttermin. Teilweise finden diese Beispiele bereits ihre Anwendungen in der Praxis. Es steht die schnell und unkomplizierte Datenübertragung in keinem Verhältnis zur Anwesenheitspflicht, die beispielsweise für das Ablesen des Gasverbrauches einmal nötig war.

Die erzeugten Daten sind nicht nur für den Anwender nützlich und kostensparend. Sie interessieren neben den Konsumenten auch Hersteller, Versicherer, Werber, Krankenkassen oder eben Journalisten. Es lassen sich Vorhersagen treffen, Produkte optimieren, Vermarktung anpassen oder interessante Daten-Geschichten entdecken. Es sollten daher die unterschiedlichen Beweggründe zur Datensammlung der verschiedenen Interessengruppen stets zusammenhängend betrachtet werden.

Verschiedene interdisziplinäre Forschungsfelder der künstlichen Intelligenz (KI), Robotik und Human Computer Interaction beschäftigen sich mit den dazu benötigten Technologien. So wird derzeit intensiv an der automatischen Gesichts-, Emotions- und Verhaltenserkennung gearbeitet.

Beispiele

Datenjournalisten können heute in ihrer Arbeitsweise auch anders vorgehen und Daten für ihre Geschichten selbst erheben. Es ist durch die technologische Entwicklung einfach geworden, anhand von Sensoren investigativ Messungen aufzustellen und Aussagen zu überprüfen. Der Reporter Bryan Christy [vgl. 2016] von National Geographic benutzte für das Projekt „Tracking Ivory" GPS-Sensoren, die Aufschluss über den Schwarzmarkt mit Elfenbein gaben, indem der gesamte Weg verfolgt wurde.

[25] In diesem Zusammenhang werden auch intelligente Agenten, Roboter oder Bots genannt. Dabei handelt es sich um Computer, deren Software nach gewissen Regeln selbstständig (autonom) ausgeführt wird.

Ein weiteres Beispiel ist die „Sensorenresidenz“[26], ein Projekt und Selbstversuch von OpenDataCity, einem Journalistenteam der Datenfreunde GmbH Berlin. In Kooperation mit dem Spiegel wurde die Wohnung des Geschäftsführers Marco Maas zum Smarthome und mit über 100 Sensoren (u.a. Bewegungsmelder, WLAN-Lampen, Steckdosen) ausgestattet. Die gemessenen ein- und ausgehenden Datenströme geben genaue Informationen über die Bewohner, Tagesabläufe und Angewohnheiten, die von den sogenannten smarten Geräten an die Hersteller gesendet werden. Sie kommunizieren sogar im Standby-Modus [vgl. OpenDataCity: 2015].

Zurzeit arbeiten die Datenjournalisten aus Berlin an einem ebenfalls auf Kontextdaten basierenden Projekt „xMinutes“[27] Ziel ist es, personalisierte Nachrichten durch Kuratierung von relevanten Informationen anzubieten, die am richtigen Ort, zum richtigen Zeitpunkt, im richtigen Medium bzw. Endgerät ausgespielt werden. Dazu werden die Kontextdaten aus den Sensoren der Endgeräte für die individuellen Nutzerprofile ausgewertet. Je mehr der Anwender bereit ist, Daten preiszugeben, um sich analysieren zu lassen, desto erfolgreicher können lernende Algorithmen die Medieninhalte abstimmen. Von den detaillierten Nutzerdaten profitieren neben den Anwendern, die optimierte, für sie relevante Nachrichten erhalten, auch die Medienanbieter und Vermarkter. Sie erhalten erweiterte Kenntnisse über das Interesse und Verhalten der Zielgruppe und können ihre Inhalte dementsprechend anpassen.

> ***„Context is where subjectivity enters into data interpretation.“***
>
> *Jonathan Stray [2016, S. 37]*

Bevor die traditionellen Medien aktuelle Nachrichten veröffentlichen, sind sie oftmals bereits über soziale Netzwerke von Leuten vor Ort verbreitet worden. Diese Berichterstattungen haben eine große, exponentiell wachsende Reichweite, weil sie, oberflächlich betrachtet, vertrauensvoll im individuellen Netzwerk ausgetauscht werden und daher sehr authentisch wirken. Die Medienhäuser haben den Auftrag, immer aktuell und informiert zu sein. Es ist zukünftig von großer Bedeutung, Lösungen zu finden, sich von der wachsenden Konkurrenz, ob professionell oder nicht, durch Reichweite und Qualität abzusetzen.

> ***„Heute sind es nur die Timelines der sozialen Netze, in denen wir mit unseren journalistischen Inhalten interessanter als die Schnappschüsse von Freunden und Bekannten sein müssen, um unser Publikum zu erreichen. Künftig werden wir mit Statusmeldungen der Waschmaschine, der Heizungsanlage, mit der Verspätungsmeldung aus dem Nahverkehr oder dem Benzinpreis-Alarm auf dem Weg zur Arbeit um die Aufmerksamkeit unserer Leser konkurrieren.“***
>
> *Marco Maas, Geschäftsführer Datenfreunde GmbH [2015, S.46]*

[26] Datenvisualisierung zum Projekt Sensorenresidenz unter: https://labs.opendatacity.de/sensorenresidenz/
[27] http://xminutes.net/

Die Zukunft liegt in der größtmöglichen Vernetzung und das Internet of Things steht für eine Digitalisierung der physischen Welt. Es wird durch „SmartCity"-Konzepte im Bereich der Stadtplanung genauso wie im Gesundheitsbereich durch die „Quantified-Self"-Bewegung umgesetzt. Big Data ist die Arbeitsgrundlage für Datenwissenschaftler und Datenjournalisten und wird durch die wachsende Anzahl von Quellen stetig größer.

Roboterjournalismus

Derzeit wächst die Angst vieler Journalisten, durch die Möglichkeiten der automatisierten Texterstellung, Roboterjournalismus genannt, überflüssig zu werden. Es handelt sich meist um kurze Berichte, die nach einem einheitlichen Muster und unveränderter journalistischer Form wiederkehren. Frühwarnsysteme, Sportprotokolle, Wetterstatistiken, Polizei- und Verkehrsmeldungen werden bereits vermehrt durch sogenannte Newsbots automatisiert hergestellt. Nach Meinung von Datenjournalist Lorenz Matzat [vgl. 2010] ist die Sportberichterstattung ein gutes Beispiel für Roboterjournalismus und mögliche Automatisierung von Textteilen, da es sich um eindeutig definierte Spielregeln handelt. Doch Matthias Müller, Ressortleiter BILD-Sport Nord beschreibt die ersten Praxistests in seiner Redaktion kritisch [2016]: „Das Ergebnis unserer Versuche mit automatisierter Spielberichterstattung war nichts anderes als Statistiken in Textform verfasst. Es war ohne Hintergründe, beispielsweise Details zu einem Spieler, der mit einer nicht auskurierten Verletzung spielt. Das ist nicht der journalistische Anspruch, mal abgesehen davon, dass reine Spielberichte so gut wie gar nicht mehr gelesen werden."

Vor allem größere Geschichten brauchen weiterhin emotionale Komponenten, die den Algorithmen fehlen. Investigativer Journalismus ist ohne die menschlichen Fähigkeiten, Informationen über den Kontext zu überblicken sowie verständlich zu erklären und zu visualisieren, nicht möglich [vgl. Howard: 2014, S. 30 ff.]. Auch die Ökonomen Frey und Osborne [vgl. 2013: S. 25ff, S. 45, S. 57ff] prognostizieren, dass zukünftig soziale und kreative Fähigkeiten eine große Rolle spielen werden und durch Maschinen nicht ersetzbar sind. Die Auswertung ihrer Studie von über 700 untersuchten Berufen ergab, dass die Arbeitsplätze von Autoren (Platz 123) und Reportern (177) von der Automatisierung weniger bedroht sind als beispielsweise von Köchen (641) oder von Programmierern (293).

Veröffentlichung der reinen Datensätze

Eine Erzählung kann auch die reine Übermittlung recherchierter Datensätze sein. Durch die Sortierung, Strukturierung, Auflistung und Bereitstellung von recherchierten Daten werden Narrationen meistens erst verständlich. In einem geordneten Format lassen sich Muster und Zusammenhänge erkennen, Vergleiche ziehen und Verhältnisse einordnen. Ein Beispiel könnte die Auflistung und dazugehörige Statistiken verschiedener Gehälter und Berufe im Vergleich sein.

Predictive Journalism

Als neuer Trend und Teilgebiet des Datenjournalismus wird der Vorhersage-Journalismus (*engl. predictive journalism*) prognostiziert. Datenanalysen werden zunehmend die Themen und Schlagzeilen von morgen generieren können. Laut Higinio Mayocotte, ehemaliger technischer Leiter vom Texas Tribune und CEO der Datenanalysefirma Umbel [vgl. Maycotte, 2016], werden Naturereignisse und menschliches Verhalten immer besser berechenbar, so dass Abläufe prognostiziert und selbst die Kriegsberichterstattung vorausschauend möglich sein wird. Außerdem lässt sich das Nutzerverhalten besser auswerten und wird durch verbesserte Datenerhebung beispielsweise aus Kontextdaten, wie bereits beschrieben, prognostizierbar. Eine Nachricht, die aus Daten analysiert wurde, könnte nach eigener Darstellung wie folgt aussehen:

> **Erhöhter Terrorverdacht**
>
> Verschiedene radikale Konversationen in sozialen Medien und statistische Berechnungen deuten auf einen geplanten Terroranschlag in Deutschland hin. Öffentliche Daten über Vorkehrungsmaßnahmen zur nationalen Sicherheit und weitere Analysemethoden zeigen ein großes Defizit an Sicherheitspersonal. Derzeit fehlen rund 10.000 Einsatzkräfte.

Die Aussage des Stratfor-Gründers George Friedmann „Journalisten erklären, was in der Welt passiert und Stratfor[28] erklärt, was passieren wird“ sollte als Kampfansage bzw. Weckruf verstanden werden [vgl. Friedmann: 2012]. Das Geschäft von Stratfor und vergleichbaren Firmen wie Statista aus Hamburg ist einfach beschrieben: Informationen aus einem guten Informantennetzwerk werden zusammen mit öffentlichen und privaten Massendaten ausgewertet. Dadurch erhalten Ereignisse einen Kontext, auf dessen Grundlage Stratfor glaubwürdige Prognosen erstellt. Ziel und Aufgabe des Journalismus sollte es zukünftig auch sein, vielfältige und nachhaltigere Prognosen aufgrund von Massendaten geben zu können.

[28] Die private US-Strategieberatungsfirma Stratfor, oft auch als „Schatten-CIA“ bezeichnet, berät erfolgreich internationale Handels- und Wirtschaftsunternehmen sowie Regierungen und gibt geopolitische Prognosen mittels Datamining und anderen Softwareanwendungen aus dem Bereich der künstlichen Intelligenz ab. Überwiegend handelt es sich um Informationen aus einem großen Informantennetzwerk und öffentlich zugänglichen Quellen, die kompiliert und gründlich ausgewertet werden. Die Prognosen wie zum arabischen Frühling zeichnen sich bisher durch eine hohe Verlässlichkeit [https://www.stratfor.com/].

Dynamische Dateneinbindung

Es ist möglich Geschichten protokollarisch zu erzählen, indem Echtzeitdaten wie Kommentare oder Ergebnisse eingebunden werden. Das geschieht zum Beispiel oft bei Wahlen oder Sportberichterstattungen, wenn sich die Erzählungen in kurzen Zeitabständen verändern und die Ereignisse relevant sind. Nach Lorenz Matzat [vgl. 2011] bedeutet Datenjournalismus in diesem Fall das gleichzeitige Sammeln von Daten und deren Aufbereitung. Durch fortschreitende Technologien zur Bild- und Gesichtserkennung, Geolokalisation und Kontextdatenanalyse können Anwendern in Echtzeit aktuell und personalisiert Zusatzinformationen gegeben werden. Mit entsprechenden Ausspielgeräten wie dem Smartphone oder einer Augmented-Reality-Brille können sich Anwender beispielsweise aktuelle und zum Standort passende Verkehrsstörungen oder freie Immobilien anzeigen lassen.

Datastorytelling

Hier kann der Anwender ganz individuell nach Interesse mit Daten experimentieren und somit die Geschichte personalisiert erfahren. Die Basis der Erzählung ist meistens ein Datensatz, der durch eine interaktive Visualisierung dargestellt wird. Viele unterschiedliche multimediale Inhalte können als Zusatzinformationen eingebunden werden und führen zu einem ganzheitlichen Storytelling auf Datengrundlage.

Crowdsourcing

> ***"[...] Wenn Maschinen nicht mehr weiterkommen, schlägt die Stunde des Crowdsourcing. Es wird auf die Effizienz der Menge gesetzt. [...]"***
>
> *Datenjournalist Lorenz Matzat [2011]*

Wie im MP-Expenses-Projekt vom Guardian bereits beschrieben, kann im Journalismus auch nach dem Wikipedia-Prinzip gearbeitet werden. Sollte die Rechercheleistung die eigenen Personalkapazitäten übersteigen, kann eine Crowdsourcing-Kampagne die Lösung sein, um mehr Quellen zu bekommen. Die freiwillige Mitarbeit der breiten Masse kann dabei zum einen der reinen Beschaffung von Informationen dienen. Zum anderen kann es auch ein Aufruf zur Teilnahme an einer Geschichte sein. Das Prinzip heißt Arbeitsteilung: Mehr Leute sehen und schaffen einfach mehr. Zu teils stupider Buchhaltertätigkeit lassen sich viele Helfer bereits durch die namentliche Nennung, Ranglisten, Wettbewerbe oder Punktesammeln motivieren. Das Prinzip wird Gamification genannt und bringt spieltypische Elemente in spielfremde Zusammenhänge, um eine Verhaltensänderung und Motivationssteigerung zu erreichen [vgl. Gabler: 2015].

Gaming

In dieser Form kann Datenjournalismus Bestandteil von Newsgames werden, die versuchen, Nachrichten und Game-Interaktivität im Journalismus zu verbinden**.** Markus Bösch [2013], Journalist und Newsgame-Pionier aus Deutschland, erklärt das Prinzip wie folgt: „Newsgames, also Spiele, die im journalistischen Kontext Verwendung finden und bei deren Erstellung journalistisch-ethische Grundregeln eingehalten werden, ermöglichen im Unterschied zu traditionellen linearen Medien die interaktive Erfahrung von Inhalten."

The New York Times hat 2010 mit „Budget Puzzle" ein beispielhaftes datenjournalistisches Newsgame umgesetzt. Die Leser konnten Vorschläge zu Budgetkürzungen des Staatshaushaltes abgeben und mit den Datensätzen experimentieren, die aus bestehenden und prognostizierten Daten verschiedener öffentlicher Einrichtungen stammten [vgl. Charter et al.: 2010]. Ein weiteres Beispiel ist das von Arte entwickelte Online Newsgame „Steuerflucht für Anfänger". Spieler versuchen darin, über verschiedene Möglichkeiten Geld am Fiskus vorbei zu schleusen und lernen nebenbei das komplexe Thema spielerisch kennen. Laut Lorenz Matzat [vgl. Gray et al.: 2012, S. 51] können Datenjournalisten besonders von der Game-Industrie lernen, die erfolgreich digitale Erzählungen, Infrastrukturen und Interfaces gestaltet. Eine Vertiefung über Crowdsourcing, Gamification und Begründung einer wachsenden Bedeutung von Gaming ist in diesem Zusammenhang in einer separaten Ausarbeitung einzusehen [vgl. Philipsen: 2016].

3 Ablauf journalistischer Datenverarbeitung

3.1 Einleitung

> ***„'We are living in the information age' is a popular saying; however, we are actually living in the data age. Terabytes or petabytes of data pour into our computer networks, the World Wide Web (WWW), and various data storage devices every day from business, society, science and engineering, medicine and almost every other aspect of daily life."***
>
> *[Han et al.: 2012, S. 1]*

Ob neueste Berufsbezeichnungen wie Business Intelligence, Data Analyst, Data Scientist oder eben Data Journalist: Alle Branchen versuchen sich aktuell Massendaten zu Nutze zu machen, sie zu analysieren und Wissen zu extrahieren. Dabei wird auf gleiche Techniken zurückgegriffen: Data Mining (*engl., Datenschürfen*) bezeichnet die Anwendung statistischer Methoden, Techniken und Algorithmen auf große Datenbestände, die selbständig neue Muster und Zusammenhänge entdecken und auswerten. Dabei können verschiedenste Muster aus Geschmack, Sprache, Produkten, Einkäufen, Nutzung, Bildern, Texten, Verhalten, Problemen oder Prozessen von Interesse sein.

Auf Grundlage mathematischen Verständnisses vereint Data Mining Elemente von Statistik, Künstlicher Intelligenz wie maschinellem Lernen[29], Datenbanken und Visualisierung/Computergrafik [vgl. Cleve, Lämmel: 2016, S. 2-3, 12-13]. Durch Wissensgewinnung, Wissensverwaltung und Wissensverarbeitung können produktions-, vertriebs- oder marketingrelevante Erkenntnisse gewonnen und umgesetzt werden. Der Gesamtprozess der Datenanalyse wird Knowledge Discovery in Databases (KDD) genannt [vgl. Fayyad et al.: 1996a]. Obwohl Data Mining (die Suche nach Mustern) genau genommen ein Zwischenschritt bzw. Teil des KDD-Prozesses ist, wird der Begriff oft mit KDD gleichgesetzt.

Neben dem KDD-Modell beschreibt auch das CRIPS-Modell (*Cross Industry Standard Process for Data Mining*) sowie das SEMMA-Modell von SAP (*Sample, Explore, Modify, Model and Assess*) den Prozess des Data Mining. Diese Modelle wurden jeweils von verschiedenen Wirtschaftsunternehmen aus Sicht

[29] Maschinelles Lernen (*engl. Machine Learning*) bezeichnet ein computerbasiertes Lernverfahren, das selbstständig Wissen aufnimmt und durch Erfahrung, neues Wissen generiert. Dabei kann das erlernte Wissen dieser intelligenten Systeme nach dem selbstständigen Lernprozess auf unbekannte Datensätze angewendet werden. Machine Learning ist ein Teilgebiet der künstlichen Intelligenz. Ziel ist das Erkennen von Mustern, Hintergründen, Zusammenhängen und Anhängigkeiten [vgl. Cleve, Lämmel: 2016, S. 14].

der Industrie für eigene Anwendungen entwickelt. Cleve und Lämmel [vgl. 2016, S. 5-6] geben einen Überblick über den gesamten KDD-Prozess und gliedern den Ablauf in einzelne Phasen:

1. **Selektion**
   Alle verfügbaren Daten werden gesichtet, Zieldaten definiert, von Datenquellen exportiert und in einer Datenbank erfasst.
2. **Datenvorbereitung**
   Alle Daten werden nach Fehlern geprüft und bei Bedarf bereinigt, korrigiert, ergänzt. Es ist im Durchschnitt mit einer ca. 5% großen Fehlerquote eines Datensatzes zu rechnen.
3. **Transformation**
   Die Daten werden formatiert und gruppiert.
4. **Data Mining**
   Durchsuchung der Datenbestände nach Mustern.
5. **Interpretation und Evaluation**
   Interpretation, Prüfung und Auswertung der Ergebnisse.

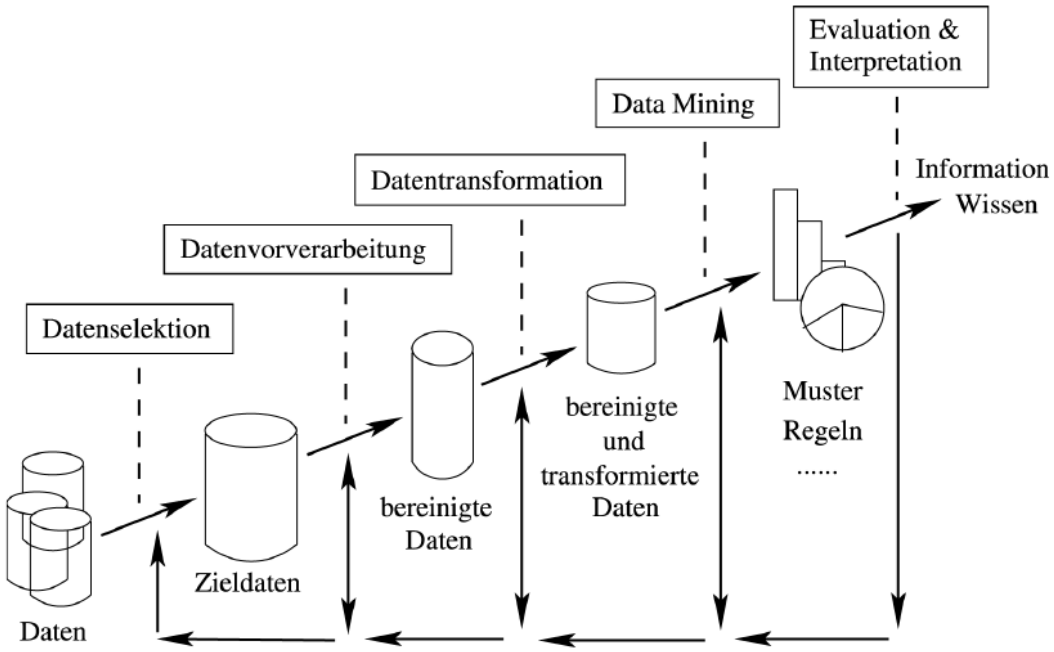


Abbildung 3: Ablauf des KDD-Prozesses nach Fayyad et al. *[vgl. 1996b]*

Die beschriebenen Schritte des KDD-Prozesses zu Datensammlung, Analyse- und Auswertungsmethoden zeigen, wie Datenanalysten der Reihenfolge nach vorgehen. Dabei können einzelne Schritte beliebig wiederholt und angepasst bzw. korrigiert werden, wenn beispielsweise die

Analyse zu keinem Ergebnis geführt hat. Die Einhaltung der Schrittreihenfolge ist dabei wichtig, allerdings ist die Anwendung unterschiedlicher Verfahren nicht zwingend.

Das folgende Kapitel geht detailliert auf den Ablauf des KDD-Prozesses ein, der die Grundlage für die Analyse von Massendaten im Zusammenhang mit datenjournalistischer Arbeit bildet. Beginnend bei der Datenrecherche und -sammlung wird der folgende Prozess der Daten- und Textanalyse beschrieben. Außerdem werden Möglichkeiten der anschließenden Präsentation aufgezeigt sowie Beispiele und Softwareanwendungen aus dem beruflichen Alltag gegeben.

3.2 Daten – Definition und Abgrenzung

Zunächst werden zum Verständnis die Begriffe *Zeichen, Daten, Nachrichten, Wissen* und *Information* beschrieben, die dem Grundverständnis dieser Arbeit dienen.

Die Begriffe stehen aufgrund ihrer Syntax[30] und Semantik[31] sowie pragmatischen Ebene[32] in hierarchischer Form zueinander. *Zeichen* bilden als Grundelemente demnach die kleinste Darstellungseinheit von Informationen. Werden diese in Beziehung zueinander gesetzt, spricht man von *Daten*[33] [vgl. Petersohn: 2005, S.4-5]. Nach Cleve und Lämmel [vgl. 2016, S. 37-38] sind daher Daten als eine „Ansammlung von Zeichen mit der dazugehörigen Syntax“ anzusehen. Sie unterscheiden zwischen drei Arten von Daten, die für die Datenanalyse entscheidend sind:

- **Unstrukturierte Daten** (Kapitel 3.11 Text Minig)
  (Bilder und Texte, deren maschinelle Interpretation aufwendiger ist, da erst eine Umwandlung in strukturierte Daten vorausgehen muss.)
- **Semistrukturierte Daten**
  (Beispielsweise Websites, die trotz unstrukturierten Dateninhalts eine Struktur aufweisen und sich somit automatisiert auslesen lassen.)
- **Strukturierte Daten** (Kapitel 3.9 Data Mining)
  (Datensätze von Datenbanken, die aufgrund ähnlicher Dateiformate, fester Datenreihenfolge, Attributen und Dateityp in einheitlichen Tabellen strukturiert werden können und so eine leichte Sortierung und Weiterverarbeitung ermöglichen.)

[30] Als Syntax wird das Regelwerk bezeichnet, dass die Zusammensetzung von Zeichen für eine Sprache bestimmt.
[31] Die Semantik definiert die Beziehungen von Zeichen und Bedeutung einer Sprache.
[32] Die pragmatische Ebene beschreibt den Mehrwert der Information für den Empfänger.
[33] Der Begriff Daten bedeutet zum einem die Mehrzahl von Datum, einer Zeit- und Kalenderangabe im Sprachgebrauch. Im Zusammenhang mit der Erklärung der oben genannten Begriffe werden Daten aber als Informationseinheit definiert.

Der Hierarchie weiter folgt die *Nachricht*. Sie bezeichnet die vom Sender übertragenen Daten an den Empfänger. Bei Daten handelt es sich um reine Fakten, die erst zu einer *Information* werden, sobald sie eine Bedeutung erhalten. Es folgt zuletzt das *Wissen*, das sich aus der menschlichen Fähigkeit zur Anwendung und Verarbeitung von Informationen ergibt [vgl. op. cit.: S.7].

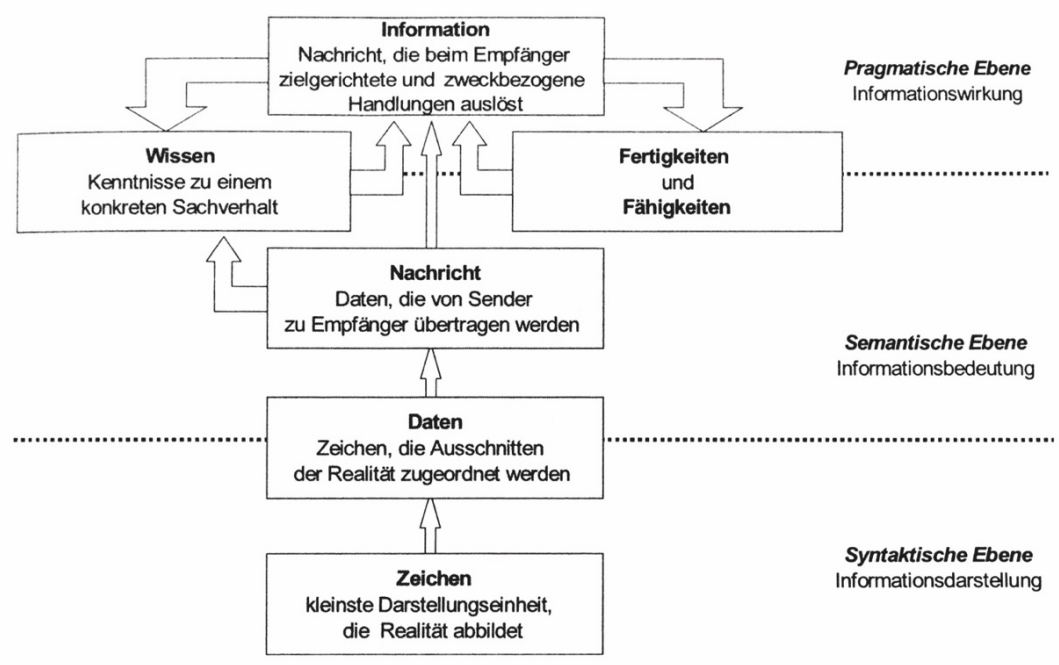


Abbildung 4: Semiotische Betrachtungsebenen des Informationsbegriffs nach Petersohn [2005, S.7]

3.3 Datentypen

Daten haben unterschiedlich viele Merkmale, auch Dimensionen genannt, und unterscheiden sich in der Struktur. Nach Jürgen Cleve und Uwe Lämmel können Daten aus Zahlen-, Ordnungs- und Rechenmerkmalen bestehen [vgl. 2016, S. 39-40]. Es entstehen unterschiedliche Herausforderungen an die Weiterverarbeitung, da beispielsweise im Falle von Farbmerkmalen wie grün und rot zunächst keine Berechnung erfolgen kann. Daher wird im ersten Schritt nach den Kriterien der Ordnung und des Rechnens unterschieden:

- **Nominale Daten**
  Die Daten sind ohne Reihenfolge, nur vergleichbar (z.B. Geschlecht, Augenfarbe).
- **Ordinale Daten**
  Es gibt zwar eine feste Ordnungsrelation, aber ohne Rechenmöglichkeit (z.B. Schulnoten 1-5 oder Abstufungen wie schlecht, durchschnittlich, gut, sehr gut).

- **Metrische Daten**
  Mit Ordnungsmerkmalen und Rechenmöglichkeit, es wird zwischen Intervallskala (Jahreszahlen, Temperatur), Verhältnisskala (Messwerte wie Strom, Entfernung, Körpergröße) und Absolutskala (Lebensjahre, Kinderanzahl) entschieden.

3.4 Daten- und Informationsqualität

Die Beurteilung von Daten und Informationen basiert auf den Forschungsergebnissen von Richard Y. Wang und Diane M. Strong [vgl. 1996: S.5-33]. Die Wissenschaftler haben 1996 nach vier Kriterien 15 Dimensionen entwickelt, die heute als Standard gelten. Ziel ist es, die Informationsqualität auf Grundlage von Verlässlichkeit, Relevanz und Richtigkeit zu definieren. Die Qualitätsansprüche beziehen sich auf Systemanforderung, Darstellungsbeurteilung, Nutzungseigenschaften und Inhalte. Die folgende Grafik zeigt anhand von vier Schwerpunkten (Kategorien) und 15 Dimensionen nach welchen Kriterien die Qualität gemessen wird:

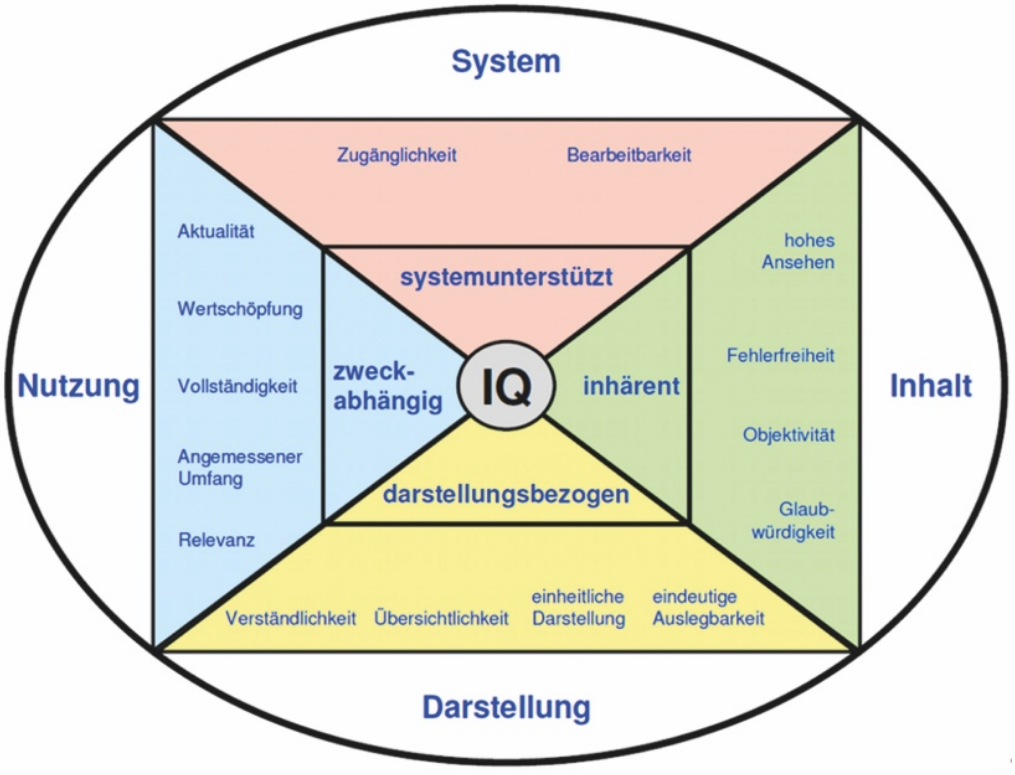


Abbildung 5: Informationsqualität [vgl. Dgiq: 2007]

3.5 Datenrecherche

> *"Actually the context comes before the data; it tells us what data is relevant, even what questions are relevant."*
>
> *Jonathan Stray [2016, S.37]*

Journalisten sollten für ihre Datenrecherche zuerst die Quellen im Hinterkopf haben, die bekanntermaßen allein aufgrund ihrer Aufgabe automatisch über umfangreiche und aussagekräftige Datensätze verfügen. Dazu zählen zum Beispiel alle öffentliche Einrichtungen und Personalabteilungen sowie viele Dienstleistungen [vgl. Egawhary et. al.: 2012, S. 4]. Im Folgenden werden neben der journalistisch tradierten Recherche zunächst verschiedene Vorgehensweisen zum Auffinden von Massendaten aus der Praxis beschrieben, die auf den Empfehlungen von Brian Boyer (Chicago Tribune), John Keefe (WNYC), Friedrich Lindenberg (Open Knowledge Foundation), Jane Park (Creative Commons) und Chrys Wu (Hacks/Hackers) in *The Data Journalism Handbook* zusammengestellt wurden [vgl. Gray et al.: 2012, S.96-100].

Da über das Internet zugängliche Datenbanken durch Google oder andere Suchmaschinen[34] indiziert und auffindbar sind, ist es möglich, neben den Inhalten auch nach Formaten zu suchen. Die Suche könnte beispielsweise nach diesen derzeit gängigen Datentypen erfolgen:

- **.xls** (Microsoft Exceldatei für die Suche nach Tabellen.)
- **.csv** (Comma-Separated Values, strukturierte Textdatei für die Speicherung von Listen oder Tabellen.)
- **.shp** (Shapefile, Format für Geodaten.)
- **.MDB** (Microsoft Access Database, Datenbankformat von Microsoft.)
- **.SQL** (Structured Query Language, der Dateityp enthält ein Abbild (*engl. Dump*) einer relationalen Datenbank.)
- **.DB** (General Database, allgemeines Datenbankformat.)
- **.Pdf** (Portable Document Format von Adobe Systems. Format für eine geräteunabhängige Darstellung von Dokumenten, unabhängig von der Auflösung.)[35]

[34] In diesem Zusammenhang wird auf die semantische Suchmaschine Wolfram Alpha hingewiesen. Der Ansatz von Wolfram|Alpha ist das Auffinden von inhaltlichen Antworten auf Suchanfragen, im Gegensatz zur üblichen Vorgehensweise der Suchmaschinen nach relevanten Inhalten.

[35] Filesuffix.com ist eine Datenbank für Dateinamenerweiterungen. Online unter: https://www.filesuffix.com/de/ (abgerufen am 9. August 2016)

Google bietet weitere Recherchemöglichkeiten mithilfe von Operatoren an. So ist auch die direkte Suche in der URL möglich. Eine downloadbare Excel-Tabelle lässt sich unter *inurl:downloads filetype:xls* suchen (z.B. „inurl:Hamburg filetype:xls“ filtert nach allen URLs, die Hamburg enthalten haben und den Dateitypen .xls). *Site:hamburg.de Kitaplätze* ermöglichen eine eingeengte Suche nach dem Begriff Kitaplätze auf der Seite www.hamburg.de. Um Orte zu finden, an denen Massenrohdaten abgelegt wurden, kann beispielsweise unter *site:“URL-Adresse“ directory listing* gesucht werden. Bewusst abgelegte Daten sind unter *site: “URL-Adresse“ database download* zu finden.

Weitere Recherchemöglichkeiten bieten zahlreiche Open-Data-Portale, Foren und Netzwerke für Datenjournalisten. So sind Datenjournalisten und CAR-Spezialisten in der Data Driven Journalism List und Nicar-L List von IRE (*Investigative Reporters and Editors*) aufgelistet. *Getthedata.org* (Auskunftsdienst zu datenrelevanten Fragen) oder *Quora.com* (allgemeiner digitaler Auskunftsdienst) sind beispielsweise Online-Portale, die unter anderem auf Fragen zur Beschaffung von Daten spezialisiert sind und mit Experten zusammenarbeiten. In Zusammenarbeit mit Newsroom-Entwicklern der Chicago Tribune entstand eine Datenbibliothek für Datenjournalisten (PANDA-Projekt), die der Verwaltung und dem internationalen Austausch von Journalisten dient. Und als spezialisierte Austauschplattform zwischen Journalisten und Technologen gilt nach Gray et al. [vgl. 2012, S.96-100] das Netzwerk Hacks/Hackers.

Es gibt zudem weitere zahlreiche praktische Anwendungen und Dienste, auf die Journalisten zurückgreifen können. So gibt „Bit.ly“ beispielsweise Auskünfte über die Verbreitung von Links und „Twitter“ über Konversationen zu Links (z.B. Twitter). Es lassen sich mit der Anwendung von „TinEye“ Fotoquellen recherchieren. „Google Trends“ analysiert Internetsuchtrends, und Homepage-Veränderungen können z.B. mit „The Waybackmachine“ über lange Zeiträume dokumentiert werden. „Whois“ ist eine Abfragemöglichkeit von Internetdomains und IP-Adressen über Name, Adresse, Email, Telefonnummer, die beispielsweise „whois.domaintools.com“ bietet. Außerdem existieren mittlerweile spezielle Plattformen zur Veröffentlichung, Analyse und Austausch von großen Datensätzen wie „The Data Hub“ oder „Document Cloud“.

3.5.1 Datenbeschaffung aus dem Web

Ein Grundverständnis von der Struktur einer Webseite erleichtert die Interpretation von Online-Dokumenten. Viele Informationen lassen sich im sogenannten Quelltext der Seite finden, der durch die weitverbreitete Auszeichnungssprache für Webanwendungen (HTML) generiert wird. Datenjournalisten sollten den Dokumentaufbau in <Head>, <Title>, <Body> und die Hierarchieregeln von sogenannten Boxen verstehen, die aus „Tags“ zur Beschreibung von Bildern oder Tabellen bestehen. Ihnen sollten Gruppen („classes“) und die Grundelemente wie zum Beispiel die Beschreibung einer Tabelle (<table>, <td>, <tr>) bekannt sein. Um gesuchte Elemente und Bezeichnungen im Quelltext schneller auffinden zu können, ist ein Allgemeinwissen über die Einbindung weiterer Dateiformate hilfreich. So werden mit der Stylesheet-Sprache CSS (*Cascading*

Style Sheets) beispielsweise einzelne HTML-Elemente angesprochen und gestaltet. Die Skriptsprache Javascript erzeugt Dynamik in einem HTML-Dokument. Datenjournalisten können diesbezüglich auf diverse kostenlose Online-Tutorials wie w3schools.com oder codecademy.com zurückgreifen. Um Daten aus dem Web zu generieren, beschreibt Friedrich Lindenberg [vgl. Gray et al.: 2012, S.106 ff.] drei Anwendungsfälle:

- **Webbasierte API** (*engl. application programming interface*)
  Dabei handelt es sich um eine Programmschnittstelle von Softwaresystemen, die eine Anbindung anderer Programme ermöglicht, um beispielsweise Zugriff auf Datenbanken zu erhalten. Diese kann generell offen (frei zugänglich) sein, wie bei dem Internet-Social-Media-Dienst Twitter, oder der Zugriff ist eingeschränkt und erfordert Authentifizierung.

- **Screen Scraping Websites für strukturierte Daten (*to scrape, engl. abschürfen*)**
  Mit dieser Anwendung wird der aktuelle Quelltext des Browsers ausgelesen und versucht, zu interpretieren. Dies setzt maschinenlesbare Daten voraus. Es besteht eine besondere Herausforderung im „scrapen" der mittlerweile meistverbreiteten dynamisch erzeugten Webseiten, die ihre Inhalte asynchron laden. Das heißt, dass die Inhalte der Webseite nicht vollständig auf einem Server liegen müssen und sich aus unterschiedlichen Quellen zusammensetzen können. Beispielsweise werden Werbeanzeigen auf Webseiten dynamisch erzeugt oder Teilinhalte wie die stets aktualisierenden Fahrzeiten der Bahn. Erst der Internet-Browser interpretiert die vollständige Seite und setzt die Inhalte in Laufzeit zusammen, ohne die Seite neu laden zu müssen. Man arbeitet in diesem Zusammenhang mit Webanwendungen wie beispielsweise Ajax (*Asynchronous JavaScript and XML*). Screenscraper können diese Seiten auswerten, aber für jede kleinste Änderung der Website muss ein neuer Code geschrieben werden. So ändert Facebook bei gleicher Browser-Darstellung regelmäßig den erzeugenden Code der Webseite, nur um ein Auslesen zu verhindern. Scraper können in unterschiedlichen Sprachen wie Ruby, Python oder PHP gebaut sein. Für die Bestimmung der richtigen Seite und exakten Elemente ist es notwendig, die Struktur der Seite und Datenbank zu verstehen. Einfach auszulesende Formate sind XML, JSON (Dateiformat für einfachen Datenaustausch), Word Dokument, HTML oder Excel. Schlecht formatierter HTML-Code, Authentifizierungssysteme (automatische Verhinderung von Zugriffen durch Paywall oder CAPTCHA[36]) sowie zeitbasierte Systeme (Usertracking durch Cookies) lassen sich dagegen sehr schlecht scrapen[37].

[36] Captcha (*Completely Automated Public Turing test to tell Computers and Humans Apart*) ist ein automatisierter Test, um zu unterscheiden, ob es sich beim Anwender um einen Computer (Roboter) oder Menschen handelt. Meistens muss der Anwender eine für Computer zu schwierige Frage beantworten oder Aufgabe lösen.

[37] Als weiterführende Literatur wird „Scraping for Journalism: A Guide for Collecting Data" von Dan Nguyen sowie der Studie von Jacob Harris „How Data Sausage is made" empfohlen.

- **Daten mit Hilfe von Programmen aus PDF-Formaten extrahieren**
  Das Dateiformat PDF ist ein Druckerformat, das Positionen von Linien und Punkten verarbeitet. Die Differenzierung von Zeichen ist bei diesem Format unwichtig und daher ist die Anwendung sehr aufwendig. Da es sich vor allem um ein visuelles Format handelt, benötigt man eine Texterkennungssoftware (*engl. Optical Character Recognition Software*) wie Abbyy Finereader oder Tabula, die durch optische Zeichenerkennung PDF-Formate in beliebige Textdateien wie Word oder Excel zur Weiterverarbeitung umformatiert.

Im Zusammenhang mit der Datenbeschaffung wird u.a. im *The Data Journalism Handbook* [vgl. Gray et al.: 2012, S. 106 ff.] auf viele bereits kostenlose Softwareanwendungen hingewiesen, die Textinhalte einer Seite sichern können, verschiedene Quellen gleichzeitig automatisch auslesen oder beispielsweise den Download mehrerer Links, Bilder und anderer Inhalte einer Website durch ein Browser-Plugin[38] ermöglichen.

3.6 Selektion

Daten liegen nicht immer automatisch aus einer Datenquelle und in einem Dokument vor. Sie müssen aus verschiedenen Datenbanken und Quellen recherchiert (Selektion) werden und in einer Tabelle zusammengetragen werden (Integration). Bei der Auswahl bestimmter Datensätze wird von vertikaler Selektion und bei der Auswahl von Attributen von horizontaler Selektion gesprochen. Die Selektion ist die Grundlage des Data-Mining-Prozesses. Folgende Probleme können dabei nach Cleve und Lämmel [vgl. 2016, S. 206] durch unterschiedliche Strukturen in Semantik und Syntax der Attribute entstehen:

- **Entitätenidentifikationsproblem**
  (Unterschiedlich ausgezeichnete Attribute mit gleicher Semantik.)
- **Redundanzen**
  (Überflüssige Informationen z. B. durch Dopplungen.)
- **Widersprüche**
  (Gleichausgezeichnete Attribute mit unterschiedlichen Merkmalen.)
- **Datenwertkonflikte**
  (Datensätze haben unterschiedliche Maßeinheiten.)
- **Verletzen der referenziellen Integrität**
  (Fehlende Informationen einer Quelle, die bei der Zusammenführung mehrerer Datensätze fehlt, aber angefordert wird.)

[38] Ein Plugin ist eine Softwareergänzung, die während der Softwareanwendung die Funktionalität erweitern kann.

3.7 Datenvorbereitung

Die Verarbeitung und das Ergebnis von Datenanalysen hängt von der Datenqualität ab. Diese wird an der Verständlichkeit, Nützlichkeit, Gültigkeit, Aktualität, Glaubwürdigkeit, Vollständigkeit, Richtigkeit und Volatilität[39] gemessen [vgl. Cleve, Lämmel: 2016, S. 221]. Besonders zur Wahrung der journalistischen Sorgfaltspflicht sind die Datensätze nach diesen Kriterien zu prüfen. Datenfehler können zum Beispiel Widersprüche bei Angaben zum Alter und Geburtsdatum, Dopplungen, fehlende Einträge oder Schreibfehler sein. Um einheitliche Daten zur Analyse zu erhalten, müssen die Daten vorher gesäubert werden. Durchschnittlich wird mit bis zu 5% technisch oder menschlich verursachten Fehlern operativer Systeme gerechnet [vgl. Redman: 1998, S. 80]. In diesem Schritt der Vorverarbeitung (*engl. Preprocessing*) sollten keine neuen Informationen hinzugefügt werden, um die Ergebnisse nicht zu verfälschen. Folgend werden dazu von Cleve und Lämmel [vgl. 2016, S. 207 ff.] Fehlerquellen und dazugehörige Lösungsansätze für die Säuberung von fehlerhaften Daten aufgezeigt.

Ausreißer und verrauschte Daten

Daten mit leichten Fehlern, die durch Messungenauigkeit und Schätzungen entstehen, nennt man verrauscht. Als Ausreißer werden vom Niveau stark abweichende Werte genannt.

- **Klasseneinteilung (*binning*)**
  (Die verrauschten Daten glättet man durch Gruppierung und anschließender Mittelwert- oder Grenzwertberechnung.)
- **Regression**
  (Die verrauschten Daten werden durch eine mathematische Funktion beschrieben und anschließend ersetzt.)
- **Verbundbildung**
  (Durch Clusterbildung beispielsweise nach dem dichtbasierten Verfahren lassen sich Ausreißer extrahieren.)
- **Kombinierte Mensch/Maschine-Untersuchung**
  (Nach einer automatisch erstellten Liste von vermeintlich falschen Werten, werden die Werte aufgrund von Erfahrungswerten manuell gefiltert.)

Falsche und widersprüchliche Daten

Im Umgang mit falschen und widersprüchlichen Werten bleibt neben einer manuellen Bereinigung und Abgleich mit richtigen Datensätzen letztlich nur das Löschen oder der Verzicht auf den Datensatz. Sollten Dopplungen in Datensätzen vorhanden sein, würde das die gesamte Gewichtung in vielen Verfahren verändern.

[39] Volatilität (*lat. volatilis für fliegend*) meint Schwankungen in Zeitintervallen.

Fehlende Daten (*engl. missing values*)

- **Attribute ignorieren**
  (Attribute werden aus der Tabelle entfernt, verfälscht jedoch das Ergebnis.)
- **Fehlende Werte manuell nachtragen**
  (Anwendung ist meistens zu zeitintensiv.)
- **Globale Konstante**
  (Alle fehlenden Attribute werden einheitlich durch z.B. „unbekannt“ aufgefüllt.)
- **Durchschnittswert**
  (Anwendung bei metrischen Attributen, angelehnt an das k-nearest-Neighbour-Prinzip, das auf Seite 52 näher beschrieben wird.)
- **Wahrscheinlichkeitswert**
  (Ermittlung einer Wertewahrscheinlichkeit mittels statistischen Methoden, nur bei wenigen fehlenden Werten anwendbar.)
- **Häufiger Wert**
  (Bei nichtnumerischen Attributen anwendbar.)
- **Relation zwischen Attributen**
  (Wenn eine Relation zwischen dem unvollständigen und einem vollständigen Attribut gefunden wurde, können mittels Regressionsfunktion oder Assoziationsanalyse Zusammenhänge ermittelt werden.)
- **Datensatz für Analyse ausschließen**
  (Wenn genug andere Datensätze vorhanden sind.)

Datenreduktion

Um den Umfang und die Komplexität der Datenanalyse zu reduzieren, ist es zum einen möglich, eine repräsentative Teilmenge aus den Datensätzen zu entnehmen, Attribute zusammenzufassen oder nicht relevante Attribute auszusortieren. Dazu gibt es vier unterschiedliche Techniken:

- **Aggregation**
  (Generalisierung, Verdichtung und Zusammenfassen von Daten zum Beispiel durch den Mittelwert.)
- **Datenkompression**
  (Verringerung der Attributanzahl und Zusammenfassung von mehreren Attributen.)

- **Dimensionsreduktion**
  (Schrittweise Trennung von wichtigen und irrelevanten Attributen durch die Zuordnung oder umgekehrt die Ausgliederung der Attribute einer Teilmenge. Dafür bietet sich beispielsweise der t-SNE-Algorithmus an, um eine Visualisierung mit einem Streudiagramm zu ermöglichen.)

- **Numerische Datenreduktion**
  (Durch verschiedene Stichproben lassen sich repräsentative Teilmengen erzeugen. Diese können unter anderem zufällig, repräsentativ, selektiv oder clustergestützt gebildet werden. Zudem besteht die Anwendungsmöglichkeit der linearen Regression, wenn die Daten durch Koeffizienten einer linearen Regression ersetzt werden.

3.8 Datentransformation

Im Transformationsschritt müssen die Daten an das ausgewählte Datamining-Verfahren angepasst werden. Die Art der Transformation ist somit abhängig von der Analyse.

In diesem KKD-Schritt werden die gesammelten und bereinigten Datensätze formatiert und in ein Datenbankschema integriert. Häufig ist eine Anpassung der Datentypen, Codierungen, Zeichenketten, Datumsangaben, Maßeinheiten oder Skalierungen nötig. Darüber hinaus können dazu die vorher beschriebenen einzelnen Schritte der Datensäuberung erneut ihre Anwendung finden. Es wird in diesem Schritt zum Beispiel entschieden, ob die Zahl 4 vom Typ „character“ oder „integer“[40] ist, sowie die Wochentagbeschreibung durch Zahlen 1-7 oder Zeichen erfolgt. So können einige Verfahren keine metrischen Attribute verarbeiten.

[40] Datentypen unterschiedlicher Programmiersprachen.

3.9 Data Mining

„Exploratory data analysis is detective work.“

John W. Turkey [1977]

Es gibt unterschiedliche Analyseverfahren, die Datensätze untersuchen. Folgend werden verschiedene Ansätze der Analyse beschrieben, die auch kombiniert werden können. Der Schwerpunkt dieser Arbeit liegt auf der Datenanalyse mittels Klassifizierung, Klassenbildung und Datenvisualisierung.

3.9.1 Klassifizierung

Für die Klassifizierung von Daten gibt es unterschiedliche Modelle, die Wissen generieren und repräsentieren können. Zum einen gibt es erklärungsfähige Modelle, die vom Menschen lesbar und abbildbar sind. Dazu zählen Graphen, Tabellen[41] und logische Regeln. Davon zu unterscheiden sind die statistischen, wahrscheinlichkeitstheoretischen und simulationsgestützten Performance-Modelle wie neuronale Netze, deren Ergebnisse nicht mehr erklärbar sind. Sie stellen eine implizite Wissensdarstellung dar, während die explizite Wissensdarstellung mittels Logik und Regeln erfolgt.

Ausgehend von bereits bekannten und klassifizierten Datensätzen können neue und unbekannte Datensätze nach zwei methodischen Verfahren klassifiziert werden. Entweder die Klassifikation erfolgt mithilfe des instanzenbasierten[42] Verfahrens durch vorhandene Beispieldatensätze oder aber nach einem Modell, das auf Basis von Beispieldatensätzen berechnet wurde. Es handelt sich um überwachtes Lernen, da eine Grundannahme vorausgesetzt wird. Im Folgenden werden nach Cleve und Lämmel [vgl. 2016] Beispiele für unterschiedliche Klassifikationsverfahren kurz vorgestellt.

TDIDT-Verfahren (*engl., Top down induction of decision trees*)
Bei dem methodischen Vorgehen mittels Entscheidungsbäumen (*Decision Trees*)[43] wird an jedem Knoten ein Attribut (univariate Bäume) und bei multivariaten Bäumen auch mehrere Attribute abgefragt. Bei jeder Ausprägung eines Attributes werden aus der Trainingsdatensatzmenge alle Daten

[41] Die tabellarische Darstellungsform heißt auch Entscheidungstabelle. Sie zeigt die existierenden Datensätze sowie die nach Regeln geltenden Eingaben (Bedingungen) und Ergebnisse (Ausgaben). Die Tabellen gliedern sich in 4 Bereiche: Oben links zeigt die Bedingung, rechts die Werte, unten links die Regel und rechts die daraus resultierenden Ergebnisse [vgl. Cleve, Lämmel: 2016, S. 68 ff.].
[42] Eine Instanzmenge bedeutet Beispielmenge oder gegebene Datenmenge.
[43] Die Ergebnisse von Entscheidungsbäume werden mithilfe von gerichteten Graphen visualisiert und repräsentieren die Zusammenhänge der Ergebnisse. Die Entscheidungen folgen hierarchisch aufeinander, wie zum Beispiel nach vordefinierten Regeln wie **[**WENN *Bedingung* DANN *Folgerung*] oder [WENN *Bedingung 1* UND *Bedingung 2* DANN *Folgerung].* Die Knoten sind Teilmengen der Objektmenge. Sie entsprechen der Anordnung von allen Attributen einer Entscheidung. Ihre Reihenfolge bestimmt auch die Größe des Baumes. Dadurch entstehen viele unterschiedliche Möglichkeiten zur Klassenzuordnung. Die Anzahl der Ausprägung eines Attributes entspricht der Anzahl der weiterführenden Kanten.

mit dem gleichen Attributwert selektiert und anschließend mit der Ausprägung abgeglichen [vgl. ebd., S. 92]. Um die effektivste Auswahl zu treffen, besteht neben der manuellen (für kleine Attributmengen) und zufälligen Bestimmung der Attribute auch die Berechnung. Dabei wird automatisch nach einem Attribut gesucht, welches den kleinsten Baumumfang erzeugt. Es wird dazu vorher mit allen anderen verglichen [vgl. ebd., S. 95]. In diesem Zusammenhang ist die Verarbeitung metrischer Daten zum Beispiel mit großen Wertebereichen bei Jahresgehältern schwierig, da nicht für jede Ausprägung eine Kante angelegt werden kann. Daher fasst man die Ausprägungen solcher Daten vorher in Intervalle zusammen oder erstellt Schwellenwerte, die sich nach Werten, die größer oder kleiner sind, orientieren. Zur Optimierung des Ergebnisses eines Entscheidungsbaumes werden während und nach der Entwicklung (Induktion) auch unterschiedliche Pruning-Verfahren angewendet. Dabei werden Knoten entfernt oder ihre Entstehung unterbunden, die aufgrund widersprüchlicher Daten oder Ausreißern nur eine geringe Aussagekraft haben [vgl. Petersohn: 2005, S. 139]. Ziel ist es, den Entscheidungsbaum möglichst klein zu halten. Im Folgenden werden zwei Beispielalgorithmen für Entscheidungsbäume gegeben:

- **ID3-Algorithmus**
  Für eine automatische Berechenbarkeit einer geeigneten Attributabfrage im Entscheidungsbaum wird das grundlegende Verfahren der Informationstheorie „Entropie von Shannon“[44] anhand des ID3-Algorithmus angewendet. Dabei wird das Attribut mit dem größten Informationsgehalt gesucht. Dieser ergibt sich aus der Anzahl der Werte und Ausprägung eines Attributes sowie der Wahrscheinlichkeit über das Vorhandensein der Trainingsmenge. Je größer der Informationsgehalt ist, desto unklarer ist die Teilmenge. Kommt z.B. in einer Teilmenge nur eine Klasse vor, so ist der Informationsgehalt auf das gesuchte Zielattribut niedrig [vgl. Cleve, Lämmel: 2016, S. 96 f.].
- **C4.5-Algorithmus**
  Da der ID3-Algorithmus keine numerischen Attribute verarbeiten kann, wurde der C4.5-Algorithmus entwickelt. Aufgrund einer Umwandlung in Intervalle werden numerische Attribute in ordinale Attribute umgewandelt. Er normalisiert zudem den Informationsgewinn im Gegensatz zum ID3-Algorithums, der stärker Attribute mit vielen Ausprägungen in der Sortierung berücksichtigt.

Weitere Beispiele für bekannte Algorithmen, die Klassifikationsbäume entwickeln, sind CLS, CHAID, AQ15, CN2, NewID, C5 und LCLR [vgl. Petersohn: 2005, S. 139].

[44] Der Mathematiker Claude Elwood Shannon charakterisiert Nachrichten anhand von ihrer Informationsdichte. Seine Übersetzung von Kommunikationsinhalten in mathematische Formeln erfolgt über Messungen von Dateneigenschaften. Es kann keine direkte Auskunft über den Inhalt gegeben werden, da semantische und pragmatische Aspekte nicht messbar sind, um in Betracht gezogen zu werden [vgl. Shannon, Weaver: 1968].

K-Nearest Neighbour (KNN)
Dieses instanzbasierte Verfahren setzt für die Verarbeitung der Daten ein Abstandsmaß voraus und gleicht anhand von Beispielobjekten die Ähnlichkeit zu unbekannten Objekten für die Klassifikation ab, um sie einzuordnen [vgl. ebd., S. 83].

Naive-Bayes-Algorithmus
Der Naive-Bayes-Algorithmus wird angewendet, um die Wahrscheinlichkeit einer Klasse vorhersagen zu können. Er wird ohne Trainingsdaten angewendet und basiert auf der Bayesschen Formel, die das Vertauschen abhängiger Ereignisse erlaubt. Das ist sinnvoll, wenn nur eine bedingte Wahrscheinlichkeit[45] verfügbar ist, und erlaubt dadurch die Berechnung der umgekehrten bedingten Wahrscheinlichkeit. Diese Methode ist effizient, da sie nur einmal die Trainingsmenge nach relativen Häufigkeiten durchsucht. Die Voraussetzung von durchweg unabhängigen Attributen ist allerdings nachteilig [vgl. Cleve, Lämmel: 2016, S. 111-117].

Neuronales Netz
Ein neuronales Netz ist eine Computersimulation, die nach dem Vorbild des menschlichen Gehirns einen Wissensspeicher durch Elemente und Vorgehensweisen aufbaut. Es handelt sich um eine Informationsverarbeitung, die aus einer Eingabeschicht, Zwischenschichten und einer Ausgabeschicht erfolgt. Alle künstlichen Neuronen einer Schicht sind mit gleicher Anzahl an Kanten zu Neuronen der nächsten Schicht verbunden. Die Stärke der Kanten und Ausrichtung der Neuronen definieren eine Gewichtung [Cleve, Lämmel: 2016, S. 47-48, 117-129]. Gelerntes Wissen wird durch die Veränderung der Gewichtung ausgedrückt und ergibt sich aus der Trainingsphase nach vier möglichen Regelanwendungen [vgl. Rey, Wender: 2010]:

- **Hebb-Regel**
  (Gleichzeitige Aktivität zweier verbunden Neuronen.)
- **Delta-Regel**
  (Vergleich zwischen gewünschten und beobachteten Ergebnis.)
- **Backpropagation of Error**
  (Fehlerrückführung bei versteckten Schichten, um die Gewichte nach gewünschten Mustern zu verändern.)
- **Competitive Learning**
  (Kategorisierung nach Ähnlichkeiten mittels Eingabeneuronen-Vergleich.)

Um die Komplexität eines solchen Perfomance-Modells zu veranschaulichen, hat Bookbytes-Blogger Stephan Selle [vgl. 2016] dazu neuronale Netze mit einem Flipper verglichen. Dabei nehmen Hindernisse und mehrere hundert kleine Magneten Einfluss auf den Lauf einer Metallkugel. Es gibt am Ende für die Kugel zwanzig verschiedene Ausgangsmöglichkeiten, die abhängig von der

[45] Eine bedingte Wahrscheinlichkeit verbindet zwei Ereignisse miteinander. Die Wahrscheinlichkeit des Eintretens eines Ereignisses setzt ein bereits eingetretenes Ereignis voraus.

Geschwindigkeit des Startabschusses und den Hindernissen sind. Ein Programm soll die Anziehungskraft der Magneten dahingehend verändern, dass die Kugel immer rechts außen landet. Zunächst schießt der Flipper die Kugeln zufällig mit unterschiedlichen Geschwindigkeiten. Das Programm merkt sich den getroffenen Ausgang und verändert daraufhin die Magnetstärken. Bei jedem Weg, den die Kugel mehr rechts verläuft, werden die Magnetstärken gespeichert und die Laufbahn optimiert. Im Endergebnis landen alle Kugeln, trotz unterschiedlicher Geschwindigkeit und Hindernissen im vorgegebenen Zielausgang, da das Programm gelernt hat, wie die Magneten zielführend reagieren müssen. Durch das eigenständige Lernen aus den Bewertungen von Datensammlungen sind die automatischen Entscheidungen, die zum Ergebnis führten, kaum nachzuvollziehen.

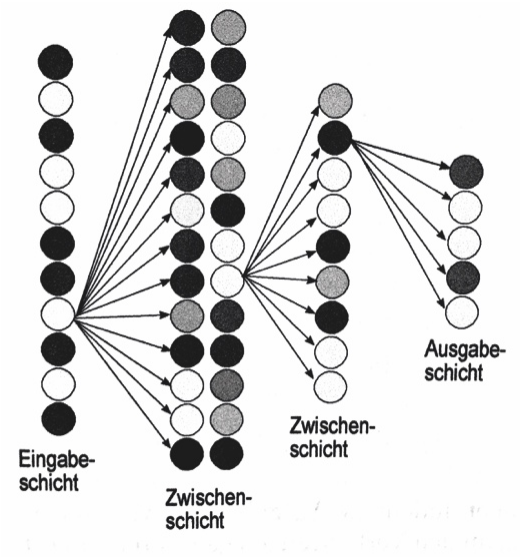


Abbildung 6: Künstliches neuronales Netz *[Cleve, Lämmel: 2016, S. 47]*

Support Vector Machines (SVM)

Ziel der Klassifizierung ist es, einzelne Klassen am effektivsten voneinander zu trennen. Die Daten werden dazu durch Geraden im zweidimensionalem Raum, Ebenen im dreidimensionalem Raum oder Hyperebenen im n-dimensionalen Raum separiert [Cleve, Lämmel: 2016, S. 130 ff.]. Da sich nur metrische Daten verarbeiten lassen, müssen alle anderen Datentypen entsprechend vorher umgewandelt werden. Mit Support Vector Machines können auch nicht-linear trennbare Daten klassifiziert werden.

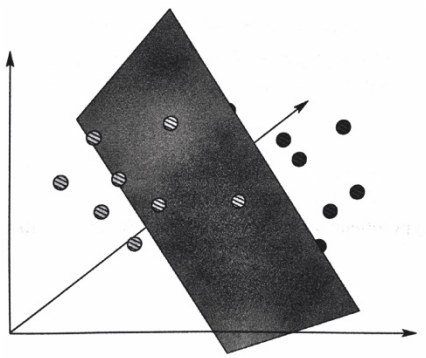


Abbildung 7: SVM im dreidimensionalen Raum *[Cleve, Lämmel: 2016, S. 131]*

Fehlerquellen und zukünftige Herausforderungen von Klassifizierungsverfahren

Da die Klassifizierung mit Trainingsmengen erfolgt, kommt es immer wieder zu Fehlern. Googles Content-ID-Algorithmus beispielsweise scannt alle auf der Internetplattform „YouTube" hochgeladenen Videos und klassifiziert sie nach urheberrechtlich geschützter Musik. Alle Rechtsverletzungen führen automatisch zur einer Löschung. Dieser Algorithmus lernt von einer Trainingsmenge, die nach Definitionen und Kriterien von Menschen erzeugt wurde. Es kann durch die subjektive Sichtweise und Voreingenommenheit so öfters zu Vorurteilen und Fehlern bei der Klassifizierungsbewertung kommen [vgl. Diakopoulos: 2015].

Auf der anderen Seite gibt es derzeit Diskussionen um vollautomatische Entscheidungen, also ohne menschliche Bewertungskomponente, die zukünftig ganz verboten werden sollen.
Ab dem nächsten Jahr treten bezüglich des Einsatzes von sogenannten „deep learning"-Algorithmen neue EU-Datenschutzrichtlinien in Kraft[46]. Intelligente, lernende Programme wie neuronale Netze kommen immer vermehrter zum Einsatz, um auf Basis von Massendaten Vorhersagen und Vorschläge zu berechnen. Sie sollen zukünftig nachvollziehbar sein, da sie, wie beispielsweise bei der Entscheidung über Kreditwürdigkeiten, zu diskriminierenden, benachteiligenden und lebensrelevanten Entscheidungen führen können. Das wird die Entwickler von neuronalen Netzen vor große Herausforderungen stellen. Durch das eigenständige Lernen der Systeme aus den Bewertungen von Datensammlungen sind die automatischen Entscheidungen kaum nachzuvollziehen.

[46] Das EU Datenschutz Grundverordnung (EU-DSGVO), Artikel 14.2 (g) zur Informationspflicht und Artikel 15.1 (h) zum Auskunftsrecht über personenbezogene Daten verpflichtet die Verantwortlichen „aussagekräftige Informationen über die involvierte Logik sowie die Tragweite und die angestrebten Auswirkungen einer derartigen Verarbeitung für die betroffene Person" zur Verfügung stellen [https://www.datenschutz-grundverordnung.eu/inhalte-der-eu-datenschutz-grundverordnung/].

3.9.2 Klassenbildung

Klassenbildung *(engl. Clustering)* bedeutet, ähnliche Objekte zu finden und in Untermengen/Gruppen zusammenzufassen. Die berechenbare Distanz zwischen den Objekten einer Menge ergibt dabei die Ähnlichkeit. Die Cluster sollten möglichst zueinander unähnlich sein. Man benennt die Teilmengen anschließend in einzelne Klassen. Die Cluster-Analyse ist ein Beispiel für unüberwachtes Lernen, da kein Ergebnis oder eine Beispielmenge vorgegeben wird, um es vergleichen zu können. Im Folgenden werden vier Clusterverfahren genauer vorgestellt.

Partitionierende Clusterbildung

Die Datensätze werden in eine vorgegebene, beliebige Anzahl von Clustern unterteilt. Die Anfangspartitionierung muss dazu bekannt sein. Anschließend werden die enthaltenden Objekte eines Clusters nach und nach mithilfe der Median-und Zentroid[47] -Berechnung zwischen den Clustern umgeordnet und das Ergebnis schrittweise optimiert. Jedes Cluster besteht aus mindestens einem Objekt, das wiederum höchstens in einem Cluster enthalten ist [vgl. Cleve, Lämmel: 2016, S.137 f.].

- **Austauschverfahren mit Zielfunktion**
  In dieser Methode wird in jedes Objekt iterativ nach Gütekriterien geprüft und neu zugeordnet, sobald die Berechnung eine Verbesserung der Klassenzuordnung ergibt. Ziel ist es, die Anfangspartitionierung schrittweise zu optimieren. Dazu wird häufig die Berechnung des Varianzkriteriums angestrebt [vgl. Petersohn: 2005, S. 95].

- **Minimaldistanzverfahren**
  Die ebenfalls iterative Optimierung der Cluster erfolgt durch die Abstands-Prüfung jedes Objektes zum Klassenschwerpunkt und anschließend neuer Zuordnung. [vgl. Petersohn: 2005, S. 96].

- **k-Means-Verfahren**
  In der k-Means-Methode wird zunächst der Klassen-Zentroid der Ausgangspartitionierung berechnet. Anschließend wird ein erstes Objekt mittels Abstandmessung zum Zentroid der Klasse zugeordnet, deren Abstandwert am geringsten ist. Anschließend wird der Vorgang wiederholt, indem erneut die Zentroide errechnet werden.

Hierarchische Clusterbildung

Die Clusterbildung erfolgt anhand von ähnlichsten Merkmalen, also kleinster Distanz zueinander, und vermengen sich paarweise bis zum Endpunkt aus einem letzten großen Cluster. Dadurch entsteht schrittweise eine hierarchisch aufgebaute Baumstruktur. Dieses Verfahren kann in beide Richtungen angewendet werden: Entweder ein großer Cluster (Gesamtmenge) spaltet sich immer weiter

[47] Bei einem Median handelt es sich um den Wert, der dem Mittelwert (Zentroid) einer Menge am nächsten liegt.

paarweise (divisives[48] Clustering) auf, oder einzelne Objekten einer Menge bilden wachsende Cluster (agglomeratives[49] Clustering). Ein Nachteil der hierarchischen Verfahren ist die schwierige Skalierbarkeit sowie eventuelle Ausreißer[50]. Ist ein Objekt einmal einem Cluster zugeordnet, kann diese Zuordnung nicht mehr aufgehoben werden. Auf der anderen Seite sagt dieses Modell viel über die Einzelbeziehungen der Cluster aus [vgl. Cleve, Lämmel: 2016, S.138 f.]. Neben der Abstandsmessung zweier Klassen mittels Zentroid und Median werden häufig folgende vier Verfahren angewendet.

- **Das Single Linkage (Nearest Neighbour)**
  Die Abstandmessung zweier Klassen erfolgt durch die Distanzmessung zu den zwei nächstliegenden Objekten der Cluster. Es entstehen zunächst wenige große und viele kleine Cluster, so dass dieses Verfahren für das Extrahieren von Ausreißern besonders nützlich ist [vgl. Petersohn: 2005, S. 92].

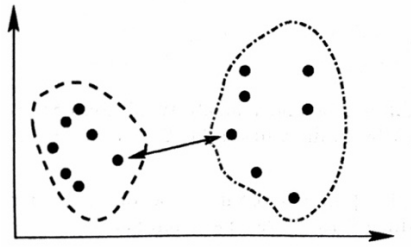


Abbildung 8: Single Linkage *[Cleve, Lämmel: 2016, S. 160]*

- **Complete Linkage (Furthest Neighbour)**
  Zwei Cluster verschmelzen, anhand der Abstandsmessung der zwei am weitesten entfernten Objekte der Cluster. Die Ergebnisse dieses Verfahrens sind oft viele kleine gleich große Cluster [vgl. Petersohn: 2005, S. 93].

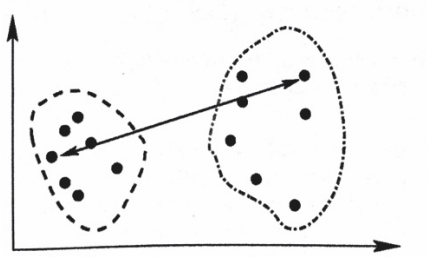


Abbildung 9: Complete Linkage *[Cleve, Lämmel: 2016, S. 161]*

[48] absteigend
[49] ansteigend
[50] Ausreißer sind Wertausprägungen, die sich stark von den anderen Werten unterscheiden [vgl. Cleve, Lämmel: 2016, S.10].

- **Average Linkage (Within Groups)**
  Alle Objektabstände in den Clustern werden berücksichtigt. Es handelt sich um eine Kombination aus Single und Complete Linkage [vgl. Petersohn: 2005, S. 93].

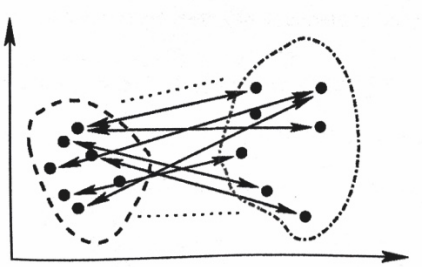


Abbildung 10: Average Linkage *[Cleve, Lämmel: 2016, S. 161]*

- **Ward**
  Diese Methode beinhaltet ein Varianzkriterium innerhalb der Gruppen. Die Abstände der im Cluster enthaltenden Objekte zum Mittelpunkt werden dazu berechnet. Die zwei Gruppen, deren Varianz-Summe den geringste ist, werden geclustert [vgl. Cleve, Lämmel: 2016, S.162].

Dichtebasierte Clusterbildung

Nicht jede Menge kann nach einem Mediod und Centroid berechnet werden, wenn sie stark unterschiedlich räumliche Strukturen haben. Es lassen sich jedoch Cluster auch anhand der Dichte des Objektvorkommens bilden. Alle Objekte werden demnach einem Cluster zugeordnet, die einen festgelegten Dichte-Schwellwert übersteigen [vgl. Cleve, Lämmel: 2016, S.139-140, S. 160-163]. Ein bekannter Algorithmus ist der DBScan (*Density-Based Spatial Clustering of Applications with Noise*). Wie in Abbildung 9 zu sehen ist, bildet der Algorithmus Schritt für Schritt einen Cluster aus den Punkten der gekrümmten Punktewolke. Noise-Punkte nennt man Punkte, die zu keinem Cluster zugeordnet werden.

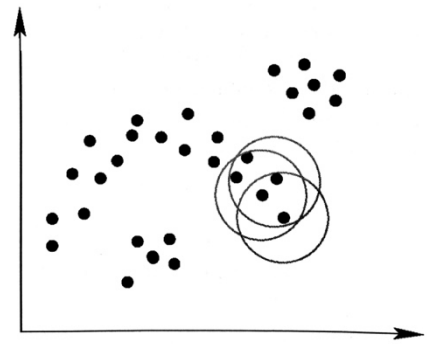


Abbildung 11: DBScan *[vgl. Cleve, Lämmel: 2016, S. 162]*

Weitere Methoden zur Clusterbildung
Aufgrund der Vollständigkeit wird an dieser Stelle auf die Möglichkeit der Clusterbildung mit Performance-Modellen hingewiesen. Dabei handelt es sich, wie bereits im Zusammenhang mit den Klassifizierungsverfahren erwähnt, um Systeme, deren Ergebnisse nicht mehr vollständig nachvollziehbar und zu erklären sind. Für die Clusterbildung stehen folgende verschiedene Methoden zur Verfügung: Self Organizing Maps (SOM), Adaptive Resonance Theory (ART), FuzzyArt oder Neuronale Netze [vgl. Cleve, Lämmel: 2016, S.140].

Im Vorangegangenen wurden nach Cleve und Lämmel [vgl. 2016] die unterschiedlichen Klassifizierungs- und Cluster-Verfahren aufgeführt. Für eine tiefere Betrachtung wird allerdings die Literatur über Data Mining von Cleve und Lämmel [vgl. 2016], Runkler [vgl. 2015] sowie Petersohn [2005] empfohlen.

3.9.3 Weitere Anwendungsklassen im Data-Mining-Prozess

Aufgrund der Vollständigkeit wird folgend auf weitere Data-Mining-Verfahren hingewiesen, die im Rahmen dieser Arbeit jedoch nicht näher betrachtet werden. Es wird daher die Literatur über Data Mining von Cleve und Lämmel [vgl. 2016], Runkler [vgl. 2015] sowie Petersohn [2005] empfohlen.

Numerische Vorhersage
Die numerische Vorhersage berechnet und prognostiziert zukünftige Werte von Datensätzen anhand von Trainingsdaten. Im Unterschied zur Klassifizierungsanalyse, die diskrete Werte[51] für ein Objekt oder Datensatz prognostiziert, bestehen die Zielattribute aus Zahlen, die aus einem unendlich großen Wertbereich bestehen können. Das Analyseverfahren kann dazu dennoch auch Methoden des Klassifikationsverfahren anwenden. Als Beispiel für ein numerisches Analyseverfahren gilt das lineare Regressionsverfahren. Es beschreibt den Trend oder durchschnittlichen Zusammenhang numerischer Attribute wie bei Aktienkursen oder Temperaturdaten [vgl. Cleve, Lämmel: 2016, S. 62].

Zeitreihenanalyse
Diese Analyseanwendung umfasst die Entwicklung eines Modelles, das einen konkreten Wert prognostiziert. Anhand vorliegender Zeitreihen werden Merkmale herausgefiltert und definiert, um ein Prognose-Modell zu entwickeln. Dazu werden Trendindikatoren wie Zeitverläufe und Wendepunkte berücksichtigt [vgl. Petersohn: 2005, 173 ff.]

Assoziationsanalyse
Bei der Assoziationsanalyse werden Beziehungen und Zusammenhänge zwischen Attributen analysiert. Es wird nach Regelmäßigkeiten gesucht, um Datenverhaltensweisen vorherzusagen. Dazu werden häufig gemeinsam auftretende Objekte extrahiert. Beispiele für die Anwendung von

[51] Diskrete Zahlen sind ganze, positive Zahlen ohne Zwischenwerte. Sie bestehen entweder aus endlich vielen oder unendlich abzählbaren Zahlen.

Assoziationsverfahren basieren ursprünglich auf Warenkorbanalysen (*recommander engines*) für Vorhersagen von Kaufverhalten. Auf dem Prinzip basieren beispielsweise auch Risikoabschätzung von Kreditvergaben, Präventionsmaßnahmen der Polizei oder die Optimierung der Homepage aus vorhergehendem Nutzer-Tracking. Faktoren zur Häufigkeit und Erfolg der Regelanwendung werden in der Analyse berücksichtigt. Es gibt hierarchische, quantitative, unscharfe (*fuzzy association rules*) und temporale Assoziationsregeln. Beispielalgorithmen zur Booleschen Assoziationsanalyse sind AIS, SetM, A-Priori, Frequent Pattern Growth [vgl. Petersohn: 2005, S.101 ff.].

3.9.4 Datenvisualisierung

Visualisierungen dienen nicht nur der graphischen Darstellungsform von Ergebnissen (*presentation graphics*), sondern können zusätzlich auch als eigenständige Analyse-Technik im Data Mining angesehen werden (*exploratory graphics*) [vgl. Cleve, Lämmel: 2016, S. 15]. Der Einsatz graphischer Verfahren, die auf statistischen Methoden basieren, dient der optischen Erkennung von Mustern, Strukturen, Abhängigkeiten und Abweichungen in unbekannten Datensätzen. Interaktive Anwendungen sind ein großer Bestandteil von visueller Analysetechnik und werden als explorative Datenvisualisierung bezeichnet. Am Beispiel der Cholera Map von 1854 lässt sich die Bedeutung von Visualisierungen verdeutlichen. Aufgrund der exakt adressierten Markierungen von auftretenden Cholerafällen in London konnte der Arzt John Snow eine Karte erstellen und eine verseuchte Wasserpumpe anstelle anderer Vermutungen als Ursache erklären.

Abbildung 12: Cholera-Map *[vgl. Wikipedia]*
Visualisierung der Cholerafälle an der Broad Street Pump in London, 1854. Die Karte zeigt Anhäufungen von Todesfällen im Umkreis der verseuchten Pumpe.

Auch der Statistiker und Grafikdesigner Edward R. Tufte verdeutlicht am folgenden Beispiel, dass manchmal erst eine Datenvisualisierung Schlussfolgerungen möglich macht [vgl. 1993, S. 13-14]:

I

x	y
10,00	8,04
8,00	6,95
13,00	7,58
9,00	8,81
11,00	8,33
14,00	9,96
6,00	7,24
4,00	4,26
12,00	10,84
7,00	4,82
5,00	5,69

II

x	y
10,00	9,14
8,00	8,14
13,00	8,74
9,00	8,77
11,00	9,26
14,00	8,10
6,00	6,13
4,00	3,13
12,00	9,13
7,00	7,26
5,00	4,76

III

x	y
10,00	7,46
8,00	6,77
13,00	12,74
9,00	7,11
11,00	7,81
14,00	8,84
6,00	6,08
4,00	5,39
12,00	8,15
7,00	6,42
5,00	5,73

IV

x	y
8,00	6,58
8,00	5,76
8,00	7,71
8,00	8,84
8,00	8,47
8,00	7,04
8,00	5,25
19,00	12,50
8,00	5,56
8,00	7,91
8,00	6,89

N=11
Mittelwert X = 9.0
Mittelwert Y = 7.5

Regressionsgerade: Y = 0.5 X + 3

Abbildung 13:
Datentabelle *[Tufte: 1993, S. 13]*
Vier Tabellen zeigen Werte, die sich scheinbar wenig unterscheiden.

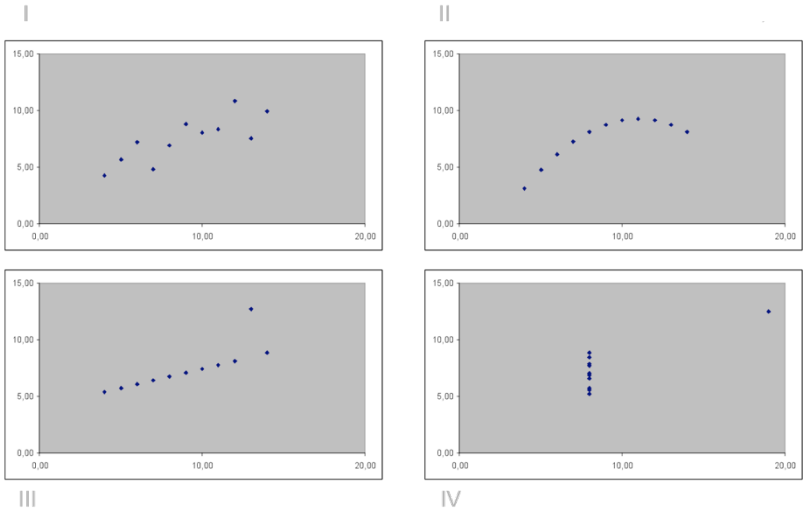


Abbildung 14:
Datenvisualisierung
[Tufte: 1993, S. 14; Anscombe, 1973]
Die graphische Darstellung ermöglicht eine Differenzierung der Tabellenwerte aus Abbildung 12 und zeigt Muster auf, die vorher nicht erkennbar waren.

Prof. Daniel. A. Keim [vgl. Keim: 2002, S. 1] erklärt den Vorteil von Visualisierungen darin, dass gegenüber der textlichen Form deutlich mehr Informationen schneller aufgenommen werden können und somit interpretierbar sind. Laut Keim können automatische Computeranwendungen der Datenanalyse durch menschliche Eigenschaften wie Kreativität, Flexibilität und Allgemeinverständnis ergänzt werden.

3.9.4.1 Visualisierungstechniken

Nachdem die Daten, wie bereits in vorigen Abschnitten beschrieben, vorbereitet wurden, wird eine geeignete Visualisierungsmethode nach Kriterien des Datentyps, der Visualisierungstechnik sowie der Interaktions- und Verzerrungsmöglichkeiten ausgewählt [vgl. Keim: 2002, S. 3].

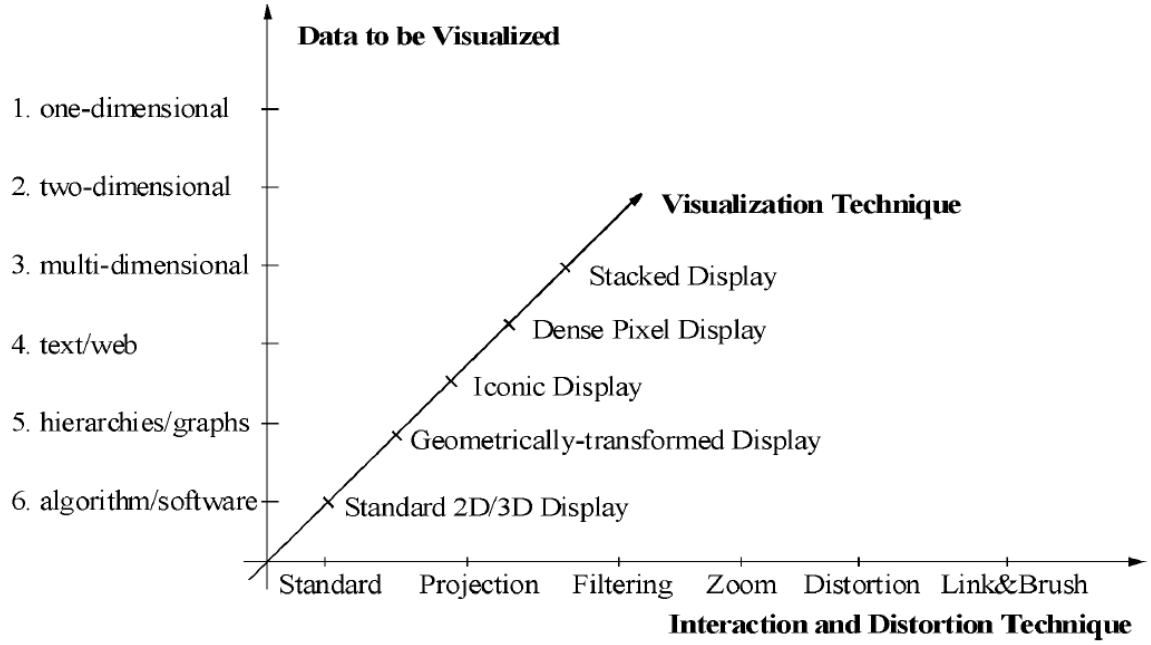


Abbildung 15: Information Visualization and Visual Data Mining *[Keim: 2002]*

Die Daten können zunächst nach verschiedenen Dimensionen (ein-, zwei- und multidimensional) abhängig ihrer Attributanzahl, ihren Beziehungen untereinander (Graphen, Entscheidungsbäume) oder der Struktur (Text) in verschiedenen Datentypen unterschieden werden [vgl. Keim: 2002, S. 5-8]. Viele bekannte Darstellungsformen wie Histogramm[52], Boxplot[53], Tortendiagramme, Liniendiagramme oder Streudiagramme[54] sind bereits aus der Statistik bekannte Methoden, die sich aufgrund der Übersichtlichkeit allerdings nur für geringe Datenmengen eignen. Daher empfehlen Cleve und Lämmel [vgl. 2016, S.253 f.] mit den Daten beispielsweise durch die Reduzierung der Dimensionsanzahl zu experimentieren und interagieren, nachdem der Datentyp bestimmt und eine favorisierte

[52] Darstellungsform für Häufigkeitsverteilungen, vorzugsweise bei gleichgroßen Klassen [vgl. Schumann, Müller: 2000, S. 135].
[53] Zusammengefasste und reduzierte Überblickdarstellung mit 5 Punkten aus Mittelwert, oberen und unteren Quantil und den zwei Extremwerten.
[54] Darstellung der Beziehung zweier Größen mittels Punkteverteilung

Visualisierungsmethode ausgewählt wurde. Dank stetig verbesserten Rechenleistungen wachsen die Realisierungsmöglichkeiten für diese unterschiedlichen Betrachtungsperspektiven.

Für die Erkundung der Daten, um z.B. Teilmengen herauszufiltern und den Fokus zu verändern, stehen folgende Interaktionsanwendungen beispielsweise zur Verfügung [vgl. op. cit.: S. 11-15]:

- **Zooming**
  (Durch interaktive Verzerrungen werden unterschiedliche Perspektiven und Detailbetrachtungen auf Teilmengen ermöglicht und repräsentiert.)
- **Linking**
  (Durch Verknüpfungen werden Abhängigkeiten und Beziehungen sichtbar.)
- **Brushing**
  (Durch unterschiedliches Einfärben werden z.B. Korrelationen verdeutlicht.)

Beispiele für Visualisierungstechniken und journalistische Projekte

- **Graphen**
  Graphen zeigen durch Knoten und Kanten die Zusammenhänge und Verbindungen einzelner Objekte zueinander auf.

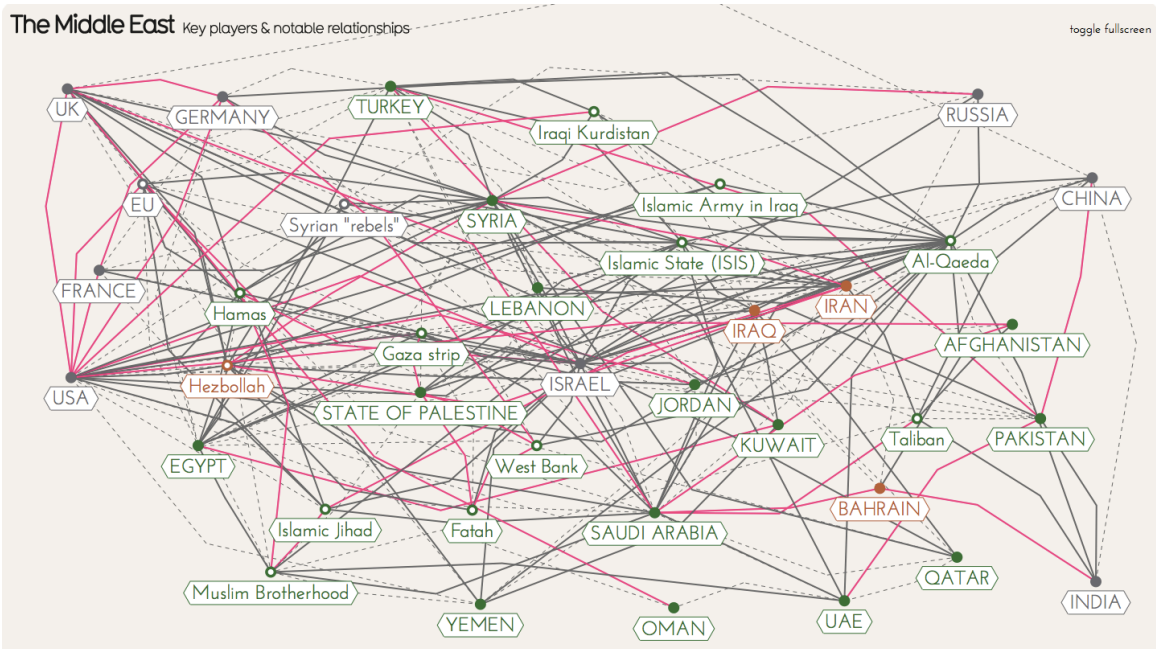


Abbildung16: The Middle East *[Mccandless: 2015]*
Die interaktive Visualisierung stellt die Beziehungen der einzelnen Länder dar. Der Betrachter kann Verbindungen auswählen, um Hintergrundinformationen zu bekommen.

- **Icons**

  Einzelne Eigenschaften werden mit vereinfachten Grafiken dargestellt. Als gutes Beispiel gelten die Chernoff-Gesichter. Da es Menschen leichtfällt, kleinste Unterschiede in menschlichen Gesichtern zu erkennen, können sie gleichzeitig 12 unterschiedliche Merkmale ausdrücken [vgl. Schumann, Müller, 2000 S. 192-194].

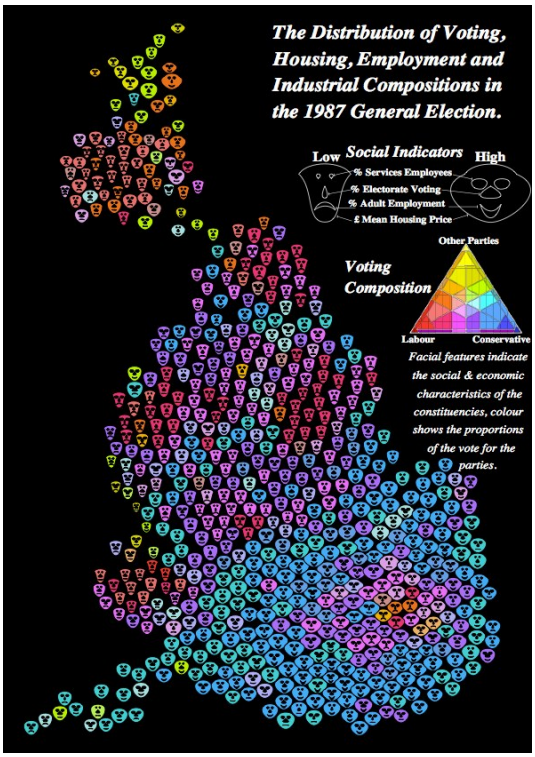


Abbildung 17:
Die Wahlkreise Großbritaniens 1987 *[Dorling: 1991]*
Das Kartogramm wird aus Gesichter mit vier Merkmalen aus sozialen Indikatoren dargestellt. Die Farben bilden die Parteien ab. Industrielle Zentren zeigen als Ergebnis aus niedrigen Mieten und geringen Beschäftigungszahlen schmale, traurige Gesichter. Im Großraum London sind die Gesichter rund und freundlich.

- **Geografische Visualisierungen**

  Kartografien dienen den Menschen schon immer zur Orientierung. Egal ob als Landkarten oder Abbildung des Streckennetzes im öffentlichen Verkehr. Durch Geolokationsdaten können genaue Bewegungsprotokolle erstellt werden.

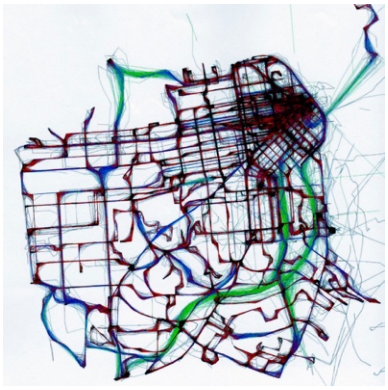


Abbildung 18: A Day of Muni *[Fischer: 2010]*
Die unterschiedlich eingefärbten Linien zeigen die unterschiedlichen Geschwindigkeiten der Municipal Transport Agency in San Francisco. Dazu wurden öffentlich verfügbarer GPS-Daten im Zeitraum von 24 Stunden ausgewertet. (Schwarz: unter 12 km/h, Rot: unter 30 km/h, Blau: unter 70 km/h, Grün: mehr als 70 km/h).

- **Interaktive-Visualisierungen**
  Diese Visualisierungen bieten dem Betrachter einen Perspektivwechsel durch die Möglichkeit einer individuellen Anpassung und Beobachtung von relevanten Teilmengen.

Abbildung 19: OCED Better Life Index *[OCED: 2015]*
Der Betrachter dieser Grafik kann Länder vergleichen und eigene Einstellungen mit bevorzugten Filterkriterien beispielsweise zu den Einkommensverhältnissen machen.

3.9.4.2 Bedeutung von Datenvisualisierung

Oft können grafische Darstellungen bei der Erkennung von Mustern, Zusammenhängen und somit Themen bzw. Narrationen helfen. Informationen aus komplexen Daten sind dadurch erfassbar und interpretierbar. Daher ist die Datenvisualisierung ein wichtiger Bestandteil des Datamining-Prozesses datenjournalistischer Arbeit. Sie dient darüber hinaus jedoch auch der Repräsentation der Ergebnisse.

Visuelle Darstellungen können gegenüber Tabellen und Texten Informationen großer und komplexer Datensätze schnell veranschaulichen und verständlich aufbereiten. Infografiken, laut Webentwickler Markus Nix „die bisherigen Stiefkinder und Lückenfüller des Journalismus“ [vgl. 2014, S.20], gewinnen immer stärker an Bedeutung. Die Vorteile individueller Interaktionsmöglichkeiten mit Grafiken werden zunehmend erkannt, deren Realisierung dank modernster Software und Computerleistung kein Problem mehr darstellen. Es gibt inzwischen erste Experimente zum Datenjournalismus im virtuellen Raum.

Simon Rodgers vom Guardian Datablog, berichtet über seine ersten Erfahrungen mit VR-Datenvisualisierungen zum Brexit unter: www.simonrogers.net/2016/06/20/how-we-made-a-vr-data-visualization/ [vgl. 2016]. Er rechnet VR zukünftig eine wachsende Bedeutung zu: „Virtual Reality is a powerful tool in making journalism more immersive to its readers [...]".

Ein zukünftig vielversprechender Anwendungsbereich für datenjournalistische Visualisierungen könnte auch Augmented Realitiy werden. Es wird zunehmend wichtiger, die Ausspielung der Inhalte dem Kontext anzupassen. Durch Augmented Reality können Inhalte den User zur richtigen Zeit und am richtigen Ort erreichen. Die Umgebung wird zum Informationsraum, in dem Daten aller Art zur Verfügung gestellt werden können [vgl. Schroll, 2010]. Dadurch wird diese Darstellungsform auch für die journalistische Arbeit interessant. Die Präsentationsflächen können dabei Oberflächen wie Displays, Wände und Plakate sein, die durch Muster-, Umriss - oder Positionserkennung mit Visualisierungen bespielt werden [vgl. Hayes: 2009].

Das Internet, mit seiner schnellen multimedialen Erzählstruktur, fordert die Aufmerksamkeit der Nutzer für viele Informationen gleichzeitig. Dies zwingt die Medienmacher, ihre Inhalte zunehmend anzupassen und attraktiver hervorzuheben. Daten in Kombination mit neuen Technologien und Softwarelösungen bieten dabei viele Möglichkeiten und das Internet eine ideale Plattform für interaktive Datenvisualisierungen. Erfolgreiche Beispiele dafür sind im Internet unter www.informationisbeautiful.net, blog.zeit.de/open-data/category/datenvisualisierung/ oder auch infosthetics.com/ einzusehen.

Für eine weiterführende Betrachtung wird die Literatur *Visualisierung: Grundlagen und allgemeine Methoden* von Heidrun Schumann und Wolfgang Müller, *Handbook of Data Visualisation* von Chun-houh Chen, Wolfgang Karl Härdle, Antony Unwin sowie *Visual Simplexity* von Markus Nix empfohlen. Darüber hinaus wird in diesem Zusammenhang auch auf die Masterarbeit von Truong Vinh Phan *Immersive Data Visualization and Storytelling based on 3D / Virtual Reality Platform: a Study of Feasibility, Efficiency, and Usability* hingewiesen. Einen allgemeinen Überblick zum Thema gibt zudem die Ausarbeitung *„(Explorative) Datenvisualisierung*" [vgl. Philipsen: 2015].

3.10 Evaluierung und Interpretation im KDD-Prozess

Die Ergebnisse und Muster des KDD-Prozesses werden in diesem Schritt nach den bereits erwähnten Kriterien der Neuartigkeit, Gültigkeit, Verständlichkeit und Nützlichkeit bewertet und interpretiert. Oft werden daraufhin einzelne Data-Mining-Anwendungen wiederholt, Parameter angepasst oder neue Verfahren verwendet, weil die Erkenntnisse in den ersten Durchläufen des KDD-Prozesses noch wenig aussagekräftig sind. Der KDD-Prozess ist ein Kreislauf aus Experimentieren, Anpassen und Optimieren.

3.11 Text Mining

Es gibt aufgrund unterschiedlicher Datentypen auch unterschiedliche Analysegebiete im Data Mining [vgl. Cleve, Lämmel: 2016, S. 65]. Text Mining umfasst die Analyse unstrukturierter Daten in Textdokumenten.

Dies geschieht nach ausgewählten Kriterien (*engl. Feature Selection*) ebenfalls beispielsweise mit Distanzfunktionen, um Attributähnlichkeiten herauszufinden. Kriterien für sogenannte Featurevektoren können zum Beispiel einzelne Wörter oder auch Kategorien sein und aus mehreren Werten (n-Dimensionen) bestehen [vgl. Weiss et al.: 2005, S. 35]. Aufgrund der Komplexität dieses Wissensgebietes wird in diesem Abschnitt der Arbeit daher nochmal detailliert auf das Themengebiet eingegangen.

Die Anwendungsbereiche sind sehr vielfältig und die unterschiedlichen Textanalysemethoden werden in Übersetzungs- und Korrekturprogrammen, Suchmaschinen, zur Spracherkennung z.B. bei Sprachassistenten wie Apples Siri und Wearables wie Applewatch, für Dialogsysteme z.B. bei Auskunftsdiensten durch den Einsatz von Bots oder auch von Sicherheitsdiensten eingesetzt. Im Journalismus können Textminingverfahren beispielsweise dem automatischen Generieren und Zusammenfassen von Texten, der Archivierung oder den Auswertungen von Interviews, Stimmungen in sozialen Netzwerken und Diskussionen dienen. Text Mining kann den Umgang mit großen Textmengen erleichtern und unterstützt Journalisten in der Recherche, Bearbeitung und Analyse. Die Analyse und automatische Auswertung von Texten bekommt zukünftig immer größere Bedeutung.

Im Gegensatz zum Information-Retrieval-Verfahren, das eingesetzt wird, um bekannte Informationen leichter zugänglich zu machen, werden in Textmining-Verfahren zuerst unbekannte Daten analysiert und neue Informationen generiert [vgl. Manning et al.: 2009]. Ziel vom Text Mining ist es, Texte (unstrukturierte Daten) zunächst in strukturierte und in numerische Daten zu transformieren, um sie nach Data-Mining-Verfahren weiterverarbeiten zu können [vgl. Weiss et al.: 2005 S. 2-4].

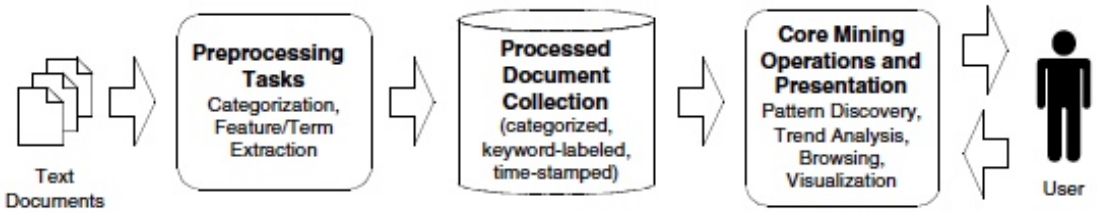


Abbildung 20: High-level text mining functional architecture *[vgl. Feldman u. Sanger: 2007, S. 1]*

Trotz unterschiedlicher Voraussetzungen ähneln sich folglich Text- und Data Mining. Zur Analyse von Textdokumenten stehen statistische Verfahren wie Differenzial-, Kookkurrenz- oder

Clusteranalysen und musterbasierte Verfahren zur Verfügung [vgl. Heyer et al.: 2006]. Um Textdokumente zu klassifizieren, werden nach relevanten Begriffen oder nach Ähnlichkeiten in Texten gesucht. Interessante Informationen werden anhand von beispielsweise Schlüsselwörtern, Häufigkeitsverteilungen, Kategorien oder Hierarchien extrahiert [vgl. Cleve, Lämmel: 2016: S. 64]. Durch viele Besonderheiten wie zum Beispiel den Umgang mit Umgangssprache, Kontextbedeutung und Mehrdeutigkeiten ist Text Mining ein eigenes Forschungsfeld aus der Wissenschaft der Computerlinguistik. Herausforderungen im Zusammenhang mit einer korrekten Sprachübersetzung können beispielsweise Homophone[55], Homografe[56], Homonyme[57], Polyseme[58] oder Metaphern[59] sein. Oder der Umgang mit Beispielsätzen wie „Peter sucht die Frau mit der Brille“. Wer hat in diesem Fall die Brille auf? Dies ist nicht eindeutig und nur im Zusammenhang zu klären. Ebenso wird es beim Beispiel „Ich will Dich nicht traurig sehen“ deutlich. Die Natural Language Software Registry bietet eine Zusammenfassung von Softwareanwendungen für derartige Herausforderungen der Textanalyse an. Das deutsche Forschungszentrum für künstliche Intelligenz (Language Technology World) gibt weitere Informationen zum Forschungsstand. Im Folgenden wird auf die geschichtlichen Grundlagen zur Textverarbeitung mittels Computer eingegangen. Es werden einzelne Schritte zur Datenvorbereitung und unterschiedliche Analysemöglichkeiten vorgestellt.

3.11.1 Grundlagen zur automatischen Sprachverarbeitung

Maschinen, Programme, Datenbanken etc. benötigen zur Datenverarbeitung spezielle Kunstsprachen – auch formale Sprachen genannt. Sie zeichnen sich gegenüber der Komplexität natürlicher Sprachen durch ihre eindeutige Definition aus, so dass eine Verarbeitung leichter möglich ist. Formale Sprachen dienen der mathematischen Verwendung anstatt der Kommunikation. Die Syntax natürlicher Sprachen, die festlegt, welche Ausdrücke zur Sprache gehören, kann somit mithilfe formaler Sprachen beschrieben werden.

Eine formale Sprache ist eine Menge von Zeichenketten, die mithilfe eines ihr zugrundeliegenden Alphabets $\sum$ gebildet werden kann. Sie ist somit eine Teilmenge aller möglicher Zeichenketten ($\sum^*$), die mit dem Alphabet ($\sum$) gebildet werden können, und wird Kleensche Hülle, nach dem amerikanischen Mathematiker Stephen Cole Kleene, genannt [vgl. Hopcroft et al.: 2002, S. 95]. Zum

[55] Homophone: Gleichklingend, aber unterschiedlich geschriebene Worte wie „Sole“, das Salzbad und „Sohle“ vom Schuh.
[56] Homografe: Gleichgeschriebene Worte mit verschiedenen Aussprachen wie bei „modern“ im Zusammenhang mit „Kompost“ oder „Mode“.
[57] Homonyme: Worte mit verschiedener Bedeutung und Herkunft wie „weiß“ als Ableitung vom Verb „wissen“ oder der Farbe.
[58] Polyseme: Gleichlautende Worte mit gleicher sprachgeschichtlicher Herkunft wie bei „Strom“ oder „Föhn“.
[59] Metapher: Andere Wortbezeichnungen wie „Wüstenschiff“ für „Kamel“ oder „Nadel im Heuhaufen“.

Beispiel benötigt die Sprache (L) Englisch das lateinische Alphabet (∑) und die Programmiersprache C (L) als Alphabet den Ascii[60]-Zeichensatz (∑).

Für die Beschreibung der Komplexität formaler Sprachen stellte der Sprachwissenschaftler Noam Chomsky 1956 vier hierarchisch abhängige Grammatiken[61] von Typ 0 bis Typ 3 auf. Seine Ausführungen bauten auf seiner Theorie der Universalgrammatik aller menschlichen Sprachen auf, die dem Menschen angeboren sei. Chomsky vertrat damit die bisher gegensätzliche Auffassung zur behavioristischen Theorie, dass der Spracherwerb nur ausschließlich durch Lernprozesse erfolgt [vgl. Chomsky: 1957].

Die Chomsky-Hierarchie teilt formale Grammatiken in vier unterschiedliche Klassen ein, die sich durch eine wachsende Ausdruckskraft unterscheiden. Anhand ihrer Regeln kann entschieden werden, ob Ausdrücke oder Zeichenreihen einer Sprache zugehörig sind oder nicht. Von Typ 0 aufsteigend nimmt ihre Einschränkung zu [vgl. Hopcroft et al.: 2002]:

- **Typ 0 –Rekursiv aufzählbare Grammatiken** [vgl. Hopcroft et al.: 2002, S. 373 f., 377 ff.]
  (Diese Ausdrücke sagen am wenigsten aus, da die Regeln keine Einschränkungen geben. Als Beispiel gilt Touring-Maschine, die nicht anhält.)

- **Typ 1 – Kontextsensitive Grammatiken** [vgl. Hopcroft et al.: 2002, S. 179 ff.]
  (Diese Grammatiken erzeugen Sprachen, die in endlicher Zeit über die Sprachzugehörigkeit einer Zeichenreihe entscheiden. Es gibt Start- und Endpunkte sowie Zwischenzustände in der Beschreibung sogenannter Keller- oder Pushdown-Automaten. Diese wissen nicht nur, wieviel Zwischenzustände gespeichert sind, sondern auch welche Ausführung (Zwischenzustände) in welchem Kontext steht.)

- **Typ 2 – Kontextfreie Grammatiken** [vgl. Hopcroft et al.: 2002, S. 202]
  (Wenn reguläre Ausdrücke nicht ausreichen, wie bei der Beschreibung vieler Programmiersprachen und natürlicher Sprache, werden kontextfreie Grammatiken verwendet. Diese sogenannten Kellerautomaten haben eine Speicherfunktion des aktuellen Zustands. Im Gegensatz zum Typ 1 ist die Reihenfolge nach dem Stapelprinzip (*engl. Stack, „last in first out")* festgelegt. Ein gutes Beispiel ist die mögliche Beschreibung eines XML-Dokument. Es können alle verschachtelten Tags nacheinander abgearbeitet werden. Typ 1 und 2 werden vornehmlich für den Compilerbau[62] angewendet.

---

[60] Ascii (*engl. American Standard Code of Information Interchange*) ist eine genormte Darstellung von u.a. Ziffern, Buchstaben und Sonderzeichen.[60]

[61] Eine Grammatik ist ein endliches Regelsystem, das eine unendliche Menge von Sätzen einer formalen Sprache erzeugen (generative Grammatik) oder erkennen (analysierende Grammatik) kann [vgl. Wissens-Portal ITwissen.info: 2016]

[62] Compiler übersetzen Quellcode in einen maschinenlesbaren Zielcode. Dazu wird die Syntax überprüft, analysiert und optimiert, um anschließend einen Code erzeugen zu können.

- **Typ 3 – Reguläre Grammatiken** [vgl. Hopcroft et al.: 2002, S. 43, 187 ff.]
  (Mit regulären Grammatiken können endliche Automaten abgebildet werden. Sie gilt als eingeschränkteste Grammatik, ist dadurch jedoch nicht so komplex. Ein Beispiel ist das Rechtschreibkorrekturprogramm von Word oder Suchmaschinen, die nach Schlagwörtern in Texten suchen. Es wird stets nach Regeln, wie beispielsweise *„nach einem Punkt muss eine Leertaste oder Großbuchstabe folgen“*, geprüft. Auf einen inakzeptablen Zustand wie ein Kommazeichen nach einem Punkt, folgt eine Fehlermeldung.)

3.11.2 Textvorbereitung

Für ein erfolgreiches Text Mining erfolgt zum Anfang eine Einordnung der Textarten. Die Zuordnung in Sprache, Sachgebiete wie Politik, Kunst oder Wirtschaft, Texttypen (Dokumentation, Werbung oder Protokolle) und die Textform (z.B. Mail, Buch, Zeitung) dienen als erste wichtige Hinweise zur Weiterverarbeitung [vgl. Heyer et al.: 2006, S. 11 f.). Es muss zwischen Fachsprachen und Allgemeinsprachen unterschieden werden. Fachsprachen zeichnen sich durch eine besondere Grammatik, spezielle Fachterminologie, charakteristische Lesearten (wie in juristischen Texten) sowie spezifischer Morphologie aus, wie beispielsweise die häufige Wortformenendung „-itis“ in medizinischen Texten [vgl. Heyer et al.: 2006, S. 45-49, 51 ff.]. Neben den Analyse-Texten werden auch Referenztexte zum Vergleich und maschinellem Lernen benötigt. Man kann dazu die unterschiedlichen Textquellen den Analyseverfahren wie folgt zuordnen:

- **Unstrukturierter Text**
  Statistische und clusterbasierte Verfahren arbeiten mit unstrukturiertem Ascii-Text. Das Internet ist die größte Quelle unstrukturierter Texte, wo jedoch häufig Mundart-Sprache verwendet wird. Für die Analyse von Standardsprachen eignen sich daher die Textsammlungen der European Language Ressource Association (ELRA), European Corpus Initiative (ECI), Linguistic Data Consortium (LDC) sowie die des Instituts für deutsche Sprache (IDS).

- **Annotierter Text**
  Annotierte Texte sind bereits klassifiziert, indem sie syntaktisch kategorisiert wurden. Sie bestehen aus einer standardisierten Anzahl von Tags (POS-Tags, grammatische Tags, siehe Kapitel 4.11.4: Statistische Methoden).

- **Maschinenlesbare Lexika**
  Dabei handelt es sich um Wortlisten, die um statistische, syntaktische, semantische, terminologische und pragmatische Angaben ergänzt sind. Weitere Informationen zur Standardisierung maschinenlesbarer Wörterbücher bietet das Consortium for Lexical Research. Das Projekt „Deutscher Wortschatz“ der Uni Leipzig bietet ein umfangreiches Lexikon deutscher Sprache an.

Zur Vorbereitung der Textanalyse werden alle Texte in Textdatenbanken gespeichert. Neben unveränderten Originaldokumenten sollten die Texte genormt, sprich in Ascii, konvertiert sein. Sie werden dazu in einzelne Sätze und Wortformen zerlegt und sollten mit einem separaten Wörterbuch aus syntaktischen und semantischen Vorgaben verknüpft werden [vgl. Heyer et al.: 2006, S. 57]. Wenn Texte aus anderen Formaten konvertiert werden, kann es zu Schwierigkeiten mit der Reihenfolge und Trennung von Sätzen, unkenntlichen Überschriften, fehlenden oder veränderten Sonderzeichen und fehlerhafter Silbentrennung kommen.

3.11.3 Bereinigung

Texte werden bereinigt, indem irrelevante Wörter, sogenannte „stopwords“[63], zunächst herausgefiltert werden. Anschließend werden die Dokumente strukturiert, indem die enthaltenden Texte in logische Einheiten, Abschnitte wie Überschrift, Kapitel, Sätze und Absätze unterteilt werden (*engl. Tokenization*). Dabei ist entscheidend, dass beispielsweise zwischen der Verwendung des Satzzeichens „.“ am Satzende, bei Aufzählungen und als Abkürzungsmarkierung unterschieden werden muss [vgl. Feldmann, Sanger: 2007, S. 60]. Schwierigkeiten treten außerdem bei mehreren zusammenhängende Sätzen auf, die wie in einer wörtlichen Rede verschachtelt sind [vgl. Heyer et al.: 2006, S. 63 ff.]. Im Gegensatz zur inhaltlich sinnvollen Satz- und Abschnittszerlegung ist die Wortformzerlegung vergleichsweise einfach, weil sich der Algorithmus an sogenannten *white spaces* (Leertaste, Tabulator oder Zeilenumbruch) orientieren kann. Probleme bereiten zusammenhängende Wortformen wie „Vor- und Nachteile“, „Beratungs-GmbH“ oder auch Groß- und Kleinschreibung wie bei „ph-Wert“, Umlaute und Zeichenmischungen wie „Audi A4“ [vgl. Heyer et al.: 2006, S. 68 ff.].

Definition von Merkmalen

Vor der Analyse werden passende Merkmale und Eigenschaften der einzelnen Elemente bestimmt. Merkmale zeichnen sich durch besonders große Zusammengehörigkeit oder Verschiedenheit aus. Jede Wortform entspricht einem Merkmal. Um die Komplexität zu verringern, wird sich nur auf besonders relevante, auffällige oder häufige Wörter beschränkt. Häufig werden zur Analyse auch nur Nomina berücksichtigt, da Verben und Adjektive das Ergebnis verschlechtern, indem sie in zu vielen unterschiedlichen Zusammenhängen auftreten. Wortvarianten, die nach grammatischen Regeln gebildet wurden, werden durch die Methode des sogenannten „Stemmings“ zurück auf den Wortstamm reduziert [vgl. Heyer et al.: 2006, S. 223 ff.]. Die daraus folgend reduzierte Wortmenge (*engl. Bag of words*) wird anschließend nach beispielsweise Häufigkeiten und Abhängigkeiten analysiert [vgl. Sharafi: 2013, S. 79-93].

[63] Stopwords sind inhaltlich irrelevante Worte wie beispielsweise „aber“, „für“, „weil“, „der“ oder „die“.

Beispiel für eine Auswahl an Angaben des Suchbegriffes „Journalist“ im Projekt Deutscher Wortschatz der Universität Leipzig

Anzahl: 6872
Häufigkeitsklasse: 11 (d.h. „der“ ist ca. 2^11 mal häufiger als das gesuchte Wort)
Sachgebiet: Presse, Motive
Grammatikangaben:
Wortart: Substantiv
Geschlecht: männlich
Flexicon: der Journalist, des Journalist, dem Journalist, den Journalist, die Journalisten, der Journalisten, den Journalisten, die Journalisten
Synonyme: Berichterstatter, Kolumnist, Korrespondent, Pressevertreter, Publizist, Referent, Reporter

Beispiel

Als die Band ihren Durchbruch erlebte, 1964, da ermunterte ihn ein **Journalist**, seine Texte als Buch erscheinen zu lassen. *(Quelle: www.come-on.de, 2010-12-25)*

Die Söhne Max (Jahrgang 1978) und Mischa (Jahrgang 1981) arbeiten als **Journalist** und als Betriebswirt. *(Quelle: www.gea.de, 2011-01-03)*

Signifikante Kookkurrenzen für Journalist:

Der (2105.94), Autor (1764.94), als (1643.05), freier (1533.32), ein (1111.92), Schriftsteller (1061.07), der (892.1), und (787.05), , (771.1), er (745.17), arbeitete (706.01), Buchautor (589.08), Buch (588.9), Ein (555.99), schreibt (444.25), Redakteur (268.98), Blomkvist (261.69), berichtete (258.43), Fotograf (258.17), hat (251.92), Tageszeitung (249.05), seine (234.22), schrieb (233.24), Historiker (231.13), « (228.64), Reporter (223.51), war (202.79), Interview (201.96), Moderator (201.4), Korrespondent (199.78), berichtet (197.92), Blogger (193.86), fragte (190.91), Mikael (190.36), Nachrichtenagentur AFP (188.07), investigativer (184.79), recherchiert (184.22), Hersh (179.91)

Signifikante linke (Wort-) Nachbarn von Journalist

Der (3487.64), als (2714.07), freier (2260.99), ein (2250.15), der (2008.14), Ein (1129.57), Als (418.95), irakische (404.85), kein (305.6), amerikanische (291.92), deutscher (285.28), freie (283.25), investigativer (275.2), junger (256.85), und (253.14), britischer (243.03), britische (236.11), gelernte (199.57), italienische (199.45), deutsche (183.63), kritischer (182.21), bekannte (176.07), investigative(147.71), ehemalige (147.31), ausländischer (147.04), libanesischer (146.14), bekannter (135.04), jeder (128.45), französischer (126.95), ehemaliger (120.84), ausgebildeter (116.23), österreichische (111.75), russische (111.56), italienischer (111.29), guter (110.26), gelernter (109.17), recherchierender (102.35), Kein (99.08), Schweizer (95.95), angehender (91.51), Freier (89.12), kanadische (88.78), österreichischer (88.56), eingebetteter (86.64), einziger (85.26), chinesischer (85.08), befreundeter (84.03), ausgezeichneter (77.05), politischer (74.26), irgendein (73.55), erfahrene (73.52), isländischer (73.45), frühere (72.94), türkischarmenische (72.53), unabhängiger (70.45), französische (68.87), hervorragender (67.29), irakischer (65.83), lebende (64.95), amerikanischer (64.71), israelische (63.65), freischaffender (62.38), polnische (61), angesehener (58.33), Jeder (57.43), renommierte (56.19), australische (56.08), bloggender (55.85), ausgezeichnete (55.55), berühmteste (55.28), geborene (54.56), israelischer (54.49), junge (52.29), norwegischer (50.83), russischer (50.71), japanischer (50.22), freiberuflicher (49.62), mancher (49.32), Kölner (48.37), engagierter (46.47)

Signifikante rechte (Wort-) Nachbarn von Journalist
und (1449.83), , (352.62), tätig (277.06), Peter (261.28), fragt (254.71), Mikael (241.52), Günter Wallraff (188.67), Seymour Hersh (185.26), John Pilger (175.81), Seymour (150.84), Muntaser (140.07),Günter (139.61), Oleg (130.64), Muntasser (126.05), Peter Scholl- Latour (119.98), Henryk (117.94), fragte (117.36), John (116.97), Hans Leyendecker (116.81), Henryk M. Broder (111.81), Daniele Mastrogiacomo (107.09), hatte (98.21), Jürgen (95.84), Michail (90.55), Hrant Dink (88.79), Hans (88.58), Hrant (87.86), Stefan (85.53), Faiza (83.58), Roland Jahn (83.25), Stefan Aust (81.33), Hajo Schumacher (80.59), Tiziano Terzani (79.33), Stefan Niggemeier (75.64), Georg Diez (75.12), Jürgen Roth (73.07), bei (73.04), Tiziano (70.63), schreibt (69.88), Thomas (69.28), war (67.2), Gideon Levy (66.6), Néji (66.44), Steffan Heuer (66.44), Michael (66.08), Daniele (64.45), Schalwa (62.73), wissen (62.71), Andreas (62.5), Paul (62.45), Adrien (62.43), Rob Savelberg (61.94), Gerhard Kromschröder (61.94), Martin (61.62), Gao Yu (59.71), getötet (57.09), Richard Kiessler (56.88), gearbeitet (56.72), Hajo (56.54), Ulrich (54.98), David Ignatius (54.98), Walter (54.57), Roberto Saviano (54.44), Montasser (54.22), Martin Bashir (54.22), Mumia (53.79), Wachtang (53.18), aus (53.08), Cal (52.61), Gerhard (52.52), Tom (52.08), Sergej (51.11), Jeffrey (50.25), Gianluigi (50.13), Gerhard Wisnewski(49.21), Sandor (48.94), Declan Hill (48.91), sollte (48.6), Richard (48.33), Harald (47.63)

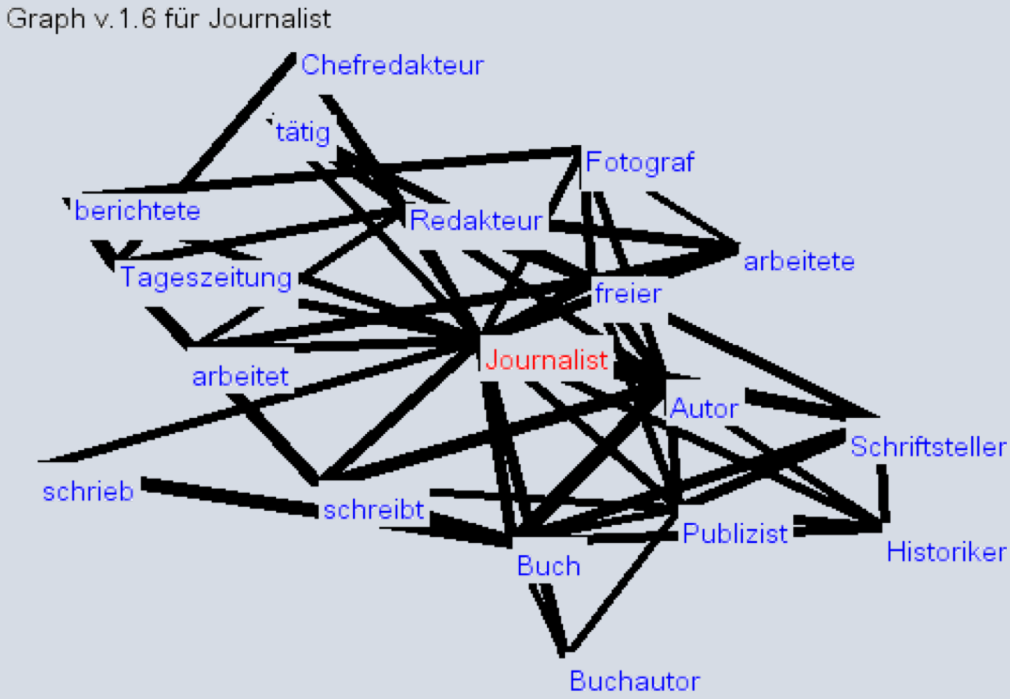


Abbildung 21: Graph zum Stichwort „Journalist“ *[vgl. Deutscher Wortschatz: 1998-2011]*

3.11.4 Analyse

Text Mining basiert auf dem Prinzip des linguistischen Strukturalismus, der Ende des 19. Jahrhunderts zur Beschreibung von fremden Sprachen entwickelt wurde. Eine automatische Wissensextraktion aus Texten ist mit dem Erlernen einer neuen Sprache vergleichbar. Dabei dienen folgende vier hierarchische Ebenen aus beispielsweise Wortformen, Sätzen und Phrasen der automatischen Ermittlung von semantischen Zusammenhängen.

- **Syntagmatische Relationen** [vgl. Heyer et al.: 2006, S. 20, 23 ff., 134 ff.]
  In syntagmatische Relationen stehen zwei zusammen auftretende Wortformen eines Satzes. Im Beispielsatz „Der Hund bellt" stehen die Wortformen „Der" und „bellt" im lokalen Kontext der Wortform „Hund". Das gemeinsame Auftreten zweier Wortformen (wie z.B. Nomen und passendem Verb) in mindestens einem lokalen Kontext nennt man Kookkurrenz bzw. die Wortformen jeweils Kookkurrenten. Um eine Zufälligkeit auszuschließen, wird ein sogenanntes Signifikanzmaß festgelegt, das die Relation bestimmt. Auffällig häufig gemeinsam auftretende Wortformen (wie „Los Angeles", „Romeo & Julia" oder „schwere Krankheit") werden signifikante Kookkurrenten genannt. Sie treten vielfach auf Satzebene als feste Redewendungen („mit freundlichen Grüßen"), Aufzählungen („rot, gelb, blau...") und Dependenzen („Blumen blühen", „Auto fährt") auf. Unter sogenannter Nachbarschaftskookkurrenz werden Mehrwortbegriffe („Vorsitzender des Vereines"), Kategorie- und Funktionsangaben („Stadt Hamburg", „Dr. Grunart") oder sogenannte „Head-Modifier"-Relations („vorgezogenes Rentenalter") verstanden.

- **Paradigmatische Relationen** [vgl. Heyer et al.: 2006, S. 20, 26 ff.]
  Die Ähnlichkeit zweier gemeinsam auftretender Wortformen in ähnlichen Kontexten aus allen Sätzen, sogenannte globale Kontexte, stehen in paradigmatischer Relation. Ein globaler Kontext besteht aus allen Wortformen, die in syntagmatischer Relation stehen. Es lässt sich daher die paradigmatische Relation aus der syntagmatischer Relation ableiten. Zwei globale Kontexte werden mit einem Ähnlichkeitsmaß und festgelegten Schwellenwert berechnet. Dazu wird ein Vergleichsprädikat verwendet. Die Tanimoto-Ähnlichkeit stellt die Gesamtzahl aller Wortformen der vergleichbaren Texte in ein Verhältnis zueinander. Je größer das Wortformvorkommen, umso ähnlicher. Darüber hinaus hilft das Cluster-Verfahren, Ähnlichkeiten festzustellen.

- **Semantische Relationen** [vgl. Heyer et al.: 2006, S. 20, 30 ff.]
  In semantischer Relation stehen Wortform-Paare, die mit der linken und rechten Nachbar-Wortform zusammenhängen. Als Vorrausetzung gilt, dass die Wortform-Paare bereits in syntagmatischer und paradigmatischer Relation stehen. Zu häufigen Wortform-Paaren zählen insbesondere Nachbarschaftskookkurrenten wie Maßeinheiten, Kategorie- und Funktionsangaben, Qualifizierungsangaben oder Muster von Ausdrücken wie Ober- und Unterbegriffe, Synonyme, „Ursache-von"-Beziehung, „Instrument-für"-Beziehung, „Teil-von"-Beziehung. Zu den vier wichtigsten semantisch relevanten syntaktischen Mustern zählen:

- **Nomen und Eigenname** („Stadt Hamburg“, „Kanzlerin Merkel“)
- **Nomen und Nomen** („Liter Milch“, „Gigabyte Daten“)
- **Adjektiv und Nomen** („geschlossene Tür“, „rote Ampel“)
- **Nomen und Verb** („Bier trinken“, „Maler streicht“)

- **Logische Relationen** [vgl. Heyer et al.: 2006, S. 20, 39 ff.]
  Logische Relationen, auch als Sinnrelationen bezeichnet, sind semantische Relationen, die logische Folgerungen zulassen.

  - **Ober- und Unterbegriffe**
    (Vogel ist der Oberbegriff für Amsel, und Vögel gehören wiederum zu den Wirbeltieren. Wenn nun Wirbeltiere Schnäbel haben, dann hat die Amsel auch ein Schnabel.) Der Oberbegriff A von B gilt, wenn die Extension von B eine Teilmenge von A ist. Ein Oberbegriff mit mehreren Unterbegriffen wird Kohynome genannt.

  - **Synonyme**
    Zwei Begriffe können die gleiche Extension haben wie beispielsweise TV-Gerät und Fernseher oder Aufzug und Fahrstuhl.

  - **Gegensätze**
    Zwei Wortformen („Hund“ und „Katze“) haben keine gemeinsame Schnittmenge der Extensionen, aber einen gemeinsamen Oberbegriff („Haustier“). Spezialfälle gegensätzlicher Begriffe bilden Komplementärbegriffe und Antonyme. Es handelt sich um Komplementärbegriffe, wenn das Komplement „Frau“ von dem Begriff „Mann“ auch das Komplement der Extension des Begriffes „Mann“ ist. Antonyme, auch als relative Gegensätze bekannt, sind Gegensätze mit vordefiniertem Bezug der vorausgesetzten Extension. Beispielsweise ist „dumm“ das Antonym zu dem Begriff „intelligent“, wenn der Bezug auf „Schachspiel“ liegt.
  - **Konverse**
    Konverse beschreiben die inhaltliche Relation von Begriffspaaren nach dem Prinzip: Wer etwas mietet, dem muss auch etwas vermietet werden.

3.11.5 Statistische Methoden

Nach einer erfolgreichen Datenvorbereitung werden wie im Data-Mining-Prozess mit unterschiedlichen Verfahren nach Mustern und Zusammenhängen gesucht, Trends analysiert und visualisiert. Die Sprachstatistik untersucht Häufigkeiten, bedingte Abhängig- und Wahrscheinlichkeiten sowie signifikante Kookkurrenzen. Allgemein ergeben sich Sprachmodelle aus sprachlichen Ereignissen, denen Wahrscheinlichkeiten durch statistische Sprachverarbeitung zugeordnet werden

können [vgl. Heyer et al.: 2006, S. 87 ff.]**.** Im Folgenden werden einige statistische Analysemethoden vorgestellt.

- **Zipfsches Gesetz** [vgl. Heyer et al.: 2006, S. 87 ff.]
  Häufigkeiten und Rangordnungen werden auf der Grundlage der Annahme berechnet, dass die Anwendung natürlicher Sprache unter dem geringsten Aufwand verwendet wird. Das Projekt „Deutscher Wortschatz" der Universität Leipzig fand unter den häufigsten zehn Wörtern ihres Textkorpus lediglich kurze Funktionswörter (der, die, und, in, den, von, zu, das, mit, sich).

- **Differenzanalyse** [vgl. Heyer et al.: 2006, S. 95 ff.].
  Dieses Verfahren dient der Einordnung zu Sachgebieten, der Verschlagwortung sowie der Terminologieextraktion mithilfe eines allgemein sprachlichen Referenztextkorpus zum Beispiel aus Zeitungstexten. Die Wortformen werden dazu nach vier Häufigkeitsklassen zugeordnet. Fachausdrücke kommen beispielsweise nur wenig im Referenzkorpus vor, stopwords dagegen sehr häufig.

- **POS-Tagging**
  Durch das POS-Tagging (*engl. Part of Speech*) werden die Wörter nach syntaktischen Informationen kategorisiert und in Wortarten zugeordnet. Üblicherweise wird unter anderem nach Nomen, Verben, Adjektiven und Präpositionen unterschieden. Je nach Anwendungsfall ergeben sich unterschiedliche Schwerpunkte. So geben beispielsweise Adjektive Indizien über Emotionen oder Bewertungen in einem Text [vgl. Feldmann, Sanger: 2007, S. 60].

  Beispiele für POS Tags:

[ART]	Artikel
[KON]	Konjunktion
[NE]	Eigenname
[NEG]	Negation
[NN]	normales Nomen
[PR]	Präposition
[PREL]	Relativpronomen
[VFIN]	finites Verb
[$]	Satzzeichen

  Beispiel für die Zerlegung eines Satzes in POS-Tags
  Lena[NE] gibt[VFIN] Stefan[NE] das[ART] Buch[NN].[$]

- **Probabilistisches Sprachmodell** [vgl. Heyer et al.: 2006, S. 100 ff.]
  Mithilfe eines Trainingskorpus werden die Häufigkeiten von Wortkombinationen geprüft, um Aussagen über die Wahrscheinlichkeit zur Zugehörigkeit einer Sprache treffen zu können. Dieses Verfahren wird vermehrt bei Rechtschreibprogrammen und Übersetzungen angewendet. Dazu

werden Wortformen, die im lokalen Kontext zueinanderstehen, in maximal Dreierkombinationen (Phrasen) berücksichtigt, da ihre Beziehungen meistens lokal geprägt sind. Mehr als drei Worte machen die Analyse komplexer und nicht genauer. Es werden Tri- bzw. Bi-gramme und Unigramme verwendet, um Wahrscheinlichkeiten zu errechnen, die durch Gewichtungsfaktoren (*engl. smoothing*) ergänzt werden können.

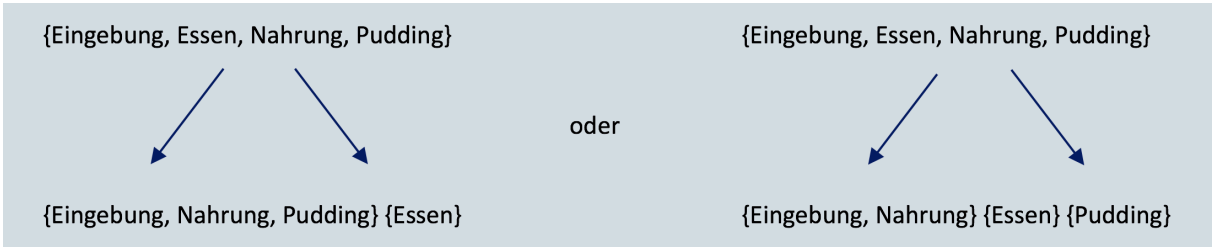
- **Markov-Sprachmodell** [vgl. Heyer et al.: 2006, S. 115 ff.]
  Diese Sprachmodelle sind laut Heyer et al. [vgl. 2006, S. 115 ff.] mit endlichen Automaten zu vergleichen. Dabei gibt es ein Start- und Endzustand sowie Zustandsübergänge, die mit Wahrscheinlichkeiten definiert werden und mehrere Folgezustände haben können.

3.11.6 Klassenbildung

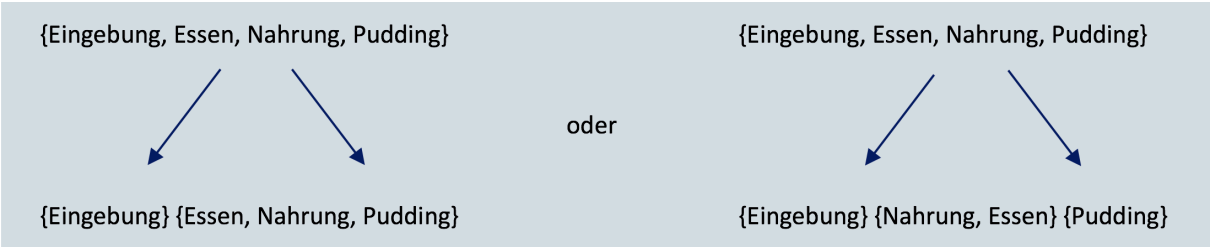
Die Klassenbildung unstrukturierter Daten bezeichnet die Ableitung der inhaltlichen Struktur und die Einteilung in homogene Teilmengen (Cluster) durch vorgegebene Kriterien [vgl. Heyer et al.: 2006, S. 195 ff.]. Clusterverfahren dienen der Ermittlung semantisch ähnlicher Texte und Wörter.

Ziel ist es, anhand eines Ähnlichkeitsmaßes eine möglichst große Homogenität innerhalb der Cluster und einer Heterogenität der einzelnen Cluster zueinander zu bilden (siehe auch Kapitel 3.9.2: Klassenbildung). Dokumente lassen sich beispielsweise nach Länge oder Fachbegriffen clustern. Wörter können nach ihrer gleichen Endung oder der semantischen Nähe eingeteilt werden.

Beispiele von Wortform-Clusters nach Endungen [vgl. Heyer et al.: 2006, S. 197 ff.]



Beispiele von Wortform-Clusters nach semantischer Nähe [vgl. Heyer et al.: 2006, S. 198 ff.].



Wie in der Clusteranalyse des Data-Mining-Verfahrens bereits beschrieben, kann hierarchisch und nicht hierarchisch vorgegangen werden. Es wird zwischen „hart“ (nur ein Element pro Cluster) und „soft“ (ein Element kann mehreren Clustern zugeordnet sein) unterschieden. Die Clusteranalyse für Dokumente geht nach einzelnen Schritte wie folgt vor [vgl. Heyer et al.: 2006, S. 202 ff.]:

1. **Identifikation der charakteristischen Merkmale**
2. **Erstellung der Dokumentvektoren**
   (Grundlage ist die Häufigkeit der Wortformen [Wortformfrequenz] mit Einbeziehung der Textlänge. Die Gewichtung ergibt eine sogenannte Term-Term-Matrix.)
3. **Auswahl des Ähnlichkeitsmaßes**
   (Berechnung z.B. mittels euklidischer Distanz, dem Skalarprodukt oder dem Cosinus-Winkel der Dokumentvektoren.)
4. **Erstellung einer Ähnlichkeitsmatrix**
   (Die Ähnlichkeiten zweier Dokumente wird Dokument-Dokument-Matrix genannt.)
5. **Clusteranalyse**
   (Die Arbeitsschritte zum Clustern von Wortformen sind ähnlich dem Dokument-Clusterverfahren. Es wird allerdings auf Textabschnitte bzw. einzelne Sätze beschränkt, da hier ein Themenwechsel unwahrscheinlich ist. Häufig vorkommende Wortformen sind thematisch schwer einzugrenzen.)

3.11.7 Musteranalyse

Die Musteranalyse sucht nach Mustern in unstrukturierten oder annotierten Texten mithilfe von regulären Ausdrücken[64] [vgl. Heyer et al.: 2006, S. 227 ff.]. Diese werden mit Atomen (Buchstaben und Wortformen) und Operatoren, die der Verknüpfungen der Atome nach grammatischen Regeln dienen, beschrieben. Der zusätzliche Einsatz von Sonderzeichen, sogenannten Wildcards, ermöglicht eine umfangreiche Suche nach Mustern in Texten. Der Asterisk (*) bedeutet beispielsweise die Suche nach dem vorangegangenen Ausdruck [vgl. Heyer et al.: 2006, S. 229 ff.]. Mit dem Ausdruck „un.*“ wird nach allen vorkommenden Wortformen gesucht, die „un“ beinhalten. Sollen nur Wortformen gesucht werden, die mit „un“ anfangen, so gilt für die Suche der Ausdruck „ un.*“. Nach einem Datum in Form von tt.mm.yyyy wird beispielsweise mit [0-9] {2} [/\.] [0-9] {2} [/\.] [0-9] {4} gesucht.

[64] Reguläre Ausdrücke beschreiben eine „Kunstsprache“, die aus festen Ausdrücken (Buchstaben und Wortformen) besteht und auf der Arbeit des Mathematikers Stephen Kleene basieren.

Übersicht der Operatoren [vgl. Heyer et al.: 2006, S. 229]

() **Gruppierungsoperator**
Zusammenfassung von „q“ und „u“ durch (qu).

[] **Zeichenbereichsoperator**
[qu] bedeutet entweder „q“ oder „u“. Mit [q-u] kann das Zeichen
„q“, „r“, „s“, „t“ oder „u“ sein.

{} Iterationsoperator
q{3} bedeutet das “q“ dreimal hintereinander vorkommt.

| Oder-Operator
(qu)|(ph) steht für entweder „qu“ oder „ph“.

^ Anfangs-Operator
^qu bedeutet, dass der Suchbereich mit „qu“ anfängt.

$ Ende-Operator
$qu definiert das Ende des Suchbereichs „qu“.

[^] Negationsoperator
[^qu] bedeutet ein beliebiges Zeichen außer „q“ und „u“.

Reguläre Ausdrücke helfen auch bei der Suche nach syntaktischen Satzstrukturen wie Redewendungen oder Ober- und Unterbegriffen. Die Textvorbereitung mittels POS-Tagging ist hier sinnvoll. So ist die Suche nach allen Eigennamen eines Textdokumentes durch *[NE] *[NE] möglich. Wenn nicht sicher ist, in welcher Form ein bestimmter Artikel und ein Nomen der Redewendung „einen Streit vom Zaun brechen“ vorkommt, so hilft die Suche nach “[ART] [NN] vom Zaun“. Morphenmuster im Speziellen bezeichnen Wortformen einer Fachsprache z.B. aus der Medizin, die sich durch viele lateinische Ausdrücke und häufig vorkommende Suffixe wie „itis“ auszeichnet. Hier erleichtert eine Grundformreduktion auf den semantisch identischen Stamm die Arbeit.

Möglichkeiten für die Zerlegung des Satzes: „Die Masterthesis behandelt verschiedene Bereiche im Datenjournalismus“ [vgl. Heyer et al.: 2006, S. 324].

Buchstaben
D-i-e-M-a-s-t-e-r-t-h-e-s-i-s-b-e-h-a-n-d-e-l-t-v-e-r-s-c-h-i-e-d-e-n-e-B-e-r-e-i-c-h-e-i-m-D-a-t-e-n-j-o-u-r-n-a-l-i-s-m-u-s

Morpheme
Die-Master-thesis-be-hand-elt-ver-schiedene-Be-reiche-im-Daten-journalis-mus

Wortformen
Die-Masterthesis-behandelt-verschiedene-Bereiche-im-Datenjournalismus

Phrasen
Die Masterthesis - behandelt verschiedene Bereiche- im Datenjournalismus

POS-Tagging
Die [ART] Masterthesis [NN] behandelt [VFIN] verschiedene [ADJA] Bereiche [NN] im [APPRAT] Datenj0urnalismus [NE]

3.11.8 Beispiele für die Anwendung von Text Mining

Überwachung verdächtiger Blogs

Anlässlich des 30. Chaos Communication Congress hielt der Linguist Joachim Scharloth (TU Dresden) einen Vortrag (30C3) „Überwachen und Sprache” [Scharloth, 2013], um computerlinguistische Vorgehensweisen der NSA zu beschreiben. Um die Gefahr ausgewählter, möglicher terroristischer Gruppen im Internet (z.B. Blogs) zu analysieren, wird zunächst zur automatischen Identifizierung von Entitäten das Verfahren „Named-Entity-Recognition” (NER)[65] angewendet. Named Entities (NE) sind Ausdrücke, die eine Entität eindeutig von Entitäten mit ähnlichen Attributen unterscheidet. Anschließende Entitätsklassen können aus Personen, Orten oder Organisationen gebildet sein, wobei entschieden werden muss, ob beispielsweise Bundestag ein Ort oder eine Organisation ist.

[65] Mit dem „Named Entity Recognition“-Verfahren werden Eigennamen extrahiert, um sie anschließend semantisch zu klassifizieren [vgl. Reznicek: 2013]. Dies geschieht nach vier Hauptklassen: Personen [PER], Orte [LOC], Organisationen [ORG] und Andere [OTH]. Am Beispiel [SV Werder Bremen] ist „Bremen“ ein Ort, aber „Werder“ und „Werder Bremen“ zusammen eine Organisation. [Max Müller] ist eine Person und [Balu] ein Eigenname. Schwierigkeiten gibt es bei [Hartz-Reformen] oder [Mercedes-Millionen].

Mithilfe der Entitäten lassen sich Kollokationen[66] berechnen, indem mit fünf Wörtern davor und danach bis zum Satzanfang und Ende der Kontext untersucht wird. Diese Wörter werden „Keywords in Kontext" (KWIC) genannt, und es wird die relative Frequenz zum Gesamttext verglichen. Wenn ein Wort häufiger als „Nachbarschaftswort" auftaucht, besteht eine Kollokation zum Schlüsselwort. Mithilfe eines Trainingskorpus können typische Entitäten und Kontexte erlernt werden und auf Texte verdächtiger Gruppen aus zum Beispiel Blogs oder Portalen angewendet werden. Die Radikalität lässt sich anschließend auf unterschiedliche Weise berechnen. Der listenbasierte Ansatz beispielsweise gleicht Texte mit einer Liste aus bekannten Vokabeln ab, die als Radikalisierungs-Indikator vordefiniert wurden. Das können skandalisierte Wörter wie Irrwitz, Dummheit oder Frechheit oder negative Adjektive wie abscheulich, abnorm oder absurd sein. Auch die Messung semantischer Taxonomien (Relationen der Wörter zueinander) wie bei vielen polaren Antonymen (z.B. wahr oder falsch, tot oder lebendig) können auf ein Schwarz-Weiß-Bild deuten. Die Messung von Gradpartikeln kann zusätzlich die Emotionen des Autors einstufen und als Indikator dienen. Dabei handelt es sich um häufig auftretende Wörter wie „absolut", „gründlich", „restlos", „total", „äußerst", „stark", oder „besonders".

GeoCollocation

In seiner Arbeit zu GeoCollocations beschreibt der Linguist (TU Dresden) Noah Bubenhofer die Vorgehensweise zur Berechnung musterhafter Attributionen zu georeferenzierbaren Orten [vgl. Bubenhofer: 2014]. Dazu unterscheidet er zunächst zwischen der „Corpus based"-Analyse (hypothesengeleitete Analyse, Klassifikationsverfahren), die der generellen Überprüfung, der Suche nach Phänomenen und der Beantwortung klassischer Fragen nach dem Wo, Wie und Wie oft dient. Die "Corpus driven"-Analyse (datengeleitete Analyse, Clusterverfahren) dagegen macht Strukturen erst sichtbar, um sie anschließend zu klassifizieren. In so einem Korpus werden alle möglichen Zeichenkonfigurationen zunächst berechnet, anstatt der Überprüfung einer Hypothese nach festgelegten Analysekategorien. Aufgrund der vielen Textdokumente verwendet Bubenhofer daher für sein Beispiel ein datengeleitetes Verfahren. Er verweist in diesem Zusammenhang auch auf einen hohen Bedarf an neuen Visualisierungsformen im Bereich der Korpuslinguistik, insbesondere an Techniken zur explorativen Datenvisualisierung.

Bubenhofer verwendete für einen Vergleich zwei Korpora: Zum einen aus Nachrichtenartikeln von „Zeit Online" und zum anderen aus Parlamentsprotokollen von Bund und Ländern bestehend. Beide Korpora wurden mit POS-Tags tokenisiert sowie zusätzlich mit Lemma- (Grundwortform) und Wortklasseninformationen versehen. Er verwendete die Software „Stanfort Named Entity Recognizer", um die Eigennamen, genauer die Toponyme, annotieren zu können. Häufige Fehlerquellen waren dabei nicht erkannte Eigennamen oder fasch erkannte Wortformen. Zu den

[66] Kollokationen sind statistisch überzufällig häufige Wortkombinationen [vgl. Evert: 2008, S. 1214]: „We define a collocation of two words that exhibit a tendency to occur near each other in natural language, i.e. to coocur." Kollokationen sind wiederkehrend und vorhersehbar.

Toponymen wurden die signifikanten Kollokationen anschließend berechnet und aufgrund der Semantik nach Wortklassen gefiltert, um häufig auftretende Orte in Verbindung mit Themen wie etwa „Frankfurt“ und „Börse“ zu finden.[67] Alle Toponyme wurden für die anschließende Visualisierung der Ergebnisse geocodiert und mit Informationen zum Längen- und Breitenmaß versehen. Die fertigen sogenannten GeoCollocations verfügten nun über Informationen über Häufigkeiten, Signifikanz der Kollokationen, präzise Angaben zum Toponym-Typen wie Ortschaft, Stadt, Kontinent, Region, Staat.

In der anschließenden Visualisierung wurde ersichtlich, dass die Berichterstattung beider Korpora völlig unterschiedlich gewichtet war (unterschiedliche Kollokationen). So wurde auf politischer Ebene über Indien in wirtschaftlichen Zusammenhängen berichtet, während in den journalistischen Texten die Menschenrechte häufiger thematisiert wurden. Es zeigte außerdem die am häufigsten erwähnten Orte sowie die lokalen und globalen Kollokationen. Interessant war, dass unter den sehr häufigen Kollokationen, einige mit sehr entfernten Toponymen („Kuba“ oder „Nordkorea“ und „Frachter“, „Drogen“, „Verdacht“ oder „veraltet“) genauso wie mit zwei auffällig zeitlich oder räumlich nahen Typonymen („Leipzig“ oder „Bad Hersfeld“ und „streiken“, „Amazon“) vorkamen. Es konnte festgestellt werden, dass Kollokationen mit vielen unterschiedlichen Toponymen eher eine allgemeine Aussagekraft haben. Im Gegensatz dazu, wiesen Kollokatoren mit einzelnen oder wenigen Toponymen auf eine vielfältige Berichterstattung einzelner Ereignisse hin und veranschaulichten den Unterschied zwischen globalen und lokalen Ereignisse. Häufige Sachkollokationen waren zum Beispiel Krieg – Gewalt und Opposition – Präsident.

Text-Datenvisualisierung

Verschiedene Strukturen, Inhalte und die Gewichtung von Texten können beispielsweise durch Größenverhältnisse von Wörtern dargestellt werden. Mit der sogenannten WordClouds lassen sich Häufigkeiten von Wörtern in einem Text visualisieren. Dabei repräsentiert die Schriftgröße die Signifikanz. Es wäre außerdem denkbar, Themen beispielsweise durch Farben hervorzuheben.

Abbildung 22: Wordcloud der Masterarbeit *[eigenhändig online unter Wordclouds.com erstellt]*

[67] Eine Zusammenstellung semantischer Klassifikationen oder Ontologien der deutschen Sprache können unter anderem bei z.B. GermaNet, oder bei Dornseiffs Wortschatz nach Sachgruppen geordnet, abgefragt werden.

3.12 Web Mining

Unter Web Mining wird die Datenanalyse aus Quellen des Internets verstanden (siehe auch Kapitel 3.5.1: Datenbeschaffung aus dem Web). Auf der einen Seite beinhaltet die Internetdatenanalyse das *Web Content Mining*. Damit ist die Extraktion von Informationen aus multimedialen Inhalten unterschiedlichen Formats und Verbindungen im Internet gemeint. Ein weiterer Bereich umfasst das *Web Usage Mining*, das die Verhaltensweisen der Internetnutzer analysiert. Dabei wird nochmals unter *Web Log Mining* (Protokolldaten der Nutzeranwendungen) und *Integrated Web Usage Mining* (Protokolldaten in Kombination mit weiteren Nutzerdaten aus anderen Quellen) unterschieden [vgl. Cleve, Lämmel: 2016, S. 66-67]. Diese Arbeit beschäftigt sich schwerpunktmäßig mit Text und Data Mining. Da Web Mining eine Kombination aus beiden Bereichen ist, die strukturierte und unstrukturierte Daten analysiert, wird das Verfahren in dieser Arbeit nicht weiter ausgeführt.

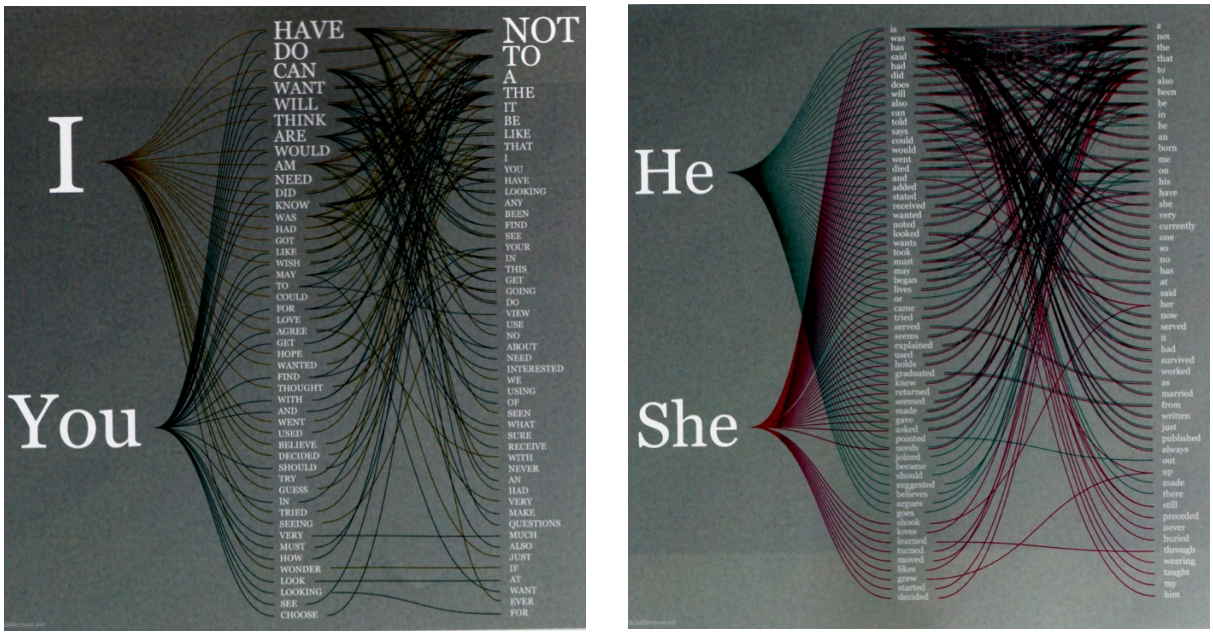


Abbildung 23: Web Triagramm *[Chris Harrison: 2006]*
Die Grafik zeigt die Google-Fundstellen im Google-Index. Die Häufigkeit der Suchphrasen aus drei Wörtern ist nach der Häufigkeit absteigend visualisiert. Es zeigt, dass beispielsweise nur „He" in Verbindung mit „argue" gesucht wurde und „She" nur mit „loves".

3.13 Arbeitswerkzeuge für Datenjournalisten

> *„'Data journalism' only differs from 'word journalism' in that we use a different kit. We all sniff out, report, and relate stories for a living. It's like 'photo journalism; just swap the camera for a laptop."*
>
> *Brian Boyer, Chicago Tribune [Gray et al.: 2012, S. 15]*

In der Praxis arbeiten an datenjournalistischen Projekten oft interdisziplinäre Teams, die mindestens aus Journalisten, Programmierern und Grafikern bestehen. Die unterschiedlichen Denkansätze und Herangehensweisen sind oft sehr konstruktiv und kreativ, weil die unterschiedlichen Perspektiven und der Austausch untereinander Projekte bereichern. Doch eine derartige Teamzusammensetzung, die Vereinigung unterschiedlichster Fachkompetenzen, erfordert eine gute Kommunikation und viel gegenseitiges Verständnis.

Als große Herausforderung für inhomogene Gruppen gilt es daher, zunächst eine grundlegend verständliche Sprache und Basis untereinander zu finden. Wie bereits beschrieben, sollten sich dafür die Entwickler grundlegend journalistische Kenntnisse und Datenjournalisten zunächst ein technisches Grundverständnis aneignen [vgl. Stencel et al.: 2014, S.11]. Neben der klassischen Arbeit in Contentmanagementsystemen können Journalisten auf immer mehr technische Softwareanwendungen für das Sammeln, Organisieren, Speichern, Analysieren, Visualisieren, Veröffentlichen und Teilen von Daten zurückgreifen, die keine besonderen Programmierkenntnisse erfordern.

> *"We try to pick tools that anyone could get the hang of without learning a programming language or having special training and without a hefty fee attached. We're currently using Google products quite heavily for this reason."*
>
> *Lisa Evans, the Guardian [Gray et al.: 2012, S. 131]*

Im Folgenden wird zunächst die Open-Source-Bewegung vorgestellt, deren Grundlagen für alle praktizierenden Berufe bedeutsam sind, die auf unterschiedliche Softwarelösungen angewiesen sind. Anschließend werden Beispiele für Softwareanwendungen in den einzelnen Arbeitsschritten gegeben.

3.13.1 Open Source

Für die journalistische Datenrecherche, Analyse und Visualisierung sind kostenlose Softwareanwendungen nach dem Open-Source-Gedanken besonders aus wirtschaftlicher Sicht nützlich, sie spielen aber auch im Umgang mit der anschließenden Publikation eine wichtige Rolle. Zum einen sollten die recherchierten Datensätze für jedermann nachvollziehbar und auch reproduzierbar sein. Darüber hinaus ist die Offenlegung der Daten ein Qualitätsmerkmal und gleichzeitig die Chance

für andere, sich zu beteiligen und weitere Geschichten zu finden oder weiterzuentwickeln. Gerade in Zeiten, wo in den Medien viele Einsparungen stattfinden, greifen Redaktionen im Datenjournalismus auf zahlreiche Open-Source-Softwareanwendungen zurück. Es spart Kosten und ermöglicht Experimente.

Open Source (*engl. für offene Quelle*) bezeichnet eine frei nutzbare Software, dessen Quellcode offen und verfügbar ist. Brasilien gilt beispielsweise als Vorbild bei Open-Source-Software-Anwendungen in der öffentlichen Verwaltung. Claudio Prado, Visionär und Vertreter des brasilianischen Kulturministeriums [Prado: 2007]: „Open Source ist die Möglichkeit, technisch autonom zu werden." Bis in die siebziger Jahre war der Quellcode jeder Software frei zugänglich. Danach erkannte man das gewinnbringende Potenzial und hütete den Quellcode als neues Geschäftsgeheimnis. Software wurde proprietär und als geistiges Eigentum mit Lizenzen gehandelt [vgl. Deterding: 2007, S. 2]. Der Programmierer Richard Stallman gilt 1985 als Gründer der Gegenbewegung Free Software Foundation (8FSF) und damit der Open-Source-Bewegung. In Hinblick auf die wachsende gesellschaftliche Abhängigkeit von Computern setzt sich die Stiftung für die Rechte der Nutzer ein. Zudem vergibt die FSF nach festgelegten Standards und Kriterien verschiedene Lizenzen für freie Software**.** Es wird die Beschränkung und Überwachung von einzelnen Unternehmen und Regierungen kritisiert:

> ***„Unsere Mission ist die Freiheit zu bewahren, zu schützen und zu fördern, um Rechnersoftware nutzen, untersuchen, kopieren, modifizieren und weiterverbreiten zu können und die Rechte von Freie-Software-Nutzern zu verteidigen."***
>
> *[Free Software Foundation: 2016]*

Der Transparenzgedanke findet in der Open-Source-Bewegung seinen Ursprung und gilt als wegweisend für weitere Offenheitskonzepte im informationstechnologischen Bereich wie beispielsweise Open Content, Open Access, Open Education und Open Data. Es handelt sich mittlerweile um eine weltweite Bewegung mit sozialen, politischen und ethischen Zielen, Wissen wie bei Wikipedia kollektiv zu sammeln und frei zugänglich zu machen. Tim O´Reilly [vgl. 2007, S.24] beschreibt in seinem Buch "What is Web 2.0" die Bewegung als *„eine Instanz der kollektiven, vernetzten Intelligenz".* Damit stehen sie stets in Konflikt mit Verfechtern des Urheberrechts und der Wirtschaft, so dass Ende der neunziger Jahre die Open-Source-Initiative entstand. Diese legt den Fokus auf den praktischen Nutzen von Open-Source-Anwendungen für alle Bereiche der Gesellschaft gleichermaßen. Um Innovationen zu fördern, sollen Nutzer wie auch Firmen von gemeinschaftlichen Softwareentwicklungen und freien Softwarelösungen profitieren [vgl. Deterding: 2007, S. 2]. Die Unterstützung der Open-Source-Bewegung ist sicherlich umstritten. Aus der Perspektive der Datenjournalisten ist die Gesamtheit der Open-Source-Bewegungen aufgrund ihrer Arbeitsanwendungen und Wirtschaftlichkeit zu unterstützen. Gleichzeit tragen Journalisten in ihrem eigenen Bereich aber auch noch den Kampf um kostenlose Inhalte und geistiges Eigentum aus.

3.13.2 Übersicht über aktuelle Softwareanwendungen aus der Praxis

> ***„My go-to tool is Excel, which can handle the majority of CAR problems and has the advantages of being easy to learn and available to most reporters. When I need to merge tables, I typically use Access, but then export the merged table back into Excel for further work.“***
>
> *Steve Doig, Walter Cronkite School of Journalism [Gray et al.: 2012, S. 133]*

Hilfestellungen und Anleitungen zu datenjournalistischer Arbeit sind u.a. unter Datadrivenjournalism.net und im „Digital Tool Catalog“ des Poynter Institute zu finden. Weiterbildungsseminare bietet auch die Non-Profit-Organisation IRE. Im Folgenden werden gängige Anwendungen aus der Praxis aufgezeigt, die für die einzelnen Arbeitsschritte nützlich sind, sie erleichtern und die Redaktionen derzeit unterstützen, datenjournalistische Projekte umzusetzen [vgl. Gray et al.: 2012, S.131ff.].

Softwareanwendungen zur Selektion, Datenvorbereitung, Transformation und Organisation

- **Google Media Tools**
  Google Docs, Sheets, Slides und Forms sind webbasierte Onlinedienste von Google für die Bearbeitung und Analyse großer Datensätze, Erstellung von Karten und Charts, Präsentationen, Tabellen zum Mischen und Zusammenstellen von Daten.
- **Libre Office**
  Programmpaket zur Textverarbeitung, Tabellenkalkulation, Präsentation und zum Erstellen von Zeichnungen ähnlich wie Microsoft Office. Zusätzlich enthält es ein Datenbankmanagementsystem und einen Formeleditor.
- **Excel**
  Umfangreiches und über alle Branchen beliebtes Tabellenkalkulationsprogramm.
- **Google Refine, Trifacta Wrangler**
  Tools zur Datenbereinigung und Transformation.
- **MySQL, SQLite**
  Relationale Datenbankverwaltungssysteme zur Speicherung, Verwaltung und Auswertung von großen Datensätzen.
- **The Overview Project**
  Durchsehen, organisieren, kategorisieren von vielen Textdokumenten und Importfunktion in DocCloud.

Visualisierung

- **Google Fusion Tables, Tableau, Many Eyes, Dipity, Google Gapminder, Chartbuilder**
  Analyse von Datensätzen durch die Erstellung von Charts und interaktiver Visualisierungen ohne Programmierung und mit Möglichkeit zur Einbindung in die Website.
- **d3.js**
  Visualisierungsbibliothek zur dynamischen Dateneinbindung in Visualisierungen.
- **GRASS**[68] **GIS**[69]
  Geografisches Informationssystem für räumliches Datamining, das Visualisierungen und Bildverarbeitungen auf Basis von räumlich-zeitlichen Daten mithilfe von Rastern und Vektoren ermöglicht.
- **Timeline JS**
  Generiert schnell und einfach interaktive Timelines, die in die Website integrierbar sind. Kann verschiedene Medien wie Tweets, Maps oder digitale Videos gruppieren.
- **Carto, Google Maps, ArcGis**
- Datenanalyse- und Visualisierungstools, die ermöglichen, Vorhersagen aufgrund von Geodaten zu machen.

Datamining

- **R, SPSS, Pandas, RapidMiner**
  Statistik-Programme für die Analyse von Datensätzen.
- **Knime**
  Graphische Analyse-Software

Veröffentlichung

- **Google Spreasheet, Document Cloud**
  Verwalten und Veröffentlichen von Dokumenten zur Analyse, Kommentierung oder Teilen.
- **Git**
  Sourcecode-Management

[68] GRASS – Resources Analysis Support System

[69] GIS – Geographical Information System

4 Datenjournalistische Arbeit am Beispiel

„Data journalism is the practice of finding stories in numbers and using numbers to tell stories."
Meredith Broussard, assistant professor of journalism at Temple University [Howard:2014, S. 5]

4.1 Einleitung

Data- und Textminingverfahren werden im Journalismus bisher noch sehr wenig angewendet. Neben den wenigen, bereits beschriebenen Projekten wird es vor allem zur automatisierten Artikelzusammenstellung auf Portalen wie Blendle, Upday oder Pocket verwendet. Mit Textminingverfahren werden dazu Inhalte indexiert und kategorisiert.

Die Medien kämpfen derzeit um ihr Ansehen. Maßnahmen sind gefordert, die mehr Professionalität fördern, Transparenz zeigen und sich klar von Kritikern mit Halbwissen abgrenzen. Journalisten können zukünftig ihre Geschichten mit starken Fakten aus Big Data bereichern und einen Mehrwert für die Konsumenten schaffen.

Im Journalismus kommt es auf die genaue Recherche, Schnelligkeit sowie unterschiedliche und verlässliche Quellen an. Der Auftrag von Journalisten ist heute wie früher unverändert: Aufspüren von Geschichten, Überprüfung, eine verständliche Aufbereitung und die Veröffentlichung von Informationen. Massendaten können ihnen dabei helfen, diese Aufgaben zukünftig zu optimieren, indem sie Geschichten offenlegen und Vergleiche ermöglichen.

Wie bereits im ersten Kapitel beschrieben, ist die wirtschaftliche Lage im Journalismus seit Jahren angespannt. Sinnvoll sind daher alle Überlegungen über automatisierte Abläufe und die schnelle, einfache Auswertung von Massendaten. Warum sollten die Medien nicht, wie alle anderen Branchen auch, eine Lösung finden, um von Big Data, dem Rohstoff des 21. Jahrhunderts, zu profitieren?

Dieser Rohstoff scheint zumindest zu wachsen, und zwar exponentiell. Nach einer Marktanalyse von International Data Corporation [vgl. IDC: 2016] verdoppeln sich die Daten alle zwei Jahre und für 2020 wird ein Datenvolumen von 44 Zettabytes, umgerechnet 44 Trillionen Gigabyte, prognostiziert.

Das große Berufsgeheimnis liegt im Finden von Geschichten. Vergleichbar mit der Wissenschaft und Forschung geht der Journalist detektivisch einer Idee oder einem Hinweis nach und versucht, diesen von mehreren Perspektiven zu beleuchten. Die Vorgehensweise ist unabhängig vom Ressort, Medium und ob der erste Anstoß aus einer Pressekonferenz, dem Newsticker oder der Eigeninitiative resultiert.

Das erste Gespür, eine These, eine Nachrichtenausgangslage oder die instinktive Vermutung zu einer Geschichte muss folglich zunächst gegeben sein.

Aber es stellt sich das Problem, dass es bisher für Journalisten keine einfache Möglichkeit gibt, selbst mit Daten zu experimentieren. Finanz- und Sportredaktionen sind gute Beispiele für die aktuellen Schwierigkeiten im Umgang mit großen Datenmengen, da ein Großteil der Inhalte schon immer aus Statistiken, Zahlen und Auflistungen, Prognosen und Entwicklungen bestanden. Im sportlichen Umfeld bietet beispielsweise die externe Firma deltatre an, Daten bzw. fertige Statistiken zu erstellen. Ein Beispiel für einen vergleichbaren Auftrag wäre, eine Fußball-Statistik über Tore einer bestimmten Spielminute in einem festgelegten Zeitraum anzufordern, die mit Wetterdaten und Trefferfuß verglichen werden. Das beste Ergebnis wäre, wenn sich der Verdacht des Redakteurs bestätigt, dass es überdurchschnittlich viele Zusammenhänge zwischen dem Sonnenstand und der linksfüßigen Trefferquote gäbe. Aber was passiert, wenn diese Statistik nichts aussagt? Die Beauftragung externer Anbieter schränkt das Experimentieren, die explorative Datenanalyse ein, und der Redakteur muss sich festlegen. Das ist die alte Arbeitswelt. Eventuell hätte ein anderer Fokus wie beispielsweise eine andere Spielminute zu einem aussagekräftigeren Ergebnis geführt. Der Anspruch sollte fortan sein, dass Redaktionen selbst mit Daten Geschichten erzählen oder in Daten Geschichten finden.

Dieses Kapitel behandelt die praktische Anwendung der in Kapitel 3 beschriebenen Möglichkeiten zur Datenanalyse. An einem generalisierbaren Beispiel soll dazu ein zukünftig denkbarer Einsatz in der journalistischen Arbeit aufgezeigt und die Rolle von Daten- bzw. Textminingverfahren in der journalistischen Arbeit beschrieben werden. Es wurden die Vorteile der technologischen Entwicklungen herausgearbeitet und die Voraussetzungen für die Ausnutzung des Big Data- Potenzials definiert. Die einzelnen Arbeitsabläufe und verbundene Schwierigkeiten wurden dazu analysiert. Ein besonderer Schwerpunkt galt der nötigen standardisierten Bereitstellung von Daten nach dem Open-Data-Prinzip.

4.2 Das Beispielprojekt – Wie belebt sind europäische Innenstädte im Vergleich

Die Ausgangsidee ist eine datengestützte Geschichte zu Veränderung der Belebtheit und Nutzung mehrerer Stadtzentren im Vergleich.

Diese Geschichte lässt sich kleinteilig innerhalb nur einer Stadt genauso wie im internationalen Vergleich zwischen Hamburg, Berlin und New York betrachten. Innerhalb der Städte gibt es weitere unterschiedliche Zentren. So wäre der Vergleich zwischen dem Berliner Kudamm, Prenzlauer Berg und Kreuzberg ebenso wie Hamburgs Reeperbahn, Eimsbüttel und der Innenstadt dafür interessant. Und

selbst innerhalb dieser Zentren gibt es wieder Unterteilungen wie die Hamburger Mönckebergstraße und Hamburger Jungfernstieg.

Ziel soll es sein, aus den Daten später Narrationen ableiten und stützen zu können. Informationen, wann sich wer, wo, wie lange, womit und warum aufhält, könnten für viele unterschiedliche Personengruppen wie Einwohner, Touristen, Einzelhändler oder Werbetreibende relevant sein. Es ergeben sich abhängig vom publizierenden Medium und der Zielgruppe verschiedene Erzählungen zum jeweiligen Stadterlebnis, das zu einer bestimmten Zeit erlebbar wird. Einige Schwimmbäder wie das Hallenbad Holstentherme in Kaltenkirchen bieten zum Beispiel bereits den Service an, die aktuellen Besucherzahlen zu veröffentlichen. So können Familien oder Sportschwimmer ihren Besuch auf die Belegung abstimmen. Eine ähnliche Dienstleistung bietet Google mit der *„Popular Times"*-Funktion. Das Unternehmen benutzt Standortdaten von Smartphones, um über die Belebtheit von Orten Auskunft zu geben. Zukünftig soll mit der *„plan your visit"*-Funktion Prognosen über einen Abgleich von Vergangenheits- und Livedaten gegeben werden. Die Standortdaten sagen aus, wie viele Personen sich an bestimmten Orten aufhalten. Zusammen mit mitgelieferten Zeitdaten kann Google beispielsweise genaue Verkehrsmeldungen geben.

Spannend werden diese Daten, wenn zusätzlich die Möglichkeit der Partizipation besteht. Durch die Einspielung von stetig aktualisierten, sogenannten dynamischen Daten und Interaktionsmöglichkeiten können die Narrationen individuell an die Bedürfnisse angepasst werden. Für das Beispielprojekt ist zu prüfen, inwiefern Google eine gute Arbeitsgrundlage bietet.

4.2.1 Team

Der Rechercheaufwand und die Umsetzung für ein solches Projekt ist groß. Es benötigt gleichermaßen Journalisten, Programmierer und Grafiker, die im ständigen Austausch miteinander stehen. Eine gute Kommunikation untereinander gilt als Grundvoraussetzung für den Erfolg des Projektes.

Verlage brauchen zukünftig interdisziplinäre Teams aus Reportern, Entwicklern, Designern, Editoren und Community Managern anstatt, wie oft vorherrschend, in einzelne Departements zu trennen. Es wird längst Zeit, das hierarchische Denken in den Redaktionen aufzugeben. Viele gute Datenjournalisten sind beispielsweise nicht unter den meist prämierten Journalisten aufgelistet [vgl. Howard: 2014, S. 5]. Zum einen genießen sie noch nicht den verbreiteten Respekt, um von den sogenannten Edelfedern der Branche ernst genommen zu werden. Zum anderen existiert noch keine Vergleichbarkeit mit der Arbeit traditioneller Medien.

Für eine erfolgreiche Teamarbeit ist es wichtig, dass es im Fachwissen Überschneidungen gibt, die notfalls angeeignet werden müssen. Auch eine stetige Zusammenarbeit mit Schulen, Universitäten, Archiven und Pädagogen ist empfehlenswert, um den aktuellen Stand der Forschung und zukünftige Trends kennenzulernen [vgl. Howard: 2014, S. 76 f.]. Der Austausch zwischen Wirtschaft und Bildungsträgern kann inspirieren und alle Seiten können davon profitieren.

Die Entwicklung in den Newsrooms zeigt, dass oft der anfänglichen Euphorie für Innovationen, die durch Weiterbildung ausgelöst wurde, schnell wieder der Alltag und die Routine folgt [vgl. Lewis, 2015, S.326-327]. Trotz einiger erfolgreicher innovativer Experimente haben sich neue Herangehensweisen, Perspektivwechsel und Veränderungen bisher wenig nachhaltig etabliert. Die Angst vor Fehlern, betrieblichen Verlusten und Risiko überwiegt bei den erfolgsverwöhnten Entscheidern der Medienhäuser, die über lange Zeit durch Anzeigenerlöse Höchstumsätze erzielten. Ohne ein Umdenken und die Bereitschaft sich zu verändern, bleibt das neue digitale Zeitalter unverständlich.

Journalisten sollten sich nicht länger von Technik und Zahlen einschüchtern lassen. Sie sollten experimentieren, spielen und überlegen, welche Geschichten in ihnen stecken könnte. [vgl. Gray et al.: 2012, S. 125]. Der allgemeine Umgang mit Zahlen, Mathematik und Statistik sind erlernbar. Informatiker dagegen sollten sich mit Erzählstrukturen beschäftigen. Um die richtigen Daten einbinden zu können, sind Kenntnisse für eine verständliche Aufbereitung von Daten notwendig. Der Endverbraucher denkt nicht in Einsen und Nullen. Für die Umsetzung derartiger Projekte sind längerfristig auch Schulungen und ständige Weiterbildungen ratsam. Da es oft keine routinierten Arbeitsabläufe und Erfahrungen gibt, sollte in solchen Projekten folglich zunächst abgeklärt werden, wer dafür zu begeistern und wer sich dafür weiterbilden und schulen ließe [vgl. Howard: 2014, S. 18].

Die Aufgabenverteilung kann wie folgt grob beschrieben werden: Es ist zunächst eine Liste mit allen Fragen zu erstellen, die beantwortet werden müssen. Hier ist es von Vorteil, wenn sich das Team aus hausinternen Mitarbeitern und ohne räumliche Trennung zusammensetzt. Im Falle externer Dienstleister kann es bei sehr komplexen Projekten sehr leicht zu Missverständnissen führen, die schnell in Rechtsstreitereien enden. Meistens ist dafür die mangelnde Kommunikation zwischen den Kreativen und der IT verantwortlich. Generell ist ein guter Projektmanager ratsam, der sich in beiden Fachrichtungen auskennt, den Überblick behält und rechtzeitig vermittelt.

Abgesehen davon, dass alle Aufgabenbereiche ineinander übergehen, sollte der Journalist sich federführend um die Datenrecherche und Beschaffung kümmern, die nicht automatisch durch beispielsweise offene APIs generiert werden kann. Die Datenzusammenführung in einer Datenbank und Bereinigung ist Aufgabe der IT-Abteilung. Das Analyseverfahren sollte zusammen ausgewählt und angewendet werden. Bei der repräsentativen Visualisierung und Interaktionsmöglichkeiten mit den Daten für die Endnutzerpublikation hilft der Grafiker.

4.2.2 Arbeitswerkzeuge

Um eine gute interdisziplinäre Arbeit zu gewährleisten, haben sich in den letzten Jahren besonders im technischen Bereich agile Arbeitsmethoden durchgesetzt. Durch die Projektaufteilung in kurze Arbeitsetappen und stetigem Feedback werden Arbeitsabläufe effektiver, indem schnell reagiert und optimiert werden kann. Unter den agilen Softwareanwendungen haben sich vermehrt Kanban und Scrum durchgesetzt. Auch die unterstützende Software MS Project und Jira werden gerne im Bereich

Projektplanung eingesetzt. Für die Umsetzung des Projektes sollte generell zunächst nach günstigen Open-Source-Anwendungen für alle Arbeitsschritte recherchiert werden. Im Laufe der Recherche können sich Kooperationen mit anderen Dienstleistern anbieten. Außerdem muss eine gute technische Infrastruktur bestehen. Ein gutes Content-Management beispielsweise führt die unterschiedlichen Inhalte später zusammen, klärt Zugriffsrechte und sollte darauf ausgelegt sein, dass gemeinschaftlich an dem Projekt gearbeitet werden kann.

Da der zeitliche Rahmen und der Aufwand am Anfang noch nicht exakt eingeschätzt werden können, sollte jedes Teammitglied für seinen Bereich dazu eine Beurteilung abgeben können.

4.2.3 Arbeitsablauf

Folgend wird der gesamtheitliche Arbeitsablauf dieses Projektes beschrieben, der sich an dem klassischen Aufbau des KDD-Prozesses orientiert. Es ist wichtig, dass alle Schritte gewissenhaft dokumentiert werden. So können die Ergebnisse jederzeit von allen Mitwirkenden nachvollzogen werden und Arbeitsabläufe gegebenenfalls rechtzeitig verändert, wiederholt oder gar abgebrochen werden.

4.2.3.1 Recherche

Interessant für das Projekt sind alle geolokalisierten Daten, die Auskunft über den Ort und Zeitraum geben. Um die Ausgangsfragestellung mit diesen Daten beschreiben zu können, muss ein Vektor aus verschiedenen Merkmalen (Featurevektor) zusammengestellt werden. Verschiedene Merkmale können dazu aus den folgenden Datenquellen in Betracht gezogen werden:

- Verkehrsaufkommen
- Verkehrsbehinderungen (Stau, Baustellen)
- Umsätze Einzelhandel
- Umsätze Gastronomie
- Parkhäuserbelegung
- Frequenz der öffentlichen Verkehrsmittel
- Hotelbelegung
- Veranstaltungen (Konzerte, Sportereignis, Märkte, Volksfeste)
- Social Media (Tweets, Fotos)
- Wetter
- Saison (Jahreszeiten, Schlussverkauf, Ferien, Feiertag)
- Taxis
- Mobilitätsverleihung (Fahrräder, Autos, Motorroller)
- Alter der Personen

Bei allen Datenquellen ist es wichtig, vorher bei der IT zu fragen, welche Dateiformate sie benötigt. Denn alle Daten eines Featurevektors werden in einer Datenbank zusammengeführt und gespeichert. Für den Aufbau einer Datenbank haben Entwickler ihre eigenen Präferenzen und jede Software andere Voraussetzungen wie bei fortlaufend dynamischer Dateneinbindung in Projekte

Die meisten Städte haben mittlerweile Transparenz- bzw. Open-Data-Portale. Wie im Hamburger Transparenzportal lassen sich viele Daten zum Beispiel über Baustellen, Fußgänger- und Radverkehrszählstellen oder Parkraumbelegung finden. Dabei kommt es darauf an, ob man an Livedaten interessiert ist oder einmal erhobene Datensätze ausreichen. Auch wenn nicht die entsprechenden Informationen dabei sind, dienen diese Portale als erster Anhaltspunkt. Wenn für die Zählung von öffentlich anonymisierten Fußgängerdaten beispielsweise bereits fest installierte Sensoren existieren, unterliegt der Zugriff darauf nur noch rechtlichen Problemen. Es ist mit der IT abzusprechen, ob es im Zweifelsfall sinnvoll ist, selbst Daten zu erheben und mit eigenen Sensoren zu arbeiten.

Als weitere Datenquellen erheben die Handelskammern regelmäßig Umsatzzahlen des Einzelhandels, und Wetterdienste liefern entsprechende Wetterdaten.

Bei der Recherche von Daten privater Firmen wie car2go, DriveNow, Hansa Taxi etc. ist man dagegen abhängig von der Zusammenarbeitsbereitschaft. Die Chancen erhöhen sich, wenn ihnen ein Mehrwert beispielsweise in Form einer Kooperation angeboten wird. Wenn eine journalistische Geschichte primär auf Daten privater Firmen aufbaut, dann kommt man an einer Zusammenarbeit nicht vorbei. Es handelt sich dann um Corporate Publishing und nicht mehr um eine unabhängig journalistische Arbeit. Es ist daher vom Journalisten zu prüfen, ob die Daten relevant sind.

Soziale Netzwerke können Meinungsbilder bei beispielsweise politischen Entscheidungen oder Diskussionen repräsentieren. Aber auch für das Projekt wären die Daten optimal. Sie sind schnell, geben Stimmungen wieder und gelten noch weitgehend als authentisch, wenn man die steigende Anzahl an (Social-)Bots hier einmal außer Acht lässt. Im Fall von Social Media bietet beispielsweise Twitter eine offene API an. Das Problem dabei ist, dass nur wenige Twitter-Nutzer geolokalisiert sind. Untersuchungen dazu ergaben, dass nur etwa 0,7% von 19,6 Millionen Tweets Geo-Koordinaten enthielten [vgl. Graham et al.: 2014]. Eine andere Studie über vier vergleichbare Großstädte ergab, dass im Zeitraum von vier Wochen nur zwischen 2% und 5% an Tweets anhand von Inhalten über die Städte geolokalisiert werden konnte [vgl. Severo et al.: 2015]. Es ist außerdem zu berücksichtigen, dass Twitter beispielsweise in Deutschland im Gegensatz zum Ausland überwiegend von Journalisten und im politischen Kontext genutzt wird. Eventuell lohnen sich aber die Suche nach Hashtags der Stadt z.B. auf Instagram und über Google-Analytics die Recherche nach entsprechenden Suchbegriffen zur Stadt und Events. Viele Social-Media-Dienste sind nicht geolokalisiert, und eine Anfrage zu anonymisierten Daten schwierig. Die verbreiteten und führenden Social-Media-Dienste sind aus dem Ausland und nicht daran interessiert. Die IT-Abteilung muss daher prüfen, ob eventuell weitere offene APIs

alternativer sozialer Netzwerke bestehen oder die Metadaten der Fotos und Tweets weiterhelfen. Bei der Beschaffung von digital vorliegenden Daten aus den Social-Media-Kanälen muss die IT eine Aufwandseinschätzung sowie eine Beurteilung über die Relevanz des Ergebnisses abgeben. Weitere allgemeine Möglichkeiten zur Datenrecherche siehe auch Kapitel 3.5: Datenrecherche.

Bei der journalistischen und IT-gestützten Untersuchung zur unterschiedlichen Belebtheit von Räumen wird größtenteils auf öffentliche Daten zurückgegriffen. Dieses Projekt ist derzeit abhängig von einer konsequenten Datenerhebung und Offenlegung seitens der Städte, denen noch nicht alle Folge leisten. Eine selbständige Erhebung von Daten durch Sensoren unterliegt dagegen strengen Datenschutzgesetzen. Im öffentlichen Raum ist eine Ausnahme mit umständlichen Behördengängen und zeitaufwendigen Prüfungen verbunden. Das ist kein einfacher Weg.

Zukünftig wird sich besonders im Umgang mit Lokationsdaten einiges ändern und viele Online-Dienstleister werden die Informationen über Standortdaten in ihren Service einbinden. Besonders im öffentlichen Sektor, ob für Auskünfte über genaue Abfahrtszeiten oder Parkplatzbelegung, werden die Daten bereits erfolgreich verwendet. Da Geodaten viele Informationen über den Kontext einer Person preisgeben, sind private Dienstleister, Entwickler, Medien, Marken oder Werber gleichermaßen interessiert. Große Halter von Lokationsdaten sind besonders die Endgerätehersteller und Telefonanbieter, und zu den größten und besten Datenquellen ohne Zugriff und Kontrolle zählen alle Privatpersonen [vgl. Gray et al.: 2012, 147f.].

Bereits anonyme Daten sind ausreichend verwertbar. Zukünftig wäre denkbar, dass Firmen besondere Angebote machen, wenn sich ein Kunde bzw. ein Smartphone in der Gegend aufhält. Oder das Medienangebot wird dem Aufenthaltsort angepasst. Einige Firmen, wie die belgische Firma Sentiance, haben sich bereits darauf spezialisiert, Geodaten aus dem Smartphone und Smarthome auszulesen. Mit den Raum-Zeit-Daten kann genau prognostiziert werden, wo Personen arbeiten und wohnen, wie sie zur Arbeit fahren und ob sie vorher noch in die Kita fahren. Diese Informationen reichen wie im Vorratsdaten-Fall von Malte Spitz aus, um eine Person genau identifizieren zu können und um zu wissen, in welchem Kontext sie steht.

Die größte Herausforderung der Zukunft ist daher eine standardisierte zugängliche Datenspeicherung und die Vereinheitlichung von Formaten. Offene Daten sind nicht nur im journalistischen Interesse. In diesem Zusammenhang werden Diskussionen und klare Richtlinien zur Datensicherheit und Datenschutz immer wichtiger.

> ***"This capability to publish data doesn't change the underlying ethics or responsibility that journalists uphold: Not all data can or should be published in such work, particularly personally identifiable information or details that would expose whistleblowers or put the lives of sources at risk."***
>
> *Alexander Benjamin Howard [2014, S. 18]*

Vergangene Auflistungen über betrunkene Fahrer, Kriminelle oder Personen mit Schusswaffen gleichen beispielsweise einem modernen Pranger. Aber die Förderung zur aktiven Teilnahme durch private Datengenerierung wäre dagegen beispielsweise im Umweltschutz denkbar. Es muss stets ein verantwortliches Abwegen zwischen öffentlichem und privatunternehmerischem Interesse stattfinden.

4.2.3.2 Datenspeicherung, Bereinigung und Umwandlung

Die Daten müssen von der IT in einer Datenbank zusammengeführt werden, die dafür eigens angelegt wird. Dazu müssen die Datensätze zunächst in die richtigen Formate umgewandelt werden. Es muss überprüft werden, ob sie aktuell, vollständig, fehler- und widerspruchsfrei sind. Vielleicht reicht auch nur eine Teilmenge der Daten zur Betrachtung aus oder eine Transformation in metrische Daten ist nötig. Der Umgang mit Massendaten erfordert große und schnelle Computerrechenleistungen und viel Speicherplatz. Daten werden oft nicht für Analysezwecke gespeichert, so dass die Datensätze (auch „dirty data“ genannt) oft zunächst um alle überflüssigen Daten reduziert werden. Zur Erleichterung und vollen Ausschöpfung des Big-Data-Potenzials wäre diese Arbeit weniger mühsam, wenn es allgemein verbindlich vorgeschriebene einheitliche Standards zur Datenspeicherung geben würde. Es handelt sich um den Schritt der Bereinigung und Analysevorbereitung im KDD-Prozess. Es kann sein, dass die einzelnen Schritte dieser Preprocessing-Phase für eine optimale Datenvorbereitung mehrmals wiederholt werden müssen. Ein Protokoll dokumentiert die einzelnen Vorgänge und vorgenommenen Veränderungen.

4.2.3.3 Analyse

> ***„[..] In order to understand a data set, it is helpful to start with understanding the people who created the data set—think about what they were trying to do, or what they were trying to discover. Once you think about those people, and their goals, you’re already beginning to tell a story.“***
>
> *Meredith Broussard, assistant professor of journalism at Temple University [Howard:2014, S. 5]*

Da es sich um Geodaten handelt, würde sich ein sogenanntes Geografisches Informationssystem (GIS-System) für die Weiterverarbeitung der Datensätze anbieten. Es ermöglicht eine räumlich-zeitliche Sortierung mit anschließender Analyse- und Visualisierungsmöglichkeit.

Im Mining-Verfahren wird in diesem System der vorher definierte Featurevektor in verschiedene Segmente eingeteilt. Sie bilden einzelne Cluster. Im Projekt Städtevergleich bestehen die Segmente aus Orts- und Zeitangaben. Um aussagekräftig zu sein, wären es für vorher ausgewählte GPS-Koordinaten (mit einem Umkreis von 2 km) beispielsweise 21 Cluster (7 Wochentage x 3 Tageszeiten

wie Vormittag, Nachmittag, Abend). Als Ergebnis erhält man Cluster, die Informationen zu allen abgesetzten Tweets, Verkehrssaufkommen, Personenanzahl etc. in einem bestimmten Raum zu einer vordefinierten Zeitanspanne enthalten.

Eine Analysemethode könnte anschließend die Suche nach Häufigkeiten (Distanzberechnung), der Vergleich zu anderen Raum-Zeit-Segmenten (Klassifizierung durch Gewichtung) oder ein Stimmungsbild (Text Mining) sein. Durch eine Klassifizierung könnten weitere Städte in Klassen eingeteilt werden. Mit der Assoziationsanalyse könnten Zusammenhänge, Regelmäßigkeiten und Verhaltensweise untersucht sowie Prognosen abgegeben werden. In unserem Projekt wären es Cluster aus Wochentagen und Tageszeiten eines Raumes und eine Klassenbildung aus ähnlichen Clustern für den Raumvergleich.

Eine Auswertung über die Häufigkeiten von abgesetzten Tweets oder Fotos in den ausgewählten Räumen gibt zunächst erste Indizien über die Belebtheit. Bei vorausgesetzt zugänglichen Social-Media-Daten könnten Textminingverfahren darüber hinaus schnell, aktuell und effizient repräsentative Stimmungsbilder erstellen. Durch eine genauere Analyse zum Beispiel von Emojis sowie der Suche nach Schlagwort- und Hashtag-Kollokationen mit bewertenden Verben und Adjektiven könnten die Texte nach emotionalen Inhalten untersucht werden. Dieses Text Mining wird als Sentimentanalyse bezeichnet. Und das entsprechende Forschungsfeld wird als emotionales Information-Retrival bezeichnet. Für das Projekt wäre es beispielsweise denkbar, die Texte nach dem Stadtnamen und zugehörigen Synonymen wie Innenstadt und City in Verbindung mit Beschreibungen wie *leer, ausgestorben, voll, wenig, viel, los, kein Platz, überfüllt, drängeln*, *ganz*, *alle, entspannt, Andrang, Tipp* zu untersuchen.

Es ist Aufgabe der IT, eine Abschätzung über den Aufwand abzugeben. Es verlangt im Arbeitsprozess eine Absprache mit der Redaktion, welche Schwerpunkte gelegt werden.

4.2.3.4 Explorative Datenanalyse

Für die explorative Datenanalyse bietet sich die Visualisierungstechnik an, die alle Daten auf einer Karte zusammenführt. Es sollten unterschiedliche Interaktionsmöglichkeiten für eine genauere Betrachtung bestehen. Das beinhaltet die Anwendung von beispielsweise Filtern, Teilmengen- und Vergleichsbetrachtungen, um Zusammenhänge und Abhängigkeiten zwischen den einzelnen Clustern und Klassen erkennen zu können. Im Beispiel-Projekt wäre es nützlich, nach Zeiträumen zu vergleichen und Ortsunterschiede zu definieren. Die Rolle besonderer Faktoren wie Wetter und Events sollten gesondert analysiert werden können. Journalisten, Programmierer und Grafiker sollten gleichermaßen mit den Visualisierungen experimentieren, da es sich um einen sehr kreativen Arbeitsschritt handelt und mit entsprechender Software keine großen technischen Hürden birgt. Jeder sieht anders auf Zahlen und Grafiken.

4.2.3.5 Präsentation und Publikation

Aus der Datenanalyse und Visualisierung ergeben sich anschließend unterschiedliche Ansätze für Narrationen. Der Journalist sollte wie üblich durch Expertenbefragungen, Interviews und Meinungsumfragen Hintergrundinformationen zur den bisher datengenerierten Geschichten recherchieren. In Absprache mit dem Redakteur kann der Grafiker durch Einfärbungen, Icons, unterschiedlichen Größenverhältnissen und Beschriftungen die Kartografie optisch endnutzerfreundlich gestalten. Die endgültige Programmierung mit Navigations- und Interaktionsmöglichkeiten, eventueller dynamischer Dateneinbindung und Publikation übernimmt die IT. Die folgenden Skizzen zeigen Beispiel-Visualisierungen für das Projekt.

Abbildung 24-25: Beispielgrafiken zur Belebtheit von Orten *[eigenhändig erstellt].*
Die mobiloptimierte Grafik zeigt die Belebtheit eines ausgewählten Ortes, in diesem Fall die Innenstadt von Hamburg. Für weitere Informationen lässt sich die Zeit und im Menü die einzelnen Featurevektoren und Untermenüs zu Social-Media-Beiträgen, Gastronomie, Verkehr, demografische Informationen und Veranstaltungen individuell auswählen. Die Datenquellen sollten aufgrund der Nachvollziehbarkeit beigefügt werden.

Für die Veröffentlichung des Projektes sollte Folgendes beachtet werden:

- Ausspielmöglichkeit auf unterschiedlichen Endgeräten (responsive)
- Feedbackmechanismus
- Personalisierbarkeit
- Diskussionsmöglichkeit
- Aktualität der Daten

Zur Veröffentlichung der Narration gehört auch die verantwortungsvolle Offenlegung der Rohdaten und bearbeiteten Datensätze unter strenger Einhaltung des Datenschutzes. Es sollte sich an dem Open-Data- und Open-Source-Gedanken orientiert werden. Es schafft Transparenz und der Leser kann die Grundlage der Geschichte nachvollziehen. Außerdem dienen die Daten beispielsweise neuen, weiterführenden Erzählungen. Der Datenjournalismus lebt von der Partizipation der Konsumenten und ist daher vor allem in digitaler Publikation sinnvoll.

> ***„In the same way that open source developers show their work when they push updated software to GitHub, data journalists are publishing updates to data sets that accompany narrative stories or news applications. “***
>
> *Alexander Benjamin Howard [Howard:2014, S. 18]*

4.3 Fazit

Im Beispielprojekt konnte der Einsatz von Softwareanwendungen im Journalismus aufgezeigt werden, die zuvor in dieser Arbeit beschrieben wurden. Der Bedarf an Data- und Textminingverfahren ist aufgrund der steigenden Informationsflut und Big Data größer denn je. Die Recherchemöglichkeiten haben sich gleichermaßen mit der Komplexität durch die digitale Vernetzung erweitert. Ohne automatisierte Abläufe und Miningverfahren sind die Massendaten nicht mehr zu bewältigen. Und in ihnen lassen sich viele Narrationen für Journalisten finden. Das Projekt ist beispielhaft und zeigt die Übertragbarkeit auf ähnlich denkbare Anwendungsfälle im Journalismus. Es zeigt die steigenden Anforderungen an Journalisten und die IT. Ohne eine funktionierende Zusammenarbeit interdisziplinärer Teams sind derart große und komplexe Projekte nicht möglich. Computer sind allgegenwärtig und es wird deutlich, dass zukünftig ein technisches Grundverständnis in jeder Berufsgruppe notwendig sein wird, um auf Augenhöhe in Teams kommunizieren zu können. Die Arbeit zeigt aber auch, dass auf menschliche Einschätzungen und die Gabe des Erzählens weiterhin nicht verzichtet werden kann. Ein wichtiger Bestandteil jeder Geschichte sind emotionale Komponenten, die bisher nicht von Maschinen glaubwürdig erbracht werden können.

5 Zusammenfassung und Ausblick

> ***"Dealing with change is hard for any incumbent organization. [..] Most of them don't have training embedded in the culture – training and learning."***
>
> *Melanie Sill, KPSS's Vice President of Content [Stencel et al.: 2014, S. 18]*

Anhand der Ausarbeitungen zum Datenjournalismus in Kapitel 2, den technischen Softwareanwendungen in Kapitel 3 sowie dem datengenerierten Beispielprojekt in Kapitel 4 kann keine ganz abschließende Einschätzung zu den Schwierigkeiten, zukünftigen Entwicklungen und denkbaren Lösungsansätzen gegeben werden.

Der Journalismus soll unabhängig informieren und das Zeitgeschehen wiedergeben. Die Aktualität spielt in der journalistischen Arbeit immer eine große Rolle. Daher ist es umso erstaunlicher, dass die Medien die Entwicklungen sehr spät gesehen und auf die Digitalisierung reagiert haben. Die Konsequenzen zeigen sich in zahlreichen Entlassungen und Redaktionsschließungen, obwohl das Arbeitspensum deutlich mehr geworden ist. Die Berichterstattung über technische Innovationen und digitale Trendthemen war über lange Zeit sehr verhalten, obwohl sie allgegenwärtig waren. Aber schwerwiegender ist die Tatsache, dass der Transfergedanke von der Digitalisierung auf die eigene Arbeit lange nicht stattgefunden hat. Lange wurde das Internet als Medienkanal gegenüber gedruckten Medien nicht ernst genommen und nicht als neue Chancen gesehen.

Früher ging es den Medien wirtschaftlich sehr gut. Sie waren gefestigt in ihrer Stellung als vierte Gewalt. Aber aus den erfolgsverwöhnten Berichterstattern werden mittlerweile Bittsteller für Fördertöpfe investigativer Recherche. Es entstehen in diesem Zusammenhang immer wieder kontroverse Diskussionen über die GEZ-Abgaben bzw. den geförderten Auftrag und die Bevorzugung der öffentlich-rechtlichen Medien. Die Medienhäuser reagieren mit Forderungen nach einem Leistungsschutzgesetz auf die angespannte Lage. Mit dem Argument des Schutzes geistigen Eigentums versuchen sie, von ihrer Machtlosigkeit gegenüber neuer Mediennutzung und sinkenden Einnahmequellen abzulenken.

Aber wie viele andere Branchen müssen auch die Medien eine neue Risikobereitschaft und den Willen zur Veränderung zeigen, Experimentieren lernen und auch Scheitern zulassen. Einen Masterplan für die Digitalisierung gibt es bisher nicht und es ist heute noch nicht abzusehen, welche zukünftigen Geschäftsmodelle sich bewähren werden. Der Datenjournalismus im Speziellen ist keine Lösung für diese Krise.

Dazu stellten die Professoren Leif Kramp und Stefan Weichert [2015] in ihrer Studie „Die Zeitungsmacher“ fest: „Zu den zweifelsohne größten Herausforderungen gehören momentan die Etablierung von tragfähigen Bezahlmodellen im Internet, die redaktionelle Verschränkung von Print- und Digital-Angebot und ein Konzept zur profitablen Einbindung der User. Das ist das Pflichtprogramm in vielen Redaktionen, doch bis zur Kür – beispielsweise der Ausreizung digitaler Erzählformen – gelangt man gar nicht erst.“

Aus ihrer Studie leiten Kramp und Weichert [vgl. 2015: S. 52-55] weitere Empfehlungen an die Redaktionen ab. Neben dem Mut zu scheitern, raten sie zu einem Change-Management unter demokratischer Beteiligung aller Mitarbeiter. Veränderungen sollten nicht allein von der Führungsspitze aus dirigiert werden. Um Trends und neue Mediennutzungsgewohnheiten rechtzeitig erkennen zu können, sollte in regelmäßige Weiterbildungen investiert werden und neue Personaleinstellungen gewagt werden. Solch ein Experimentieren kann auch zu Berufsbildern wie „Maker“ oder „Programmer-Journalist“ führen. Nach der Auffassung von Kramp und Weichert spielen die Nutzer und die Kommunikation zu ihnen eine zunehmend wichtige Rolle. Sie sollten in den Kreativprozess einbezogen werden.

Eine wichtige Frage der Zukunft wird es dabei deshalb auch sein, welche Rolle neue konkurrierende Berufs- und Freizeitpublizisten wie Blogger, Hobbyfotografen oder Leserreporter dabei spielen werden. Sie unterliegen keiner Verpflichtung zur Unabhängigkeit, Einhaltung ethischer Grundsätze und Richtlinien zur Beherrschung eines professionellen Umgangs mit Quellen. Bisher ist es Aufgabe der Journalisten, den Standard an Genauigkeit zu gewährleisten.

Es ist aus heutiger Sicht noch nicht abzuschätzen, ob unabhängiger Journalismus, Aushängeschild für freiheitlich demokratische Strukturen, weiterhin in seiner bisherigen Form bestehen kann und welchen Stellenwert kritischer Journalismus in der Gesellschaft haben wird. Die aktuellen Ereignisse im Zusammenhang des Politischen Wechsels in Amerika durch Präsident Donald Trump, die vermehrten Inhaftierungen von Journalisten in der Türkei, die Einschränkungen der Pressefreiheit in Russland und Polen zeigen deutlich, welchen Konflikten die Medien ausgesetzt sind. Eine Wertschätzung zeigt sich unter anderem auch in der Bereitschaft, für Inhalte zu bezahlen.

Neben der traditionellen Auffassung des Berufsbildes sollten sich Journalisten in Zukunft auch als Vermittler und Interpreter von Quellen verstehen [vgl. Neumüller und Kahn: 2015]. Diese Aufgabe gewinnt in Zeiten von Big Data eine wachsende und unverzichtbare Bedeutung, denn es fällt zunehmend schwer, individuell relevante Informationen herauszufiltern.

Es wird sich zeigen, ob Datenanalysten oder Informatiker Journalisten teilweise zukünftig ersetzen können und wo Einsparungen vorgenommen werden. Es ist ein kreativer Prozess, eine Kunst und Handwerk, eine Geschichte zu finden, sie verständlich, interessant und spannend zu erzählen. Die menschliche Komponente sollte dabei nicht unterschätzt werden, denn sie macht Geschichten erst lebendig.

Es ist eine spannende Zeit für Journalisten. Sie bekommen derzeit eine technologische Spielwiese geboten, in die sie investieren sollten. Computerunterstütze Emotionserkennung, Virtual- und Augmented Reality, Drohnen, Audio- und Sprachsteuerungssysteme – all diesen Entwicklungen bedarf es an kreativen Ideen für innovative Narrationen. Journalisten sollten sich hier selbst besser vertrauen, denn gerade sie sind es, die genau dieses Handwerk eigentlich beherrschen.

> ***„To become a good data journalist, it helps to begin by becoming a good journalist. Hone your storytelling skills, experiment with different ways to tell a story, and understand that data is created by people. [..]“***
>
> *Meredith Broussard, assistant professor of journalism at Temple University [Howard:2014, S. 5]*

6 Literaturverzeichnis

Amerland, Andrea: Wenn Journalisten in die PR wechseln. Wiesbaden: Springer Professional, Springer Fachmedien Wiesbaden GmbH, 2013. Online unter: https://www.springerprofessional.de/public-relations/wenn-journalisten-in-die-pr-wechseln/6603052
(abgerufen am 16. Juli 2016)

Barthel, Michael: Newspaper: Fact Sheets. State oft the News Media 2016. Washington: The Pew Research Center's Project for Exellence Journlism, 2016. Online unter: http://www.journalism.org/2016/06/15/newspapers-fact-sheet/
(abgerufen am13. Juli 2016)

Baum, Gerhart: Abrüsten! In: Zeit Online, 2016. Online unter: http://www.zeit.de/2016/04/datenschutz-internet-sicherheit
(abgerufen am 2. August 2016)

BDZV: Leistungsschutzrecht für Verlage. Berlin: Deutscher Zeitungsverleger e.V., 2016a. Online unter: http://www.bdzv.de/medienpolitik/leistungsschutzrecht-fakten/
(abgerufen am 16. Oktober 2016)

BDZV: Paid Content Angebote deutscher Zeitungen. Berlin: Bundesverband Deutscher Zeitungsverleger e.V., 2016b. Online unter: http://www.bdzv.de/maerkte-und-daten/digitales/paidcontent/
(abgerufen 14. Juli 2016)

Beeck, Torsten: Wer kümmert sich eigentlich um die tausenden Userkommentare, die Medien täglich bekommen? In: Zeit Magazin, Zeit Online, Hamburg 2016. Online unter: http://www.zeit.de/zeit-magazin/2016/31/kommentare-internet-medien-community-redakteur
(abgerufen am 4. August 2016)

Bösch, Marcus: Journalismus zum Spielen. Berlin: Stiftung Digitale Spielekultur GmbH, 2013. Online unter: http://stiftung-digitale-spielekultur.de/artikel/journalismus-zum-spielen
(abgerufen am 3. August 2016)

Boyle, James: The Public Domain: Enclosing the Commons of the Mind. London: Yale University Press, 2008. Online unter: http://thepublicdomain.org/thepublicdomain1.pdf
(abgerufen am 23. Juli 2016)

BPB: Medien - Die "vierte Gewalt"? Bonn: Bundeszentrale für politische Bildung, 2016. Online unter: http://www.bpb.de/politik/grundfragen/deutsche-verhaeltnisse-eine-sozialkunde/138737/medien (abgerufen am 13. Juli 2016)

Bubenhofer, Noah: GeoCollocations – Diskurse zu Orten: Visuelle Korpusanalyse. In: Sondernummer Mitteilungen des Deutschen Germanistenverbandes 1/2014: Korpora in der Linguistik – Perspektiven und Positionen zu Daten und Datenerhebung. Göttingen: V&R Unipress GmbH, 2014.

Bundesgerichtshof: Internet-Suchdienst für Presseartikel nicht rechtswidrig. Pressemitteilung Nr. 96/03 vom 17.7.2003. Karlsruhe: Bundesgerichtshof, 2003, S. 25.
Online unter: http://juris.bundesgerichtshof.de/cgi-bin/rechtsprechung/document.py?Gericht=bgh&Art=en&Datum=2003&Sort=3&Seite=5&client=3&anz=1706&pos=171&nr=27035
(abgerufen am 16. Oktober 2016)

Breners-Lee, Tim: 5-Sterne Modell. In: Open Goverment Data Weißbuch. Österreich: Universitätsverlag der Donau-Universität Krems, Kaltenböck, Martin; Thurner, Thomas (Hrsg.), 2011. Online unter: http://open.semantic-web.at/display/OGDW/6.3+Open-Data-5-Stern-Modell+von+Tim+Berners-Lee
(abgerufen am 2. August 2016)

Carter, Shan; Leonhardt, David; Marsh, Bill; Quealy, Kevin: Budget Puzzle: You Fix the Budget. In: The New York Times. New York: The New York Times Company, 2010. Online unter: http://www.nytimes.com/interactive/2010/11/13/weekinreview/deficits-graphic.html?_r=0
(abgerufen am 3. August 2016)

Caspar, Johannes: Tätigkeitsbericht Informationsfreiheit 2014 / 2015. Hamburg: Der Hamburgische Beauftragte für Datenschutz und Informationsfreiheit (Hrsg.), 2015. Online unter: https://www.datenschutz-hamburg.de/uploads/media/Taetigkeitsbericht_Informationsfreiheit_2014-2015.pdf
(abgerufen am 2. August 2016)

Chomsky, Noam: Syntactic Structures. Massachussets Institute of Technology, 1956. Paris: Mouton Publishers, The Hague. 1957.

Christy, Bryan: How Killing Elephants Finances Terror in Africa. Washington: National Geographics Magazine, 2015. Online unter: http://www.nationalgeographic.com/tracking-ivory/
(abgerufen am 8. August 2016)

Cisco: The Zettybyte Era Trends and Analysis. Cisco Visual Networking Index (VNI). 2. Juni 2016.
Online unter: https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/vni-hyperconnectivity-wp.html
(abgerufen am 2. Dezember 2016)

Cleve, Jürgen; Lämmel, Uwe: Data Mining. Berlin: De Gruyter Oldenbourg Wissenschaftsverlag, 2. Auflage, 2016.

Coddington, Mark: Clarifying Journalism's Quantitative Turn. A Typology for Evaluating Data Journalism, Computational Journalism, and Computer-Assisted Reporting. Austin, USA: School of Journalism, University of Texas, 2014.
Online unter: http://www.tandfonline.com/doi/pdf/10.1080/21670811.2014.976400
(abgerufen am 14. November 2016)

Cooper-Wright, Matt: Are You a Good Driver? How Designers Use Data to Get to the Truth.
In: Design X Data. London: Medium, 2015.
Online unter: https://medium.com/design-x-data/are-you-a-good-driver-how-designers-use-data-to-get-to-the-truth-3c534fcaf9d2#.235ngmv7r
(abgerufen am 2. August 2016)

Creative Commons: Icons. Mountain View: Creative Commons, 2016.
Online unter: https://creativecommons.org/about/downloads/
(abgerufen am 17. Juli 2016)

Daniel, Anna; Flew, Terry: The Guardian Reportage of the UK MP Expenses Scandal:
a Case Study of Computational Journalism. In: Communications Policy and Research Forum.
Sidney: Queensland University of Technology, 2010.
Online unter: http://eprints.qut.edu.au/38701/2/38701.pdf
(abgerufen am 10.02.2016)

Deterding, Sebastian: Into the Great Wide Open, Open Source, jenseits von Software. Bonn: Bundeszentrale für politische Bildung, 2007.
Online unter: http://www.bpb.de/gesellschaft/medien/opensource/63921/into-the-great-wide-open
(abgerufen am 18. Juli 2016)

Deutscher Wortschatz: 2016, Abteilung Automatische Sprachverarbeitung, Universität Leipzig.
Online unter: http://wortschatz.uni-leipzig.de/abfrage/
(abgerufen am 12. September 2016)

DGIQ: Informationsqualität. 15 Dimensionen, 4 Kategorien. Lünen: Deutsche Gesellschaft für Informations- und Datenqualität e.V. (DGIQ), 2007.

Diakopoulos, Nicholas: Agorithmic Accountability Reporting: In the Investigation of Black Boxes. In: The Tow Center for Digital Journalism. A Tow/Knight Report. New York: Columbia University, 2014. Überarbeitete Fassung: Algorithmic Accountability. Journalistic investigation of computational power structures. University Maryland, USA, 2014. In: Digital Journalism, Ausgabe 3. London: Routledge, Taylor & Francis Group, 2015.

Dietrich, Daniel: „Was sind offene Daten?“, Bundeszentrale für politische Bildung, 2011a.
Online unter: http://www.bpb.de/gesellschaft/medien/opendata/64055/was-sind-offene-daten
(abgerufen am 15.07.2016)

Dietrich, Daniel: Open Data - Offene Daten In Deutschland. 2011b. Online unter:
http://www.bpb.de/gesellschaft/medien/opendata/64061/ offene-daten-in-deutschland. 2011.
(abgerufen am 2. August 2016) .07.2016)

Dorling, Daniel: The Visualisation of Spatial Social Structure. The Distibution of Voting, Housing, Employment and Industrial Compositionsin 1887. England: 1991.
Online unter: http://www.dannydorling.org/books/visualisation/Graphics/Pages/Figures.html#156
Abgerufen am 15. September 2016

Dörr, Dieter; Schwartmann, Rolf: Medienrecht. Heidelberg: C.F. Müller, 4. Auflage, 2012.

Drepper, Daniel: Jahrestagung von Netzwerk Recherche im Livestream Journalismus an der Grenze. In: Investigativer Datenjournalismus: Zur exklusiven Story mit Daten, Dokumenten und Leaks. Hamburg: Spiegel Online, 2016, (Minute 14:08 f).
Online unter: http://www.spiegel.de/netzwelt/netzpolitik/pressefreiheit-und-populisten-journalismus-an-der-grenze-a-1101670.html
(abgerufen am 8. Juli 2016)

Dumbill, Edd: What Is Big Data? In: Dumbill, Edd (Hrsg.): Planning for Big Data – A CIO’s Handbook to the Changing Data Landscape. 1. Ausgabe. Sebastopol, California: O’Reilly Media, 2012.
Online unter: http://eecs.wsu.edu/~yinghui/mat/courses/fall%202015/resources/planning-for-big-data.pdf
(abgerufen am 16. Juli 2016)

Egawhary, Elena; O’Murchu, Cynthia: Data Journalism. London, City University, The centre for investigative journalism (CIJ), 2012.
Online unter: http://www.tcij.org/sites/default/files/u4/Data%20Journalism%20Book.pdf
(abgerufen am 14. Juni 2016)

Elmer, Christina; Wormer, Holger: Datenjournalismus – was ist das? In: Das Blog zur nr-Jahreskonferenz 2014 — Berlin: netzwerk recherche e.V., 2014.
Online unter: http://netzwerkrecherche.org/wordpress/blog14/datenjournalismus-was-ist-das/
(abgerufen am 10. Juli 2016)

Evert, Stefan: Corpora and Collocations. In: Lüdeling, Anke; Kytö, Merja (Hrsg.): Corpus Linguisitcs: an international handbook. Berlin: Mouton de Gryter, 2008.

Fayyad, Usama M; Piatetsky-Shapiro, Gregory; Smyth, Padhraic: The KDD-Process for Extracting Useful Knowledge from Volumes of Data. ACM 39, 1996a.
Online unter: https://dl.acm.org/citation.cfm?id=240464&CFID=656668054&CFTOKEN=10455264
(abgerufen am 12. Juli 2016)

Fayyad, Usama M.; Piatetsky-Shapiro, Gregory; Smyth, Padhraic: From Data Mining to Knowledge Discovery: An Overview. In: Fayyad, Usama M., Piatetsky-Shapiro, Gregory; Smyth, Padhraic; Uthurusamy, Ramasamy (Hrsg.): Advances in Knowledge Discovery and Datamining. London: Cambridge, Menlo Park, MIT Press, 1996b.

Feldman, Ronen; Sanger, James: The Text Mining Handbook- Advanced Approaches in Analyzing Unstructured Data. Cambridge: Cambridge University Press, 2007.

Felten, Edward W.: Written Testimony of Edward W. Felten. As part of: United States Senate, Committee on the Judiciary: Hearing on Continued Oversight of the Foreign Intelligence Surveillance Act, October 2, 2013.
Online unter: http://www.cs.princeton.edu/~felten/testimony-2013-10-02.pdf
(abgerufen am 17. Oktober 2016)

Fiedler, Steffen: OECD Better Life Index. Berlin: Studio NAND.
Online unter: http://www.oecdbetterlifeindex.org/de/
(abgerufen am 2. November 2016)

Finberg, Howard; Klinger, Lauren: Core Skills for the future of journalism. St. Petersburg, USA: The Poynter Institute for Media Studies. 2014.
Online unter: http://www.newsu.org/course_files/CoreSkills_FutureofJournalism2014v5.pdf
(abgerufen am 29. August 2016)

Fischer, Eric: A Day of Muni, 2010. In: Visual Simplexity – Die Darstellung großer Datenmengen; Nix, Markus (Hrsg.). Frankfurt: Entwickler.press, 2014, S. 113.

Frees, Beate; Koch, W.: Internetnutzung: Frequenz und Vielfalt nehmen in allen Altersgruppen zu. Ergebnisse der ARD/ZDF-Onlinestudie 2015. In: Fachzeitschrift Media Perspektiven. Frankfurt am Main: AS&S GmbH, S. 369, 2015. Online unter: http://www.ard-zdf-onlinestudie.de/fileadmin/Onlinestudie_2015/0915_Frees_Koch.pdf
(abgerufen am 2. Juni 2016)

Free Software Foundation: The Free Software Foundation (FSF) is a nonprofit with a worldwide mission to promote computer user freedom. Boston, USA: 2016.
Online unter: https://www.fsf.org/about/
(abgerufen am 18. Juli 2016)

Frey, Carl B.; Osborne, Michale A.: Online unter: The Future of Employment: How susceptible are jobs to Computerisation? Oxford, United Kingdom: University of Oxford, Department of Engineering Science, 2013. Online unter:
http://www.oxfordmartin.ox.ac.uk/downloads/academic/The_Future_of_Employment.pdf
(abgerufen am 5. Juli 2016)

Friedmann, George: About Stratfor: Intelligence vs. Journalism. Austin (Texas/USA) : Stratfor, 1. Januar 2012. Online unter: https://www.stratfor.com/video/about-stratfor-intelligence-vs-journalism
(abgerufen am 1. November 2016)

Graham, Mark; Hale, Scott A.; Gaffney, Devin: Where in the world are you? Geolocation and language identification in Twitter. Oxford Internet Institute, University of Oxford, 2013.
Online unter: https://arxiv.org/pdf/1308.0683v1.pdf
(abgerufen am 2. November 2016)

Gray, Jonathan; Bounegru, Liliana; Chambers, Lucy: The Data Journalism Handbook – How Journalists Can Use Data to Improve the News. Sebastopol: O'Reilley Media, 2012.

Gabler, Wirtschaftslexikon: Springer Gabler Verlag, Stichwort: Gamification.
Online unter: http://wirtschaftslexikon.gabler.de/Archiv/688938796/gamification-v4.html
(abgerufen am 2.August.2016)

Hackmack, Gregor: Demokratie einfach machen – Ein Update für unsere Politik.
Hamburg: edition Körber-Stiftung, 2014.

Han, Jiawei; Kamber, Micheline; Pei, Jian Datamining: concepts and techniques.
Amsterdam: Morgan Kaufmann, 2012.

Harrison, Chris: Web Trigrams, Visualizing Google's Tri-Gram Data. Pittsburgh, 2006.
Online unter: http://chrisharrison.net/index.php/Visualizations/WebTrigrams
(abgerufen am 2. November 2016)

Hayes, Gary: 16 Top Augemented Reality Business Models. In: Trends der IT. Anett Mehler-Bicher und Lothar Steiger (Hrsg.). Mainz, University of Applied Sciences, 2012.

Heise, Christian: Erfolg für Open Data. Datenlizenz Deutschland Version 2.0 ist eine offene Lizenz (Update). Berlin: Open Knowledge Foundation, 2014a.
Online unter: https://okfn.de/en/blog/2014/09/erfolg-fuer-open-data-datenlizenz-deutschland-version-2-0-ist-eine-offene-lizenz/
(abgerufen am 28. Juli 2016)

Heise, Christian: Sehr geehrten Damen und Herren Abgeordneten, tun Sie endlich etwas für offene (Verwaltungs-)Daten! In: Jahrbuch Netzpolitik 2014. Lüneburg: Center for Digital Cultures, Leuphana Universität, 2014b. Online unter:
http://christianheise.de/files/2014/12/heise_open_data_preprint_netzpolitik_2014.pdf
(abgerufen unter 28. Juli 2016)

Heise, Christian: Open Data – Die wichtigsten Fakten zu offenen Daten. Berlin: Konrad-Adenauer-Stiftung, 2016.
Online unter: http://www.kas.de/wf/doc/kas_44530-544-1-30.pdf?160315122244
(abgerufen unter 18. Juli 2016)

Holovaty, Adrian: A fundamental way newspaper sites need to change, 2006.
Online unter: http://www.holovaty.com/writing/fundamental-change/
(abgerufen am 20. Juni 2016)

Hopcroft, John E.; Motwani, Rajeev; Ullman, Jeffrey D.: Einführung in die Automatentheorie, Formale Sprachen und Komplexitätstheorie. Bonn: Addison-Wesley Longman Verlag, 2. Auflage, 2002.

Howard, Alexander B.: The Art and Science of Data-Driven Journalism. When Journalists combine new technology with narrative skills, they can deliver context, clarity, and a better understanding of the world around us. In: The Tow Center for Digital Journalism. A Tow/Knight Report. New York: Columbia University, 2014. Online unter: http://towcenter.org/wp-content/uploads/2014/05/Tow-Center-Data-Driven-Journalism.pdf
(abgerufen am 2. Juni 2016)

IDC: The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things. Framingham: Whitepaper sponsored by EMC Digital Universe with Research & Analysis, 2014.
Online unter: http://www.idcdocserv.com/1678
(abgerufen am 20. Juni 2016)

Jakubetz, Christian: DDJ: Als die Daten laufen lernten...Königswinter: ABZV Universalcode, 2013.
Online unter: http://universal-code.de/2013/10/17/ddj-als-die-daten-laufen-lernten/
(abgerufen am 9. August 2016)

Keim, Daniel: Datenvisualisierung und Data Mining. Konstanz/Florham Park; Universität Konstanz, AT&T Shannon Research Labs, 2002.
Online unter: http://fusion.cs.uni-magdeburg.de/pubs/spektrum.pdf
(abgerufen am 22. Juli 2016)

Krabina, Bernhard: Vorgehensmodell für Open Government Data. In: Hilgers, Dennis / Schauer, Reinbert / Thom, Norbert (Hrsg.): Public Management im Paradigmenwechsel. Linz: Trauner Verlag, 2012, S. 279-287.

Kramp, Leif; Weichert, Stephan: Die Zeitungsmacher – Aufbruch in die digitale Moderne. Verlag Springer VS, 2015, S. 52-55. In: medium magazin: Schafft neue Berufsrollen! Freilassing: Johann Oberauer GmbH, 2015.

Kuzev, Dr. Pencho: Open Data – Die wichtigsten Fakten zu offenen Daten. Berlin: Konrad-Adenauer-Stiftung, 2016. Online unter: http://www.kas.de/wf/doc/kas_44530-544-1-30.pdf?160315122244
(abgerufen unter 18. Juli 2016)

La Roche, Walther von: Einführung in den praktischen Journalismus. Berlin: Econ/Ullstein Buchverlage GmbH, 18. Aufl., 2008.

Lewis, Seth C.: Journalism in a Era of Big Data. Cases, concepts and critiques. In: Digital Journalism, Vol. 3., Routledge Taylor Group, 2015.
Online unter: http://www.tandfonline.com/doi/full/10.1080/21670811.2014.976399
(abgerufen am 6. August 2016)

Loukides, Mike: Overfocus on Tech Skills Could Exclude the Best Candidates for Jobs. Sebastopol: O'Reilly Media, 2012. Online unter: https://www.oreilly.com/ideas/overfocus-on-tech-skills-could-exclude-the-best-candidates-for-jobs
(abgerufen am 5. Juli 2016)

Lucke, Jörn von; Geiger, Christian P.: Open Government Data – Frei verfügbare Daten des öffentlichen Sektors (Gutachten für die Deutsche Telekom AG zur T-City Friedrichshafen) / Friedrichshafen: Deutsche Telekom Institute for Connected Cities, Zeppelin University, 2010.

Lucke, Jörn von: Innovationsschub durch Open Data, Datenportale und Umsetzungswettbewerbe. In: Schauer, Reinbert (Hrsg.) ; Thom, Norbert (Hrsg.) ; Hilgers, Dennis (Hrsg.): Innovative Verwaltungen – Innovationsmanagement als Instrument von Verwaltungsreformen. Linz: Trauner Verlag, 2011.

Lynch, Dianne: Above & Beyound. Looking at the Future of Journalism Education. Miami: Knight Foundation, 2015. Online unter: http://www.knightfoundation.org/media/uploads/publication_pdfs/KF-Above-and-Beyond-Report.pdf
(abgerufen am 5. Juli 2016)

Maas, Marco: Lernt endlich Technik! In: medium magazin. Freilassing: Johann Oberauer GmbH, 2015.

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

Manyika, James et al.: Big Data: The Next Frontier for Innovation, Competition, and Productivity. McKinsey, Global Institute, 2011. Online unter: www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation (abgerufen am 1. Juni 2016)

Matzat, Lorenz: Datenjournalismus vor dem Internet: Wetterbericht, Finanzdaten und Co. In: Datenjournalist data-driven-journalism & interaktiver Journalismus. Berlin: Lorenz Matzat, 2010. Online unter: http://datenjournalist.de/datenjournalismus-vor-dem-internet-wetterbericht-finanzdaten-und-co/
(abgerufen am 20. Juni 2016)

Matzat, Lorenz: Open Data - Datenjournalismus. Berlin: Bundeszentrale für politische Bildung, 2011. Online unter: http://www.bpb.de/gesellschaft/medien/opendata/64069/datenjournalismus. 2011. (abgerufen am 2. August 2016)

Matzat, Lorenz: Datenjournalismus: Methoden, Prozesse und Kompetenzen. In: Fachjournalist. Fach-Journalismus, Fach-PR und Fachmedien. Berlin: DFJV Deutscher Fachjournalisten-Verband AG, 2014. Online unter: http://www.fachjournalist.de/datenjournalismus-methoden-prozesse-und-kompetenzen/
(abgerufen am 10. Juli 2016)

Matzat, Lorenz: Aus Datenjournalismus sollte Journalismus über Daten werden. In: Datenjournalist data-driven-journalism & interaktiver Journalismus. Berlin: Lorenz Matzat, 2015.
Online unter: http://datenjournalist.de/aus-datenjournalismus-sollte-journalismus-ueber-daten-werden/
(abgerufen am 3. August 2016)

Maycotte, Higinio: Big Data Triggers Predictive Journalism. Cambridge: Nieman Foundation at Harvard University, 2016. Online unter: http://www.niemanlab.org/2015/12/big-data-triggers-predictive-journalism/
(abgerufen am 15. Juli 2016)

Mccandless, David: The Middle East – Key players and notable relationships.
In: Knowledge is Beautiful project. London: Universlab.
Online unter: http://www.informationisbeautiful.net/visualizations/the-middle-east-key-players-notable-relationships/
(abgerufen am 18. Oktober 2016)

McLaughlin, Catriona: Die Besserwisser. Acxiom hortet Informationen über 300 Millionen Amerikaner und bereits 44 Millionen Deutsche. Aber was genau macht das Unternehmen damit? In: Zeit Online, Datenschutz, 2013. Online unter: http://www.zeit.de/2013/28/acxiom
(abgerufen am 3. August 2016)

Müller, Matthias: Personal Communication, Hamburg, 2. November 2016.

Nix, Markus: Visual Simplexity, Die Darstellung großer Datenmengen. Frankfurt: Entwickler.press, 2014.

Neumüller, Fritz; Kahn, Eram: Perlentauer im Datenmeer. Sieben Thesen zu Datenjornalismus im New News Process. In: kommunikation.medien. Onlinejournal des Fachbereichs Kommunikationswissenschaft Universität Salzburg, 2015.
Online unter: http://journal.kommunikation-medien.at/wp-content/uploads/2015/03/Ausg5_Khan_Neum%C3%BCller.pdf
(abgerufen am 2. Dezember 2016)

Obermayer, Bastian; Wormer, Vanessa; Jaschensky, Wolfgang: Panama Papers – die Geheimnisse des schmutzigen Geldes. München: Süddeutsche Zeitung GmbH, 2016.
Online unter: http://panamapapers.sueddeutsche.de/articles/56ff9a28a1bb8d3c3495ae13/
(abgerufen am 14. Juli 2016)

OpenDataCity: Wie das Smarthome unser Leben aufzeichnet. Datenfreunde GmbH, Berlin 2015.
Online unter: https://opendatacity.de/project/wie-das-smarthome-unser-leben-aufzeichnet/
(abgerufen unter: 10. Juli 2016)

Open Knowledge Foundation Deutschland: Offene Daten. Berlin: Open Knowledge Foundation Deutschland e.V., 2016. Online unter: https://okfn.de/themen/offene-daten/
(abgerufen am 15. Juli 2016)

O´Reilly, Tim: What is web 2.0 ?, Sebastopol, USA: O'Reilley Media, 2007.

Parlamentarischer Rat von 1949: Grundgesetz der Bundesrepublik Deutschland. Berlin: Deutscher Bundestag, 2015. Online unter: https://www.btg-bestellservice.de/pdf/10060000.pdf
(abgerufen am 10 Juli 2016)

Peters, Diane: Improving Access to the Public Domain: the Public Domain Mark. Creative Commons Blog. 2010. Online unter: https://creativecommons.org/2010/10/11/improving-access-to-the-public-domain-the-public-domain-mark/
(abgerufen am 26. Juli 2016)

Petersohn, Helge: Data Mining. Verfahren, Prozesse, Anwendungsarchitektur. München: Oldenbourg Wissenschaftsverlag, 2005.

Peukert, Alexander: Die Gemeinfreiheit – Begriff, Funktion, Dogmatik. Tübingen: Mohr Siebeck, 2012.

Philipsen, Quirine: Explorative Datenvisualisierung. Hamburg: HAW, Fakultät Technik und Informatik der Hochschule für Angewandte Wissenschaften, 2015. Online unter: http://users.informatik.haw-hamburg.de/~ubicomp/projekte/master-nm-rv-2015/philipsen.pdf
(abgerufen am 10. September 2016)

Philipsen, Quirine: The Game-Universe for Non-Gamer. Das 21. Jahrhundert voller Spielkinder. Hamburg: HAW, Fakultät Technik und Informatik der Hochschule für Angewandte Wissenschaften, 2016. Online unter: http://users.informatik.haw-hamburg.de/~ubicomp/projekte/master-nm-2016-sem/philipsen/bericht.pdf
(abgerufen am 10. September 2016)

Prado, Claudio: "Eine Peer-to-Peer Gesellschaft ist möglich". In: Dossier, Open Source, Interview von Deterding, Sebastian. Bonn: Bundeszentrale für politische Bildung, 2007.
http://www.bpb.de/gesellschaft/medien/opensource/63914/claudioXprado
(abgerufen am 18. Juli 2016)

Pressekodex: Publizistische Grundsätze. Richtlinien für die publizistische Arbeit nach den Empfehlungen des Deutschen Presserats, Berlin: Pressrat, 2015. Online unter:
http://www.presserat.de/fileadmin/user_upload/Downloads_Dateien/Pressekodex_BO_2016_web.pdf
(abgerufen am 10. Juli 2016)

Reda, Julia: IGEL – Initiative gegen ein Leistungsschutzrecht. Berlin: Dr. Till Kreutzer. 17.12. 2014. Online unter: http://leistungsschutzrecht.info
(abgerufen am 16. Oktober 2016)

Reznicek, Marc: Guidelines NER. Linguistische Annotation von Nichtstandardvarietäten — Guidelines und „Best Practices". Berlin: Humboldt Universität, 2013. Online unter: https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/nosta-d/nosta-d-ner-1.5
(abgerufen am 27. Oktober 2016)

Redman, Thomas. C.: The Impact of Poor Data Quality on the Typical Enterprise. In: Communications of the ACM, Vol. 41 No. 2 1998, S. 79-82. Online unter: http://dl.acm.org/citation.cfm?id=269025&CFID=826748054&CFTOKEN=75798409
(abgerufen am 17. August 2016)

Rey, Günter Daniel; Wender, Karl F.: Neuronale Netze – Eine Einführung in die Grundlagen, Anwendungen und Datenauswertungen. Göttingen: Hogrefe, vorm. Verlag Hans Huber; Auflage: 2, 2010. Online unter: http://www.neuronalesnetz.de
(abgerufen am 2. November 2016)

Robinson, David G./Yu, Harlan (2012): The New Ambiguity of „Open Government". 2012.
Online unter: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2012489
(abgerufen am 30. Juli 2016)

Rodgers, Simon: Open Data Journalism. London, United Kingdom, 2013.
Online unter: https://simonrogers.net/2013/01/24/open-data-journalism/
(abgerufen unter: 1. August 2016)

Rodgers, Simon: How we made a VR Data Visualization. In: Data Journalism and other curiosities. London: Simon Rodgers, 2016.

Scharloth, Joachim: Überwachen und Sprache. How do things with words. Vortrag im Rahmen des 30. Chaos Communication Congress [30c3], Hamburg (CCH), Chaos Computer Club [CCC]. 29.12. 2013.
Online unter: https://www.youtube.com/watch?v=NxJ1uT1DlVo
(abgerufen am 16. Oktober 2016)

Schneller, Dr. Johannes: Allensbacher Markt- und Werbeträgeranalysen (AWA). Veränderung der Mediennutzung. Institut für Demoskopie Allensbach: 2016. Online unter: http://www.ifd-allensbach.de/fileadmin/AWA/AWA_Praesentationen/2016/AWA_2016_Schneller_Medien.pdf
(abgerufen am 17. September 2016)

Schroll, Willi.: Augmented Reality – Ein Urknall steht bevor, 2010. In: Trends der IT.
Anett Mehler-Bicher und Lothar Steiger (Hrsg.). Mainz, University of Applied Sciences, 2012.

Schumann, Heidrun; Müller, Wolfgang: Visualisierung. Grundlagen und allgemeine Methoden. Heidelberg: Springer Verlag, 2000.

Schwenke, Thomas: Basiswissen Journalismus: Presserecht für Journalisten und Blogger. In: Upload Magazin, E-Business, Social Media und die Internetwirtschaft. Berlin, 2013.
Online unter: http://upload-magazin.de/blog/715-basiswissen-journalismus-presserecht-fur-journalisten-und-blogger/
(abgerufen am 23. Juli 2016)

Selle, Stephan: „Big Data" werden Grenzen gesetzt. In: Bookbytes. Blog für Digitales, Börsenblatt.net, 2016. Online unter: https://www.boersenblatt.net/bookbytes/artikel-neue_eu-datenschutzrichtlinien.1232559.html
(abgerufen am 27. November 2016)

Severo, M.; Giraud, T.; Pecout, H.: Twitter data for urban policy making: an analysis on four European cities. In C. Levallois (Ed.), Handbook of Twitter for Research. Lille: Emlyon, 2015. Online unter: https://www.researchgate.net/publication/279175120_Twitter_data_for_urban_policy_making_an_analysis_on_four_European_cities
(abgerufen am 10. November 2016]

Shannon, Claude E.; Weaver, Warren: The Mathematical Theory of Communication. Urbana, Chicago: University of Illinois Press. 1968.

Sharafi, Armin: Knowledge Discovery in Databases - Eine Analyse des Ände- rungsmanagements in der Produktentwicklung. Wiesbaden: Springer Gabler, 2013.

Sonderman, Jeff: Programmers Explain How to Turn Data into Journalism & Why That Matters. St. Pertersburg, Florida: The Poynter Institute, 13 Jan. 2013. Online unter: http://www.poynter.org/2013/programmers-explain-how-to-turn-data-into-journalism-why-that-matters-after-gun-permit-data-publishing/199834/
(abgerufen am 16. Juli 2016)

Spitz, Malte: Sechs Monate meines Lebens in 35.000 Datensätzen. In: Malte Spitz Blog, Berlin, 2011a. Online unter: http://malte-spitz.de/2011/02/24/sechs-monate-meines-lebens-in-35-000-datensaetzen/
(abgerufen am 2. August 2016)

Spitz, Malte: Dem Speicherwahn ein Ende setzen. Berlin. In: Pressemitteilung vom Pressedienst Bündnis 90 / Die Grünen, Berlin, 2011b. Online unter: http://www.artikel-presse.de/dem-speicherwahn-ein-ende-setzen.html
(abgerufen am 2. August 2016)

Stencel, Mark; Adair, Bill; Kamalakanthan Prashanth: The Goat Must Be Fed – Why digital tools are missing in most newsrooms. Durham: Duke Reporter'Lab, Sanford School of Public Policy, 2014. Online unter: http://www.goatmustbefed.com/resources/pdf/goat-must-be-fed.pdf (abgerufen am 24. Juni 2016)

Stray, Jonathan: The Curious Journalist's Guide to Data. In: The Tow Center for Digital Journalism. A Tow/Knight Report. New York: Columbia University, 2016. Online unter: https://www.gitbook.com/book/towcenter/curious-journalist-s-guide-to-data/details (abgerufen am 2. Juni 2016)

Tillack, Hans-Martin: Transparenz in Politik und Medien. In: Medien müssen draußen bleiben! Wo liegen die Grenzen politischer Transparenz? (Beiträge zur 8. Fachtagung des DFPK), Roger, Franziska Bravo / Henn, Philipp / Tuppack, Diana (Hrsg.). Berlin: Frank & Timme GmbH Verlag für wissenschaftliche Literatur, 2013.

Thurner, Thomas: Open Goverment Data Weißbuch. Östereich: Universitätsverlag der Donau-Universität Krems, Kaltenböck, Martin; Thurner, Thomas (Hrsg.), 2011.

Tufte, Edward R.: The Visual Diplay of Quantitative Information. Cheshire, Connecticut: Graphics Press, 1983.

Turkey, John W.: Exploratory Data Analysis. Boston: Addison Wesley, 1977.

Wang, Richard Y.; Strong, Diane M.: Beyond accuracy: what data quality means to data consumers. In: Journal of Management and Informations Systeme, Ausgabe 12. New York: M. E. Sharpe Inc., 1996, 5-33. Online unter: http://mitiq.mit.edu/Documents/Publications/TDQMpub/14_Beyond_Accuracy.pdf (abgerufen am 20. August 2016)

Weiss, Sholom M.; Indurkhya, Nitin; Zhang, Thong; Damerau, Fred J.: Text Mining. Predictive Methods for Analyzing unstructured Information. New York: Springer, 2005.

WikiLeaks: In Wikipedia, Die freue Enzyklopädie. Bearbeitungsstand: 25. Juni 2016. Online unter: https://de.wikipedia.org/w/index.php?title=WikiLeaks&olsid=155614882 (abgerufen am 4. Juli 2016)

Wikipedia: Cholera Map. Eintrag John Snow. Online im Internet: http://de.wikipedia.org/wiki/John_Snow_(Arzt) (abgerufen am 10. August 2016)

Wissens-Portal ITwissen.info: Chomsky-Hierarchien. Peterskirchen: DATACOM Buchverlag GmbH. 2016. Online unter: http://www.itwissen.info/definition/lexikon/Chomsky-Hierarchie.html (abgerufen 15.Oktober 2016)

Wolf, Prof. Dr. Patricia: Interdisziplinäre Innovation braucht Freiraum. Frankfurt am Main: Zukunftsinstitut GmbH, 12.2015. Online unter: https://www.zukunftsinstitut.de/artikel/tup-digital/06-innovation-gap/02-shortcuts/interdisziplinaere-innovation-braucht-freiraum/ (abgerufen am 17. Oktober 2016)

Zörner, Hendrik: Sind doch nur ein paar Füller. In: DJV Blog. Berlin, Deutscher Journalisten Verband, 2016. Online unter: https://www.djv.de/startseite/service/blogs-und-intranet/djv-blog/detail/article/sind-doch-nur-ein-paar-fueller.html (abgerufen am 20. August 2016)

7 Abbildungsverzeichnis

Versicherung über Selbstständigkeit

Hiermit versichere ich, dass ich die vorliegende Arbeit im Sinne der Prüfungsordnung ohne fremde Hilfe selbstständig verfasst und nur die angegebenen Hilfsmittel benutzt habe.

Hamburg, 06.03.2017

Ort, Datum

Unterschrift