Hochschule für Angewandte Wissenschaften Hamburg
Hamburg University of Applied Sciences

Bachelorarbeit

Vasilis Ikonomou

Open Data basierte digitale narrative Strukturen

Fakultät Technik und Informatik
Studiendepartment Informatik

Faculty of Engineering and Computer Science
Department of Computer Science

Vasilis Ikonomou

Open Data basierte digitale narrative Strukturen

Bachelorarbeit eingereicht im Rahmen der Bachelorprüfung

im Studiengang Bachelor of Science Angewandte Informatik
am Department Informatik
der Fakultät Technik und Informatik
der Hochschule für Angewandte Wissenschaften Hamburg

Betreuender Prüfer: Prof. Dr. Kai von Luck
Zweitgutachter: Dr.-Ing. Sabine Schumann

Eingereicht am: 2. März 2015

Vasilis Ikonomou

Thema der Arbeit
Open Data basierte digitale narrative Strukturen

Stichworte
Open Data, Open Government Data, Datenjournalismus, Digital Storytelling, Datenvisualisierung, Data Mining

Kurzzusammenfassung
Seit der Digitalisierung sind Daten zum neuen Rohstoff der modernen Gesellschaft geworden, der ein hohes Potenzial in sich birgt. Dieses Potenzial haben inzwischen auch Unternehmen, Organisationen und Einrichtungen des öffentlichen Sektors entdeckt, die aus diesem Anlass begonnen haben, ihre Daten der Öffentlichkeit als Open Data zur freien Weiterverwendung und -verbreitung anzubieten, um Raum für Innovationen zu schaffen. Im Bereich des Journalismus hat sich seitdem der Begriff Datenjournalismus als neues Genre der Berichterstattung geformt, der Gebrauch von diesen Daten macht und sie als Material zur Erzählung von Geschichten einsetzt. Diese Arbeit untersucht aus der Sicht der IT den Stand der Entwicklung von Open Data (insbesondere aus dem öffentlichen Sektor) sowie des Datenjournalismus. Sie versucht außerdem, die Zuständigkeitsbereiche und Rollen von IT und Journalismus innerhalb von diesem Kontext zu erfassen, sowie auch die noch offenen Herausforderungen in diesen Gebieten aufzuzeigen.

Vasilis Ikonomou

Title of the paper
Open Data based digital narrative structures

Keywords
Open Data, Open Government Data, Data-driven journalism, Digital Storytelling, Data visualization, Data Mining

Abstract
Since the digitization, data has turned into the new raw material of modern society which carries a lot of potential for different fields. This potential has been discovered by companies, organizations and public sector institutions, who have started for this occasion to offer their data to the public as Open Data for free reuse and redistribution, in order to make room for innovation. In the field of journalism, the term data-driven journalism has formed since then as a new genre of reporting, that makes use of this data by employing it as a material for the telling of stories. From the perspective of IT, this thesis studies the current state of development of Open Data (especially from the public sector) and data-driven journalism. Furthermore, it attempts to understand the scope and the roles of IT and journalism within this context, and to highlight outstanding emerging challenges in these areas.

Inhaltsverzeichnis

1. Einleitung **1**
- 1.1. Hintergrund . . . 1
- 1.2. Motivation und Zielsetzung . . . 2
- 1.3. Abgrenzung . . . 3
- 1.4. Aufbau der Arbeit . . . 4

2. Open Data **6**
- 2.1. Einleitung . . . 6
- 2.2. Grundbegriffe . . . 6
  - 2.2.1. Daten, Informationen und Wissen . . . 6
  - 2.2.2. Open Data . . . 8
  - 2.2.3. Open Government und Open Government Data . . . 9
  - 2.2.4. Public Sector Information und Informationsfreiheit . . . 10
- 2.3. Potenziale von Open Data . . . 12
- 2.4. Situation offener Verwaltungsdaten in Deutschland und im Ausland . . . 13
  - 2.4.1. Situation in Deutschland . . . 14
  - 2.4.2. Situation auf internationaler Ebene . . . 15
- 2.5. Herausforderungen und Fazit . . . 16

3. IT-Maßnahmen zur Bereitstellung von Datenbeständen **19**
- 3.1. Einleitung . . . 19
- 3.2. Umgang mit großen Datensätzen . . . 20
  - 3.2.1. Big Data . . . 21
  - 3.2.2. Die Dimensionen von Big Data . . . 22
  - 3.2.3. Technologien . . . 23
    - 3.2.3.1. NoSQL Datenbanken . . . 23
    - 3.2.3.2. Apache Hadoop Framework . . . 24
  - 3.2.4. Schlussfolgerung . . . 25
- 3.3. Plattformen zur Bereitstellung von Open Data . . . 25
  - 3.3.1. Erfassung der allgemeinen Anforderungen . . . 26
  - 3.3.2. Vorhandene Softwarelösungen für Open Data Portale . . . 27
  - 3.3.3. Comprehensive Knowledge Archive Network . . . 28
    - 3.3.3.1. Architektur und Features von CKAN . . . 29
    - 3.3.3.2. Zugriff auf Datensätze . . . 32
    - 3.3.3.3. Skalierbarkeit . . . 33

3.3.4. Zwischenfazit . . . 35
3.4. Verknüpfung von Datenbeständen . . . 35
3.4.1. Linked Data und Linked Open Data . . . 36
3.4.2. Das Datenmodell . . . 37
3.4.3. Technologien und Standards . . . 38
3.4.3.1. Uniform Resource Identifier (URI) . . . 38
3.4.3.2. Hypertext Transfer Protocol (HTTP) . . . 39
3.4.3.3. Resource Description Framework (RDF) . . . 39
3.4.3.4. RDF Schema (RDFS) und Web Ontology Language (OWL) . . . 39
3.4.3.5. SPARQL Protocol and RDF Query Language . . . 39
3.4.4. Linked Data in der Praxis . . . 40
3.4.5. Zwischenfazit . . . 41
3.5. Fazit . . . 42

4. Interpretation von Datenbeständen 44
4.1. Einleitung . . . 44
4.2. Knowledge Discovery in Databases und Data Mining . . . 45
4.3. Ablauf einer Datenanalyse . . . 46
4.3.1. Selektion . . . 48
4.3.2. Datenvorverarbeitung . . . 48
4.3.3. Transformation . . . 49
4.3.4. Data Mining . . . 49
4.3.5. Interpretation und Evaluation . . . 50
4.3.5.1. Bedeutung von Visualisierung . . . 51
4.4. Anwendungsklassen . . . 52
4.4.1. Bedeutung von Abstands- und Ähnlichkeitsmaßen . . . 53
4.4.2. Cluster-Analyse . . . 53
4.4.3. Klassifikation . . . 55
4.4.4. Assoziationsanalyse . . . 55
4.4.5. Text Mining . . . 56
4.4.6. Web Mining . . . 57
4.5. Fazit . . . 57

5. Daten-gestützte narrative Strukturen 59
5.1. Einleitung . . . 59
5.2. Die Medien im Wandel . . . 60
5.3. Der Journalismus im Wandel . . . 61
5.4. Datengetriebener Journalismus . . . 62
5.4.1. Definition von Datenjournalismus . . . 62
5.4.2. Bedeutung von Datenanalyse und -aufbereitung . . . 63
5.4.3. Erzählformen und Methoden . . . 64
5.4.3.1. Datastorytelling . . . 65

5.4.3.2. Echtzeitdaten 65
5.4.3.3. Datensätze 66
5.4.3.4. Crowdsourcing 67
5.4.3.5. Hyperlokal 67
5.4.3.6. Newsgames 67
5.5. Die Rolle von IT in der Redaktion 68
5.6. Die Rolle des Journalisten 69
5.7. Fazit 70

6. Schlussbetrachtung **72**
6.1. Zusammenfassung 72
6.2. Ausblick 74

A. Anhang **76**
A.1. API-Beispiel für den Abruf eines Datensatzes in CKAN 76

Abbildungsverzeichnis **83**

Literaturverzeichnis **84**

1. Einleitung

1.1. Hintergrund

Die Verbreitung von Wissen hat in unseren Gesellschaften schon immer eine wichtige und vor allem einflussreiche Rolle gespielt. Die heutigen Entwicklungen und wissenschaftlichen Erkenntnisse sind nicht zuletzt dem kontinuierlichen Informationsfluss zu verdanken, der sich durch alle Zeitalter und Generationen hin erstreckt. Seit der Digitalisierung sind Daten zu einem modernen Rohstoff geworden (Barnickel und Klessmann, 2012, S. 127), der in vielfacher Weise eine gewisse Veredelung durch die Weiterverarbeitung und -verteilung erfährt, und somit zur Generierung von neuem Wissen beiträgt. Durch die Entwicklungen der Informationstechnologie vorangetrieben, sind besonders in den vergangenen Jahren die Mengen an erzeugten Daten ins Unermessliche gestiegen; nicht umsonst wird deshalb auch gerne von „Datenbergen“ gesprochen. Das Interesse an den Daten ist ohne Frage groß, denn die dahinter verborgenen Informationen besitzen nicht nur ein hohes Potenzial, das sich in den verschiedensten Anwendungsbereichen entfalten kann, sondern es lässt sich mit ihnen auch Profit erzielen.

Unterdessen ist in den modernen Gesellschaften ein ausgeprägtes Verlangen nach „Offenheit“ und Transparenz zu beobachten. Die Open Source Bewegung hat in diesem Zusammenhang als eine der ersten Initiativen im Bereich der Informationstechnologie den Horizont für eine transparentere digitale Welt geöffnet. Sie hat den ungehinderten Zugang zu Software und den dahinter liegenden Quellcode unterstützt (Open Source Initiative, 2015), und damit den Grundstein für das Entstehen weiterer Konzepte gelegt, die alle in ihrem Kern auf Offenheit abzielen; Open Content, Open Access, Open Education und Open Data sind nur ein paar entsprechende Beispiele. Die grundlegende Idee dieser Offenheitskonzepte ist es, den Zugang zu Werken und dessen anschließende Weiternutzung einfacher zu gestalten, um im Weiteren die Beteiligung der Gesellschaft zu fördern und Raum für Innovationen zu schaffen. So ist bereits heute zu beobachten,

dass immer mehr Dienstleistungen auf der Basis von offenen Werken entstehen. Die Roadworks Database[1] nutzt beispielsweise offene Daten der britischen Regierung, um auf einer Karte aktuelle Baustellen und damit verbundene Verkehrsbehinderungen in Großbritannien aufzuzeigen. Neben dem genannten Beispiel gibt es noch eine Vielzahl weiterer Informationsprodukte und -dienstleistungen, die auf den Möglichkeiten der Analyse und Visualisierung von Daten beruhen. Im Bereich des Journalismus zum Beispiel, hat sich in den letzten Jahren der Begriff Datenjournalismus als ein neues Genre der Berichterstattung geformt, der Gebrauch von offenen Daten macht und diese als Ausgangsbasis und Gegenstand zur Erzählung einer Geschichte einsetzt. Einen Durchbruch für dieses Genre gab es den Ausführungen von Lorenz Matzat zufolge im Jahr 2010, mit der Veröffentlichung der Kriegstagebücher der US-Armee in Afghanistan durch Wikileaks (Matzat, 2014). Durch die Verwendung interaktiver Grafiken und Karten wurden der Öffentlichkeit die großen Mengen an Informationen digital veranschaulicht und zugänglich gemacht.

1.2. Motivation und Zielsetzung

In dem genannten Zusammenhang ist gegenwärtig zu beobachten, dass speziell im Bereich des öffentlichen Sektors weltweit immer mehr Regierungen und Behörden ihre Daten als Open Data zur freien Einsicht und Weiternutzung zur Verfügung stellen. Gleichzeitig nimmt aber auch die Anzahl der Redaktionen zu, die an diese Daten anknüpft, um mit ihnen neue Fragestellungen zu beantworten, die im Rahmen eines journalistischen Beitrags mit neuen Erzählformen an den Leser gebracht werden. Die vorliegende Arbeit nimmt diese erhöhte Aufmerksamkeit zum Anlass, um die noch relativ neue Entwicklung des Open Data Konzepts aber auch des Datenjournalismus zu untersuchen und strukturiert wiederzugeben. Dabei wird vom Standpunkt der Informationstechnologie aus der Blick auf die verschiedenen Aktivitäten bzw. Teilaspekte gerichtet, die in dem gesamten Kontext von Belang sind – begonnen von der Bereitstellung von Open Data, dessen Akquise, Analyse und Interpretation bis hin zu der Verwendung der gewonnenen Informationen in narrativen Strukturen. Der Fokus liegt dabei vor allem auf der technischen Infrastruktur, d.h. den Arbeitswerkzeugen und -umgebungen, die hier zum Einsatz kommen. Neben einem Bericht zum Stand der Umsetzung bzw. Entwicklung, versucht diese Arbeit auch einen Überblick über die

[1] http://www.roadworks.org/

Zuständigkeitsbereiche sowie die Rollen von IT und Journalismus innerhalb des genannten Kontextes zu geben, und schließlich die noch offenen Herausforderungen in diesen Gebieten aufzuzeigen.

Diese Arbeit soll nicht nur Informatikern, sondern auch angehenden sowie bereits im Beruf stehenden Journalisten, die Potenziale und möglichen Herausforderungen innerhalb des thematischen Rahmens dieser Arbeit vor Augen führen. Ziel ist schlussendlich auch die Schaffung von einem gegenseitigen Verständnis zwischen den Gebieten IT und Journalismus.

1.3. Abgrenzung

Diese Arbeit legt den Fokus primär auf Open Data aus dem Bereich des öffentlichen Sektors, und befasst sich aus diesem Grund im zweiten Kapitel mit den technischen Herausforderungen, die sich speziell in diesem Bereich bei der Umsetzung des Open Data Konzepts ergeben. Open Data aus dem privatwirtschaftlichen Bereich werden hingegen nicht betrachtet. Ebenso wenig werden jene Daten betrachtet, die im Zeitalter des *Ubiquitous Computing* (dt. „Rechnerallgegenwart“) von Einzelpersonen etwa durch die Nutzung moderner Smartphones erzeugt werden (Dumbill, 2012, S. 3). All diese Daten können in einer unendlichen Breite von Produkten und Dienstleistungen eingesetzt werden, und dementsprechend auch Verwendung im Datenjournalismus finden, wo sie aggregiert und analysiert werden können, um als Grundlage für einen Bericht zu dienen. Die vorliegende Arbeit befasst sich an dieser Stelle allerdings nur mit datenjournalistischen Werken, die Open Data zur Berichterstattung heranziehen. Eine Beurteilung aus medienwissenschaftlicher Sicht ist des Weiteren nicht Gegenstand dieser Arbeit.

Es wird abschließend darauf hingewiesen, dass in dieser Arbeit nur ein begrenzter Einblick in die verschiedenen behandelten Themenbereiche gegeben werden kann, da eine umfangreichere Erläuterung dieser den Rahmen der Arbeit übersteigen würde. Für eine Vertiefung wird deswegen die Auseinandersetzung mit der herangezogenen Literatur empfohlen.

1.4. Aufbau der Arbeit

In Kapitel 2 werden die Grundlagen für diese Arbeit behandelt. Anfänglich werden allgemeine Grundbegriffe im Zusammenhang mit Open Data und speziell Daten des öffentlichen Sektors vermittelt. Anschließend werden die Potenziale von Open Data sowie die aktuelle Situation von offenen Verwaltungsdaten in Deutschland wie im Ausland am Beispiel von Online-Plattformen erläutert, die bereits aktiv Open Data bereitstellen. Im letzten Abschnitt sollen zuletzt die Herausforderungen aufgezeigt werden, die sich im öffentlichen Sektor bei der Bereitstellung von Open Data ergeben.

Kapitel 3 befasst sich mit den technischen Maßnahmen die für eine geeignete Bereitstellung von Open Data benötigt werden, sodass Bürger, Unternehmen, und sonstige externe Akteure auf diese zugreifen und Gebrauch davon machen können. Hierzu wird zunächst die Rolle von Big Data und entsprechenden Technologien umrissen, die einen effizienten Umgang mit großen unstrukturierten Datensätzen ermöglichen. Das darauf folgende Unterkapitel befasst sich anschließend mit Plattformen zur Bereitstellung von Open Data. Nach einer Erfassung der Anforderungen, die an geeignete Plattformen gestellt werden, werden im Weiteren Softwarelösungen für Open Data Portale genannt, die aktuell im Einsatz sind. Die bekannteste Lösung wird zum Schluss in einem gesonderten Abschnitt genauer vorgestellt. In einem weiteren Abschnitt wird zuletzt das Linked Data Konzept behandelt, das durch die Vernetzung von Datenbeständen die Wiederverwendung und Nachnutzung der Daten nachhaltig fördern soll.

Kapitel 4 beschäftigt sich mit der Wissensentdeckung in Datenbanken und geht hierzu auf das Thema Data Mining ein, das eine Reihe von bewährten Verfahren zur Interpretation von Datenbeständen mitbringt. Nach einer grundlegenden Definition des Begriffes sowie Vorstellung des zugrundeliegenden Data Mining-Prozesses, werden im Weiteren die Vorgehensweisen der verschiedenen Data Mining Verfahren übersichtsartig aufgezeigt und anhand von kurzen Beispielen näher gebracht. Ziel dieses Kapitels ist es, ein grundlegendes Verständnis für die Analyse und Interpretation von Daten zu vermitteln, und die Bedeutung der verschiedenen einzelnen Teilschritte hervorzuheben.

In Kapitel 5 wird auf das Feld des noch jungen Datenjournalismus eingegangen, der Open Data als mögliches Material zur Erzählung einer Geschichte einsetzt. Nach einer kurzen Darstellung des heutigen Medienwandels, werden anschließend der Begriff Datenjournalismus definiert und seine verschiedenen Erzählformen und Methoden be-

trachtet. Im weiteren Verlauf wird die Rolle der IT sowie auch die Rolle des Journalisten in diesem Feld diskutiert.

In Kapitel 6 findet abschließend die Schlussbetrachtung der Inhalte dieser Arbeit statt. Neben einer Zusammenfassung der gewonnenen Erkenntnisse, werden im Rahmen eines Ausblickes Fragestellungen für mögliche weiterführende Arbeiten vorgebracht.

2. Open Data

2.1. Einleitung

Das Thema Open Data zählt zu den Grundlagen dieser Arbeit. Dieses Kapitel soll dem Leser hierzu einen grundlegenden Überblick verschaffen, und dabei konkret am Beispiel von Daten öffentlicher Behörden aufzeigen, wie weit zum jetzigen Zeitpunkt der Grundsatz von Offenheit und Transparenz (insbesondere auch aus technischer Sicht) umgesetzt wurde. Der Fokus der gesamten Arbeit beschränkt sich dabei ausschließlich auf Open Data des öffentlichen Sektors.

In Abschnitt 2.2 werden zunächst die wichtigsten Begriffe im Zusammenhang mit Open Data und Daten des öffentlichen Sektors definiert. Welche Potenziale in der öffentlichen Bereitstellung von Daten liegen, und wie daraus sowohl der öffentliche Sektor selbst, als auch die Wirtschaft Nutzen ziehen können, wird darauf in Abschnitt 2.3 diskutiert. Abschnitt 2.4 versucht anschließend einen Überblick über den aktuellen Status von öffentlichen Verwaltungsdaten auf nationaler sowie internationaler Ebene zu geben, und führt hierzu verschiedene Beispiele aus der Praxis auf. Abschnitt 2.5 führt abschließend die noch ausstehenden Herausforderungen auf, und liefert ein kurzes Fazit zu den behandelten Inhalten.

2.2. Grundbegriffe

2.2.1. Daten, Informationen und Wissen

An dieser Stelle werden zunächst die zentralen Begriffe *Daten*, *Informationen* und *Wissen* definiert, die für ein Verständnis der Inhalte dieser Arbeit erforderlich sind. Die Definitionen dazu basieren im Kern auf der Publikation von Cleve und Lämmel (2014, S. 37-38).

Die in dieser Arbeit verwendete Begriffshierarchie stützt sich hierbei auf diejenige, die sich in der Fachliteratur weitgehend durchgesetzt hat. *Daten* stellen dazu zunächst die Grundlage der IT-gestützten Verarbeitung dar. Besitzt ein Datum eine Bedeutung, dann wird dieses zu einer *Information*. Steht zum Beispiel als Datum die Zahl 2 für den prozentualen Anstieg der Einwohnerzahlen einer Region im Vergleich zum Vorjahr, so wird diese Zahl zu einer Information. Gibt es nun zum Beispiel bestimmte Regeln, die ein Wachstum dieser Einwohnerzahlen auslösen, so ist an dieser Stelle die Rede von *Wissen*.

Der Begriff *Daten* wird in der Fachliteratur als eine Menge von Zeichen mit der dazugehörigen Syntax definiert. In der deutschen Sprache sind *Daten* der Plural des Wortes *Datum*, welches die Bedeutung einer Kalender- oder Zeitangabe trägt. In dieser Arbeit wird diese Bedeutung jedoch nicht verwendet; stattdessen wird die ebenso gültige Interpretation eines Datums als eine *Informationseinheit* verwendet.

Daten teilen sich dabei in *unstrukturierte*, *semistrukturierte* und *strukturierte* auf. *Unstrukturierte* Daten zeichnen sich dadurch aus, dass sich die maschinelle Extraktion von Wissen aus diesen als schwierig erweist. Typische Beispiele hierfür sind etwa Schriftstücke, die als Texte in PDF-Dateien oder in eingescannter Form als Bilder vorliegen. Webseiten die üblicherweise Text und Bilder enthalten, fallen hier allerdings aufgrund ihrer Struktur die ein automatisiertes Auslesen ermöglicht, unter die Kategorie der *semistrukturierten* Daten. Der Begriff *strukturierte* Daten bezeichnet wiederum meist relationale Datenbank-Tabellen oder Daten in Datei-Formaten mit einer ähnlich festen Struktur wie zum Beispiel CSV. Die in diesen Formaten enthaltenen Daten besitzen eine feste Reihenfolge und haben definierte Attribute und Datentypen. Durch die eindeutige Datenstruktur wird eine Sortierung, Filterung und Weiterverarbeitung der Daten erleichtert.

Weist man nun Daten eine Bedeutung zu, entstehen aus ihnen *Informationen*. Informationen stellen somit eine für den Zweck entsprechende Interpretation von Daten dar. Die Fakten, aus denen Daten bestehen, wandeln sich erst dann in Informationen um, wenn ihnen unter Berücksichtigung des Kontexts eine Bedeutung zugeordnet wird.

Die Fähigkeit von Personen, eine Information einzusetzen, wird schließlich als *Wissen* bezeichnet. Probst u. a. (2012, S. 23) definieren den Begriff noch präziser als einen Komplex von Kenntnissen und Fähigkeiten, den Personen zur Lösung von Problemen einsetzen. Daten und Informationen werden hier des Weiteren als die Grundlage

für Wissen benannt, sind aber im Gegensatz zum Wissen nicht immer an Personen gebunden.

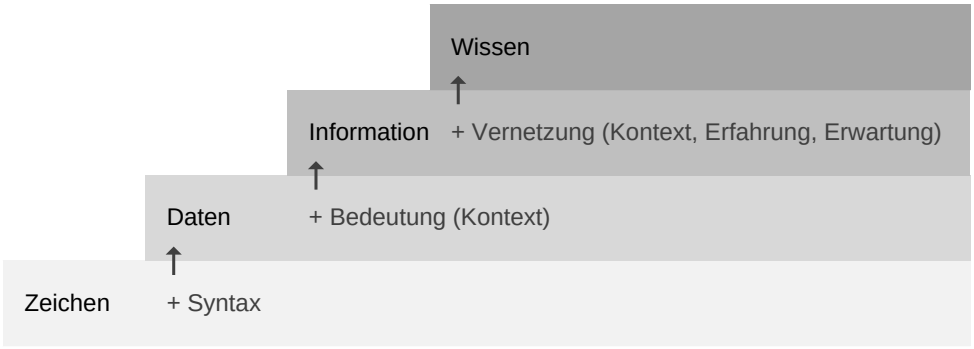

Abbildung 2.1.: Wissenstreppe nach North (2011, S. 36).

Die in Abbildung 2.1 dargestellte Wissenstreppe veranschaulicht den Zusammenhang zwischen den Begriffen Daten, Informationen und Wissen.

2.2.2. Open Data

Der Begriff *Open Data* (oder auch „offene Daten“) wird von der Open Knowledge Foundation Deutschland (2014) wie folgt definiert:

> „Offene Daten sind Daten, die von jedermann frei verwendet, nachgenutzt und verbreitet werden können – maximal eingeschränkt durch Pflichten zur Quellennennung und ‘sharealike’.“

Was unter „Offenheit“ in Bezug auf Inhalte und Daten verstanden wird, gibt die Definition für *offenes Wissen* der Open Knowledge Foundation (2014g) im Genaueren vor. Dieser zufolge ist *offenes Wissen* ein Gegenstand oder Werk, das verschiedene Kriterien erfüllt, die mehrfach auf Klauseln der Open Source Definition aufbauen. Ein solches Werk muss als Ganzes (und vorzugsweise als entgeltfreier Download im Internet) zugänglich sein, und ebenfalls in einer dem Zweck entsprechenden und modifizierbaren Form verfügbar sein. Das Werk muss einer Lizenz unterliegen, die eine Weiterverbreitung unter den Lizenzbedingungen des ursprünglichen Werks erlaubt, ebenso wie Modifikationen und Derivate. Die Weiterverwendung und -verteilung darf hierbei nicht durch technische

Hindernisse eingeschränkt sein, ebenso darf auch keine Diskriminierung von Einzelpersonen oder Gruppen stattfinden. Eine Namensnennung der Urheber und Mitbeteiligten kann bei der Weiterverwendung des Werkes Bedingung der Lizenz sein. Weiterhin muss die Lizenz des Werkes bei einer Weiterverteilung übergehen und sie darf nicht die Verbreitung anderer Werke behindern. Außerdem darf die Lizenz keine Einschränkung der Einsatzzwecke vorsehen. Eine Nutzung des Werkes für kommerzielle Zwecke darf somit auch nicht ausgeschlossen werden.

Die Arten von offenen Daten sind vielfältig und umfassen beispielsweise Geodaten, Daten über kulturelle Werke und Artefakte, Statistiken, Umwelt- und Verkehrsdaten, wissenschaftliche Veröffentlichungen, Forschungsergebnisse aus dem Bereich der Medizin oder Hörfunk- und Fernsehsendungen. Unter Open Data fallen somit Daten sowohl aus dem Bereich des öffentlichen Sektors, als auch von privaten Unternehmen, Nichtregierungsorganisationen, Non-Profit-Organisationen oder Bildungseinrichtungen (Open Knowledge Foundation Deutschland (2014); von Lucke (2011, S. 5)).

2.2.3. Open Government und Open Government Data

Open Government steht grundsätzlich für die Öffnung von Staat und Verwaltung gegenüber der Bürger. Unmittelbar nach dem Amtseintritt von US-Präsident Barack Obama im Januar 2009 wurde das „Memorandum on Transparency and Open Government“ unterzeichnet, durch das der Begriff Open Government seine heutige Bedeutung erhalten hat, die sich auf eine moderne Kooperation zwischen Politik, öffentlichen Behörden, Industrie und Bürgern bezieht, die durch die Förderung von Transparenz, Mitbestimmung und Zusammenarbeit ermöglicht werden soll (Bauer und Kaltenböck, 2012, S. 9). In dem Memorandum von Obama (Obama, 2009) findet sich genauer wieder:

> „My Administration is committed to creating an unprecedented level of openness in Government. We will work together to ensure the public trust and establish a system of transparency, public participation, and collaboration. Openness will strengthen our democracy and promote efficiency and effectiveness in Government.“

Als eine der wichtigsten Maßnahmen zur Ermöglichung von Open Government sind dabei laut Bauer und Kaltenböck (2012, S. 10) der freie Zugang zu Informationen des Staates und der Verwaltung zu nennen, genauso wie auch die Möglichkeit einer freien

Weiterverwendung und -verbreitung dieser Informationen. In diesem Zusammenhang kommt der Begriff *Open Government Data* ins Spiel, der sich auf offene Daten bezieht, die konkret im Bereich des öffentlichen Sektors anfallen und für die Öffentlichkeit frei zur Verfügung und weiteren Nutzung gestellt werden. In von Lucke und Geiger (2010, S. 6) wird für den Begriff Open Government Data (hier auch als „offene Verwaltungsdaten“ bezeichnet) folgende Definition gegeben:

> „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.“

Die Datenbestände um die es hier im Genauen geht, sind also explizit dem öffentlichen Sektor zuzuordnen. Ihre Veröffentlichung muss im Interesse öffentlicher Belange liegen und sie dürfen ohne vorherige Einholung einer Genehmigung weder sensible Daten enthalten, die etwa Rückschluss auf einzelne Personen geben, noch Betriebs- oder Geschäftsgeheimnisse preisgeben. Einige Beispiele für Open Government Data sind etwa Statistiken, Geodaten, Bebauungspläne, Materialien der Parlamente, Ministerien und Behörden, Haushaltsdaten, Gesetze, Verordnungen, richterliche Beschlüsse und sonstige Veröffentlichungen (von Lucke und Geiger, 2010, S. 6).

Heutzutage wird für Open Government Data auch häufig der Begriff Open Data synonym verwendet (Dietrich, 2011b). In dieser Arbeit wird folgend der Begriff Open Data bevorzugt verwendet, da er als allgemeiner Oberbegriff offene Daten aus sämtlichen erdenklichen Bereichen einschließt.

2.2.4. Public Sector Information und Informationsfreiheit

Im Bereich des öffentlichen Sektors existieren zum Teil schon seit mehreren Jahren verschiedene weitere Maßnahmen zur Erhöhung der Transparenz und einfacheren Wiederverwendung von Daten, die von Staat und Behörden erzeugt werden. Open Government Data ist in diesem Zusammenhang noch ein relativ neues Konzept, das für die Erfüllung der obigen Zwecke herbeigerufen wurde. Basierend auf der Publikation von Barnickel und Klessmann (2012, S. 132) sollen in diesem Abschnitt die weiteren Aktivitäten übersichtsartig vorgestellt werden, um eine Abgrenzung der verschiedenen Begriffe und Konzepte möglich zu machen (siehe auch Tabelle 2.1).

Bereits vor dem Open Data Konzept wurden in vielen Ländern und auch auf europaweiter Ebene gesetzliche Regelungen verabschiedet, die die Weiterverwendung von Informationen des öffentlichen Sektors sowie die Veröffentlichung von Verwaltungsunterlagen betreffen; beides im Zusammenhang mit Anfragen zur Aushändigung von Verwaltungsdaten im Rahmen der Informationsfreiheitsgesetze.

Bei dem Informationsfreiheitsgesetz handelt es sich um ein Bürgerrecht, das sich auf den Ansatz der Informationsfreiheit stützt, der Bürgern die Möglichkeit geben soll, Einsicht in Informationen der öffentlichen Verwaltung nehmen zu können. Mit der Umsetzung dieses Anspruchs durch rechtliche Bestimmungen können Bürger Anfragen an öffentliche Einrichtungen zur Bereitstellung von bestimmten Informationen stellen (Kubicek, 2008, S. 9). Die Einrichtungen sind an dieser Stelle dann verpflichtet, die entsprechenden Dokumente bzw. Informationen bereitzustellen, es sei denn es handelt sich bei den angeforderten Informationen um Gegenstände, die unter verschiedene Ausnahmetatbestände fallen (Barnickel und Klessmann, 2012, S. 132). Der Zugang zu solchen Informationen wird heute bereits in über 95 Staaten durch entsprechende Informationsfreiheitsgesetze gewährleistet (Right2Info.org, 2013).

Durch die Weiterverwendung der Informationen des öffentlichen Sektors soll vor allem ein volkswirtschaftlicher Nutzen entstehen. Diese Informationen (in der Literatur auch geläufig unter dem Begriff *Public Sector Information*, kurz: PSI) besitzen verschiedenen Untersuchungen zufolge einen Wert, der nicht nur in der primären Nutzung zum Tragen kommt, sondern besonders auch in der Weiterverwendung durch Dritte Potenziale für die Wirtschaft verspricht (Dekkers u. a. (2006); Fornefeld u. a. (2009); Pollock (2009)). Die Untersuchungen sprechen hier unter anderem von geschätzten Auswirkungen die Beträge im bis zu zweistelligen Milliardenbereich herbeiführen.

	Open Government Data	Public Sector Information	Informations-freiheit
Ziele (in absteigender Bedeutung)	Transparenz, zivilgesell-schaftliches Engagement, Wirtschafts-förderung, Effizienz	Wirtschafts-förderung, Transparenz	Transparenz
Bereitstellung	Proaktiv	Proaktiv	Auf Nachfrage
Kosten für die Nutzer	Nach Möglichkeit ohne Kosten	Marginalkosten – Profitmaximie-rung	Üblicherweise Marginalkosten
Umfang der Datensätze	Große Mengen	Große Mengen	Eher kleine bis mittlere Mengen
Digital/analog	Digital	Digital	beides
Weiterverwendung durch Dritte	Ja	Ja	Derzeit häufig nicht vorgesehen
Art der Daten	Überwiegend quantitativ	Überwiegend quantitativ	Überwiegend qualitativ

Tabelle 2.1.: Abgrenzung der Begriffe Open Government Data, Public Sector Information, Informationsfreiheit anhand verschiedener Kategorien (nach Barnickel und Klessmann (2012, S. 133)).

2.3. Potenziale von Open Data

Open Data bietet sowohl für Unternehmen und Wirtschaft, als auch für den öffentlichen Sektor selbst, enorme Potenziale. Die Informationen, die etwa konkret in den öffentlichen Einrichtungen als Open Data bereitgestellt werden, können Weiterverwendung in neuen Produkten und Dienstleistungen finden, und parallel auch für Effizienzsteigerungen in den Verwaltungen sorgen (Europäische Kommission, 2011, S. 2). Durch die Bereitstellung von Verwaltungsdaten als Open Data wird außerdem die Transparenz erhöht, da den Bürgern ein Einblick in bestimmte Vorgänge der Behörden und Gebietskörperschaften ermöglicht wird. Liegen diese Daten in einem leicht weiterzuverarbeitenden Format vor, vereinfacht dies nach Barnickel und Klessmann (2012, S. 134) „die Aufbereitung der Daten für unterschiedliche Fragestellungen, die Organisationen oder interessierte Bürger haben“. Barnickel und Klessmann erwähnen als zusätzlichen

Effekt auch eine verstärkte Beteiligung der Bürger an politisch-administrativen Abläufen (Barnickel und Klessmann, 2012, S. 133-134).

Was die Potenziale im Bereich der Wirtschaft betrifft, ist durch die Bereitstellung von Open Data durch öffentliche Einrichtungen vor allem ein großer Wachstumsschub zu erwarten. Einer europäischen Studie aus dem Jahr 2008 zufolge, wurde die Summe der direkten und indirekten wirtschaftlichen Auswirkungen, die sich aus Public-Sector-Informations-Anwendungen und deren Nutzung in den 27 EU-Mitgliedstaaten ergibt, auf jährlich 140 Milliarden Euro geschätzt (Europäische Kommission, 2011, S. 3). Diese positiven Auswirkungen sind schließlich darauf zurückzuführen, dass Open Data in einer Vielzahl von Geschäftsmodellen Weiterverwendung finden kann, und somit die Entstehung eines breiten Spektrums an neuen Informationsprodukten und -dienstleistungen begünstigt bzw. fördert. Die strukturierte Veröffentlichung von Verwaltungsdaten kann zuletzt auch die Arbeitsabläufe in den Einrichtungen des öffentlichen Sektors positiv beeinflussen. Mit Hilfe eines Portals, auf dem die Daten öffentlich zugänglich gemacht werden, können beispielsweise die verschiedenen Abteilungen einer Organisation ihre verfügbaren Daten besser verwalten und auffinden, und infolgedessen die Arbeitseffizienz erhöhen (Barnickel und Klessmann, 2012, S. 136).

2.4. Situation offener Verwaltungsdaten in Deutschland und im Ausland

Die Entwicklung und praktische Umsetzung des Konzeptes und der damit verbundenen Prinzipien von Open Government Data in Einrichtungen des öffentlichen Sektors sowie Gebietskörperschaften ist noch lange nicht in allen Ländern gleichermaßen vollzogen. Laut Barnickel und Klessmann (2012, S. 136-137) hat im Ausland eine Auseinandersetzung mit dem Ansatz teilweise wesentlich früher begonnen, und entsprechend lässt sich dies an der fortgeschritteneren Umsetzung feststellen. Hier spielen nicht nur der zeitliche Faktor oder verwaltungsorganisatorische Gründe eine Rolle; auch das kulturhistorische und gesellschaftliche Verständnis von Offenheit und Transparenz und die damit in Verbindung stehenden Einflüsse auf die verfassungsrechtlich definierten Grundprinzipien der Staaten können das zügigere Aufgreifen des Open Government Data Ansatzes unterstützt haben (Barnickel und Klessmann (2012, S. 136-137); Blumauer u. a. (2011, S. 18-19)). Der Literatur nach wird Transparenz in einigen Ländern

schon lange großgeschrieben, und entsprechend steht hier der digitalen Öffnung von Verwaltungsdaten wenig im Wege. Die Ausgestaltung des staatlichen Aufbaus wird hier ebenfalls als einflussreicher Faktor beschrieben. Stärker zentralistisch orientierte Länder wie zum Beispiel Großbritannien können eine Implementierung von neuen transformativen Konzepten im öffentlichen Bereich schneller vornehmen als Länder mit föderalistischer Ausprägung (Barnickel und Klessmann, 2012, S. 136-137).

Die folgenden Abschnitte sollen anhand von Praxisbeispielen einen groben Überblick über den aktuellen Stand in Deutschland in Bezug auf die Bereitstellung von Open Data geben, und ebenfalls entsprechende Beispiele von Ansätzen auf internationaler Ebene aufzeigen. Dabei liegt der Fokus auf Aktivitäten, welche bewusst dem Konzept von Open Government Data folgen bzw. diesem zugeordnet werden können.

2.4.1. Situation in Deutschland

Die Bereitstellung von offenen Verwaltungsdaten wird in Deutschland heute sowohl auf landesweiter Ebene als auch auf regionaler bzw. kommunaler Ebene verfolgt. Eine umfangreiche Studie zu Open Government Data in Deutschland, die 2012 im Auftrag des Bundesministerium des Innern vom Fraunhofer-Institut für Offene Kommunikationssysteme (FOKUS) durchgeführt wurde (Klessmann u. a., 2012), bestätigt das Interesse an der Öffnung von Daten des öffentlichen Sektors in Deutschland.

Auf Bundesland- bzw. landesweiter Ebene stellen bereits verschiedene Bundesbehörden ihre Daten bereit. Das Bundesamt für Statistik[1] stellt zum Beispiel eine Vielzahl von amtlichen statistischen Informationen über Deutschland zur Verfügung (Statistisches Bundesamt, 2015). Im Umweltbereich bietet des Weiteren das Portal NUMIS[2] des Niedersächsischen Ministeriums für Umwelt, Energie und Klimaschutz Zugang zu Katalogen, Fachdatenbanken, Dokumenten, Metadaten und digitalen Karten zum Thema Umwelt (Niedersächsisches Ministerium für Umwelt, Energie und Klimaschutz, 2015). Die Liste weiterer bundesweiter Praxisbeispiele ist lang, und wird dem Überblick halber nicht weiter fortgeführt.

Auf lokaler Ebene sind in Deutschland vor allem in Großstädten entsprechende Aktivitäten zur Bereitstellung von Open Government Data zu beobachten. In der Freien

[1] http://www.destatis.de/
[2] http://numis.niedersachsen.de/

Hansestadt Bremen wurden zum Beispiel im eigenen Informationsfreiheitsgesetz proaktive Veröffentlichungspflichten für bestimmte Dokumente der öffentlichen Einrichtungen festgelegt (daten.bremen.de, 2015). Auf dem Portal daten.bremen.de werden diese Dokumente schließlich als reine Rohdaten zum Download bereitgestellt. In der Freien und Hansestadt Hamburg wurde in Bezug auf die Veröffentlichung von Verwaltungsdaten ebenfalls der Rechtsrahmen erweitert: Zum 6. Oktober 2012 wurde das bis dahin geltende Informationsfreiheitsgesetz durch das Hamburgische Transparenzgesetz abgelöst, durch das die Behörden der Stadt zur unverzüglichen Veröffentlichung einer Vielzahl von Dokumenten und Daten verpflichtet wurden (Senat der Freien und Hansestadt Hamburg (2012); Behörde für Justiz und Gleichstellung der Freien und Hansestadt Hamburg (2012)). Unter den Begriff Behörde fällt dabei nach dem Gesetz nicht nur die gesamte Verwaltung der Freien und Hansestadt Hamburg, sondern es werden auch viele sogenannte öffentliche Unternehmen erfasst, die öffentliche Aufgaben wahrnehmen und der Kontrolle der Stadt unterliegen. Die veröffentlichungspflichtigen Informationen werden seit 2014 als Rohdaten in dem sogenannten Transparenzportal[3] veröffentlicht, das zum Zeitpunkt der Verfassung dieser Arbeit etwa 24.000 Datensätze aufführt (Stand: Februar 2015). Neben dem Zugriff auf den Katalog über das Web Interface, kann auch über eine Programmierschnittstelle (API) maschinell auf die gespeicherten Datensätze zugegriffen werden. Abschließend betrachtet, unterscheiden sich die rechtlichen Rahmenbedingungen in der Freien Hansestadt Bremen und der Freien und Hansestadt Hamburg von vergleichbaren Normen in anderen Bundesländern und auf Bundesebene, weswegen die Städte damit eine Vorreiterrolle bei der proaktiven Veröffentlichung von Open Government Data in Deutschland einnehmen.

2.4.2. Situation auf internationaler Ebene

Auch im Ausland ist bereits seit ein paar Jahren eine steigende Tendenz zur Öffnung von Daten des öffentlichen Sektors zu beobachten. Zum Zeitpunkt der Verfassung dieser Arbeit (Stand: Februar 2015) führt das CTIC (Center for the Development of Information and Communication Technologies in Asturias, Spanien) weltweit mehr als 280 Open Government Data Projekte auf (CTIC, 2015). Hierbei gelten vor allem die USA mit dem Portal data.gov und Großbritannien mit dem Portal data.gov.uk als Vorreiter bei der praktischen Umsetzung von Open Government Data. Neben diesen wurden

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

u.a. auch in Japan[4], Kanada[5], Australien[6] und Neuseeland[7] entsprechende Projekte, meist direkt von den Regierungen oder Kommunen selbst, gestartet (Forsterleitner und Gegenhuber, 2011, S. 238).

Das Open Government Data Portal data.gov.uk, das 2009 als eines der ersten als geschlossenes Beta-Projekt von der britischen Regierung gestartet wurde (data.gov.uk, 2014), hebt sich dabei von den bisherigen genannten Portalen bedeutend ab. Die aktive politische Einbringung des früheren Premierministers Gordon Brown sowie die Unterstützung von Sir Tim Berners-Lee, dem Erfinder des World Wide Web, haben dabei besonders geholfen, die Entwicklung des Datenportals voranzutreiben (Blumauer u. a., 2011, S. 17). Das Portal umfasst aktuell (Stand: Februar 2015) etwa 23.500 Datensätze, die von den Ministerien Großbritanniens bereitgestellt werden, und bietet zudem eine Programmierschnittstelle (API) für den maschinellen Zugriff auf die Daten an. Darüber hinaus wird für die IT-gestützte Verarbeitung auch ein Teil der Datensätze mit semantischen Annotationen als sogenannte Linked Data veröffentlicht. Zum Zeitpunkt der Verfassung dieser Arbeit listet der Katalog etwa 260 Einträge zu Datensätzen auf, die als Linked Data veröffentlicht wurden. Weiterhin ermöglicht data.gov.uk über ein integriertes Formular einen Online-Antrag auf Bereitstellung eines gewünschten Datensatzes zu stellen. Auf diese Weise haben Bürger die Möglichkeit, fehlende Datensätze einer öffentlichen Behörde anzufordern, ohne diese vor Ort aufsuchen und um Herausgabe der Daten bitten zu müssen. Auf einer gesonderten Übersichtsseite im Portal lassen sich außerdem bereits gestellte Anträge öffentlich einsehen und anhand ihres Status verfolgen.

2.5. Herausforderungen und Fazit

Das Konzept Open Data bzw. Open Government Data wird zum jetzigen Zeitpunkt bereits von vielen Regierungen und Kommunen der Welt verfolgt, doch mit Blick speziell auf Deutschland, scheint dieses noch nicht vollständig ausgereift bzw. geeignet umgesetzt zu sein. Im Bereich des öffentlichen Sektors weist die praktische Umsetzung

[4] http://www.data.go.jp/
[5] http://open.canada.ca/en
[6] http://data.gov.au/
[7] https://data.govt.nz/

des Konzeptes Schwachstellen hinsichtlich der rechtlichen und technischen Rahmenbedingungen auf, mit der Folge, dass die Nutzung der Daten erschwert wird.

Die Heterogenität der Zuständigkeiten für die Erfassung, Aufbereitung und Bereitstellung von Daten, die speziell im öffentlichen Sektor anfallen, wirkt sich in Deutschland auf die Datenbereitstellung aus (Dietrich, 2011a). Dies betrifft demnach auch andere Länder, die wie Deutschland eine föderale Struktur besitzen. Die vorhandenen Formate der bereitgestellten Daten sind nur eingeschränkt maschinell lesbar und auswertbar, und eine Einbindung in bestehende Web-basierte und mobile Anwendungen wird ebenfalls erschwert. Dies liegt vor allem daran, dass die Daten durch unterschiedliche Akteure meist in nicht-standardisierten bzw. -proprietären Formaten, wie etwa PDF- oder XLS-Dateien, zur Verfügung gestellt werden. Diese Schwierigkeiten hebt auch Daniel Dietrich, offizieller Repräsentant der Open Knowledge Foundation in Deutschland, hervor (Dietrich, 2011a). Er weist auch auf die unterschiedlichen und teils inkompatiblen Lizenzen hin, die auch heute noch ein Problem für die Weiterverwendung der Daten darstellen. Dietrich erwähnt außerdem als eine weitere Herausforderung die „Heterogenität der verwendeten Vokabulare und Klassifikationen zur semantischen Beschreibung der Daten“, die sich speziell bei der Aggregation und Auswertung von Open Data auf übergreifender Ebene stellt. In dieser Hinsicht ist zu beobachten, dass die Bereitsteller auch heute nach wie vor unterschiedliche Vokabulare und Klassifikationen einsetzen, und die praktische Durchsetzung eines global einheitlichen Standards aussteht.

Die genannten ausstehenden Herausforderungen wirken sich letztendlich kontraproduktiv auf die Nutzung und Weiterverwendung von Open Data aus, sodass dessen Potenziale noch nicht vollständig genutzt werden. Die eigentliche Herausforderung für die Bereitsteller von Open Data liegt somit in der Aufgabe, einen dezentralen, föderierten Ansatz zu entwickeln, der nach Dietrich (2011a) auf offene Standards, Formate und Lizenzen setzt, und „eine übergreifende Bereitstellung und Weiterverwendung von öffentlichen Daten in Deutschland ermöglicht“.

Durch die aktuellen rechtlichen Rahmenbedingungen, die in Deutschland aber auch im Ausland gegeben sind, werden noch lange nicht von allen Behörden sämtliche verfügbare Informationen als Open Data freigegeben, da die Gesetzeslage in den meisten Ländern noch keine pro-aktive Veröffentlichung von Behördendaten vorschreibt. In dieser Hinsicht bleibt somit noch viel zu tun. So lange sich die rechtlichen Rahmenbedingungen nicht entsprechend geändert haben, müssen interessierte Bürger vom

Informationsfreiheitsgesetz Gebrauch machen, und auf klassische, bürokratische Art die gewünschten Daten bei den öffentlichen Einrichtungen anfragen.

3. IT-Maßnahmen zur Bereitstellung von Datenbeständen

3.1. Einleitung

Aus der Definition des Begriffes Open Data, die von der Open Knowledge Foundation geprägt und bereits in Kapitel 2 erläutert wurde, geht eine besonders zentrale Aufgabe für die Produzenten von Open Data hervor: Sie müssen ihre Daten frei zugänglich bereitstellen, sodass diese ohne Einschränkungen wiederverwendet, nachgenutzt und verbreitet werden können. Dieses Kapitel widmet sich in diesem Zusammenhang der technischen Infrastruktur, die benötigt wird, um Open Data bereitzustellen und wiederverwendbar zu machen, sodass externe Akteure darauf zugreifen und Verwendung davon machen können.

Dazu wird zunächst in Abschnitt 3.2 der Aspekt *Big Data* behandelt, der sich grundlegend mit der Thematik auseinandersetzt, wie eine stetig wachsende Menge von Daten, sowohl heute als auch in Zukunft, von der Informationstechnologie bewältigt werden soll. Welche Dimensionen werden im Rahmen von Big Data betrachtet, und wie können Daten effizient persistiert, verwaltet und weiterverarbeitet werden? Welche Technologien spielen in diesem Zusammenhang eine Schlüsselrolle?

Im Kontext der Bereitstellung von Open Data ist neben der effizienten Verwaltung der Daten aber auch vor allem die Art des Zugangs zu diesen von Interesse. Datenjournalisten, Anwendungsentwickler und andere Interessierte benötigen passende Schnittstellen, über die sie gezielt nach veröffentlichten Datensätzen suchen können, die sie dann im weiteren Schritt abrufen und weiterverarbeiten können. Abschnitt 3.3 beschreibt hierzu die Anforderungen, die eine geeignete Plattform für Open Data erfüllen muss, und geht im Weiteren auf bereits vorhandene Softwarelösungen ein, die auf die Bereitstellung von Open Data ausgelegt sind.

Abschnitt 3.4 befasst sich weitergehend mit technischen Maßnahmen, die eine Wiederverwendung und Nachnutzung von Daten nachhaltig fördern sollen. Hierzu hat sich das *Linked Data* Konzept bewährt, das eine Vernetzung der im Web veröffentlichten Datenbestände vorsieht, um Zusammenhänge abzubilden, die der Beantwortung von konkreten Fragestellungen dienen sollen.

Abschnitt 3.5 beinhaltet abschließend eine kurze Zusammenfassung zu den obigen Teilaspekten.

Die nachfolgenden Abschnitte sollen dabei jeweils nur einen Überblick zu den Teilaspekten liefern. Für eine Vertiefung in die zugehörigen Themengebiete empfiehlt sich die Auseinandersetzung mit der entsprechenden Literatur, die für diese Arbeit herangezogen wurde.

3.2. Umgang mit großen Datensätzen

Der Bereich der Informationstechnologie verzeichnet ein massives Aufkommen von Daten, das durch den Einsatz neuester technologischer Entwicklungen und deren Nutzung durch alle Ebenen der Bevölkerung angetrieben wird. Dieses Aufkommen umfasst Informationen die von einer Vielzahl von Quellen stammen, sowohl öffentlicher als auch privater (Dumbill, 2012, S. 3). Hierzu zählen Daten des öffentlichen Sektors, organisations- oder unternehmensspezifische Daten sowie auch Informationen von Privatpersonen, die beispielsweise durch die Nutzung von sozialen Netzwerken oder Smartphones erzeugt werden. Versendete Chat-Nachrichten, Status-Updates und „Check-In's“ auf Facebook, mittels GPS ermittelte Standort-Informationen von Smartphones, die zur Verkehrsstau-Anzeige in Kartendiensten dienen, Informationen, die Webbrowser auf dem Server einer Internetseite hinterlassen, Haushaltspläne von Regierungen, Wetterberichte oder Wirtschaftsindikatoren sind nur ein winziger Bruchteil von Beispielen, wo Daten im Spiel sind. Einer Studie der International Data Corporation (IDC) zufolge, wurde das in 2012 weltweit erzeugte Volumen an Daten auf 2,7 Zettabytes (entspricht 2,7 Milliarden Terabyte) geschätzt. Es wurde außerdem prognostiziert, dass sich dieses etwa alle zwei Jahre verdoppelt und somit bis 2015 insgesamt 8 Zettabyte an Daten erzeugt wurden (Gens, 2011).

Unter dieser massiven Zunahme an Daten, die u.a. auch in der Open Data Landschaft erzeugt werden, wird das populäre IT-Buzzword *Big Data* hauptsächlich verwendet,

um große Datensätze zu beschreiben, die verglichen mit herkömmlichen Datensätzen, typischerweise Massen an unstrukturierten Daten enthalten, die von traditionellen Datenbanksystemen kaum oder nur ineffizient verarbeitet werden können. Diese Datensätze besitzen meist ein finanzielles Potenzial, das darauf wartet, entdeckt zu werden (Klein u. a., 2013, S. 319). Um dieses Potenzial ausschöpfen zu können, muss jedoch zunächst eine technische Infrastruktur gegeben sein, die eine effiziente Speicherung, Verwaltung und Verarbeitung der Massen an Daten sicherstellt. Die folgenden Abschnitte befassen sich mit diesem Aspekt näher, und führen abschließend die Rolle und Herausforderungen von Big Data zusammen.

3.2.1. Big Data

Der Begriff *Big Data* stellt ein abstraktes Konzept dar, dessen Definition aufgrund unterschiedlicher Belange fachübergreifend nicht eindeutig festgelegt ist (Chen u. a., 2014). Dumbill (2012, S. 9) beschreibt Big Data als Daten, die aufgrund ihrer Menge, Schnelllebigkeit oder inkompatiblen Struktur die Verarbeitungskapazitäten von konventionellen Datenbanksystemen übertreffen. Es besteht demnach der Bedarf an modernen Technologien, die mit solchen Daten effektiv umgehen können. Aus der Definition von Mills u. a. (2012, S. 10) wird dieses Verlangen ebenfalls ersichtlich:

> „Big Data is a term that describes large volumes of high velocity, complex and variable data that require advanced techniques and technologies to enable the capture, storage, distribution, management, and analysis of the information.“

Über die Eigenschaften von Big Data ist sich die Fachliteratur wiederum weitgehend einig. Diese bezieht sich auf das 3V-Modell von Gartner (2011), das wiederum auf einen Forschungsbericht von Laney (2001) zurück geht, der den Zuwachs an Daten mit drei Dimensionen beschreibt: Umfang (engl. *volume*), Varietät (engl. *variety*) und Geschwindigkeit (engl. *velocity*). In der Fachliteratur wird des Weiteren noch die Eigenschaft Zuverlässigkeit (engl. *veracity*) als vierte Dimension erwähnt (Mills u. a. (2012); Klein u. a. (2013)), auf die zusammen mit den bereits genannten Dimensionen im nachfolgenden Abschnitt näher eingegangen wird.

3.2.2. Die Dimensionen von Big Data

Folgend werden die vier Dimensionen von Big Data *volume*, *variety*, *velocity* und *veracity* genauer erläutert. Die Beschreibungen basieren auf der Literatur von Mills u. a. (2012) und Klein u. a. (2013).

Die Dimension *volume* bezeichnet das ansteigende Volumen an Daten, das aufgenommen und verwaltet werden muss. Abhängig von der Anzahl der Quellen und der gewählten Datentiefe bzw. Auflösung, die Bereitsteller von Daten (und somit auch von Open Data und speziell Open Government Data) einsetzen, erhöht sich auch entsprechend der Datenumfang. Der erhöhte Umfang an Daten stellt jedoch für traditionelle Datenbanksysteme eine Herausforderung dar. Es existieren zwar Datenbanksysteme, die Datenbestände im Petabyte Bereich aufnehmen können, doch hier muss abgewägt werden, welchen Wert die Daten haben und ob sie die hohen Kosten für diese Datenbanksysteme ausgleichen.

Variety bezieht sich auf die Vielzahl und Vielfalt der Daten, die wegen ihrer unterschiedlichen Art und oftmals fehlenden Struktur von traditionellen relationalen Datenbanksystemen nicht effizient verarbeitet werden können. Im Open Government Data Bereich zum Beispiel, werden Datensätze aus den verschiedensten Quellen bzw. Bereichen des öffentlichen Sektors bezogen und veröffentlicht. Hierbei handelt es sich meist um eine Mischung aus unstrukturierten, semistrukturierten und strukturierten Daten, die sich in zusammengefasster Form nur schlecht analysieren lassen.

Der Aspekt *velocity* beschreibt einerseits die Rate, mit der Daten in den verschiedensten Anwendungsbereichen produziert werden. Er beschreibt andererseits aber auch die Geschwindigkeit, in der die Daten empfangen und weiterverarbeitet werden. So gibt es beispielsweise Anwendungsfelder, in denen die Daten möglichst zeitnah bzw. in Echtzeit verarbeitet werden müssen (als bekanntestes Beispiel hierfür gelten Suchmaschinen wie Google, Yahoo etc.).

Die Dimension *veracity* bezieht sich zuletzt auf die Zuverlässigkeit bzw. Qualität der erhaltenen Daten, die häufig aufgrund ihrer unterschiedlichen Herkunft unvollständig, inkonsistent oder mehrdeutig sein können. Auch die kurzen Verarbeitungszeiten von Daten, die in bestimmten Anwendungen gefordert werden, können zu einer Ungenauigkeit in den gesammelten Daten führen.

3.2.3. Technologien

Im Bereich von Big Data gibt es bereits moderne Technologien, die mit der Erfassung, Speicherung, Verteilung, Verwaltung und Analyse von Daten jeglicher Art und Anzahl zurecht kommen. Vermehrt werden hier neue Konzepte eingesetzt, die im Gegensatz zu traditionellen Technologien (wie z.B. relationalen Datenbanksystemen) einen effizienteren Umgang mit unstrukturierten und großen Datenbeständen bieten. Diese Konzepte haben sich in dem Bereich von industriellen Big Data inzwischen als sehr brauchbar erwiesen, und werden etwa von Unternehmen wie Facebook oder Twitter erfolgreich eingesetzt (Borthakur u. a. (2011); Twitter, Inc. (2012)). Zu den bekanntesten Beispielen zählen NoSQL Datenbanken und das Apache Hadoop Framework, die in den folgenden Abschnitten übersichtsartig vorgestellt werden.

3.2.3.1. NoSQL Datenbanken

Not Only SQL (NoSQL) Datenbanksysteme verfolgen einen nicht-relationalen Ansatz und sind für Einsatzbereiche konzipiert, bei denen große Mengen von unstrukturierten Daten verwaltet und verarbeitet werden müssen. Sie sind für Einsatzbereiche konzipiert, bei denen laut Klein u. a. (2013, S. 321-322) „relationale Datenbanksysteme an ihre Grenzen stoßen“. Klein u.a. betonen weiterhin, dass NoSQL Datenbanksysteme in verteilten Systemen Verwendung finden, bei denen insbesondere vorhandene Werte nicht oft modifiziert werden, jedoch stets neue hinzugefügt werden. Ein Beispiel für solche Werte können etwa Chat-Nachrichten sein, die Nutzer in sozialen Netzwerken untereinander versenden.

NoSQL Datenbanksysteme teilen sich, basierend auf den Ausführungen von Klein u. a. (2013, S. 321-322), in drei Arten auf: Dokumenten-orientierte Datenbanken, Graphen-Datenbanken und Key-Value-Datenbanken. Dokumenten-orientierte Datenbanken eignen sich für die Persistenz von semistrukturierten Daten und erlauben ein Durchsuchen der Dokumentinhalte. Bekannte Beispiele hierfür sind die Systeme MongoDB und Apache CouchDB. Graphen-Datenbanken sind Klein u.a. zufolge „für das Speichern von Beziehungen zwischen verschiedenen Entitäten optimiert“ (Beispiele hierfür sind die Datenbanksysteme Neo4j und ArangoDB), während zuletzt Key-Value-Datenbanken die Speicherung von beliebigen Werten mit Hilfe von Schlüsseln realisieren. Key-Value-Datenbanken teilen sich hierbei in zwei Varianten auf: Bei der In-Memory Variante

liegen die Daten vollständig im Arbeitsspeicher, was in der Praxis zu einer hohen Leistung führt. Bei der On-Disk Variante hingegen werden die Daten auf einer Festplatte gespeichert, sodass hier gut die Rolle als Datenspeicher zum Tragen kommt. Beispiele für Key-Value-Datenbanken sind die Systeme Apache Cassandra und BigTable.

3.2.3.2. Apache Hadoop Framework

Der folgende Abschnitt basiert auf dem veröffentlichten Dokument des Intel IT Center (2013), welches Apache Hadoop als ein Open Source Framework beschreibt, das zur verteilten Verarbeitung von großen Datensätzen auf Computerclustern eingesetzt wird. Es basiert auf der Programmiersprache Java (Klein u. a., 2013, S. 322) und beinhaltet ein verteiltes Dateisystem, ein Framework zur parallelen Verarbeitung (genannt Apache Hadoop MapReduce, das weiter unten erläutert wird), sowie verschiedene weitere Komponenten, die die Erfassung von Daten, Koordinierung von Workflows und Aufgaben sowie die Überwachung der Cluster unterstützen.

Im Vergleich zu traditionellen Ansätzen, wie zum Beispiel relationalen Datenbanksystemen, ist Hadoop kosteneffizienter im Umgang mit großen unstrukturierten Datensätzen. Hadoop stellt jedoch anders als NoSQL kein Datenbanksystem dar, sondern ist ein hochskalierbares Speicher- und Datenverarbeitungssystem, das als Ergänzung für existierende Systeme dient, die bedingt durch ihre eingeschränkten Datenverarbeitungskapazitäten an ihre Grenzen stoßen. Hadoop ist in der Lage, beliebige Daten von einer Vielzahl an Quellen gleichzeitig zu erfassen und zu speichern, und kann sie weiterhin für beliebige Zwecke aggregieren, weiterverarbeiten und ausliefern.

MapReduce stellt hierbei das Programmiermodell und -framework zur Verfügung, das Software-Entwicklern die Möglichkeit zur Bestimmung und Einfädelung von komplexen Berechnungen auf Computerclustern bietet. Das eingesetzte Verfahren arbeitet dabei laut Klein u. a. (2013, S. 322) nach dem *Divide and Conquer* („Teile und Herrsche“) Paradigma: In dem Map-Task werden die Datensätze in voneinander unabhängige Einheiten bzw. Fragmente unterteilt, die dann parallel verarbeitet werden. Die Ausgaben aus dem Map-Task werden anschließend sortiert und an den Reduce-Task gereicht, der aus den einzelnen Zwischenergebnissen das Gesamtergebnis berechnet. Die Eingabe- und Ausgabedaten werden dabei im Apache Hadoop Distributed File System (HDFS) oder in anderen Speichersystemen gespeichert. Da die Daten typischerweise auf dem

selben Knoten verarbeitet und gespeichert werden, können die Aufgaben effektiver durchgeführt werden.

3.2.4. Schlussfolgerung

Der Begriff Big Data steht nicht nur für eine weltweit immer größere Häufung von Daten, sondern es geht genau genommen um das Potenzial, das in diesem unüberschaubaren Datenberg steckt und mit Hilfe der Informationstechnologie herausgeholt werden muss. Für die Informationstechnologie geht es hier letztlich nicht nur um Fragen bezüglich der Skalierbarkeit von Persistenz und Rechnerleistung, sondern im engeren Sinne auch um Herausforderungen bei der algorithmischen Deutung sowie semantischen Aggregation von großen Datenbeständen. So sind Big Data genau dann gut, wenn die Verfahren, die darauf ablaufen sollen, dieses Gut auf algorithmischer Ebene tatsächlich als Grundlage nehmen können. Dazu ist aus technischer Sicht die Wahl von geeigneten Technologien erforderlich, die mit diesen Daten effizient umgehen können. Zu diesem Zweck wurden in diesem Unterkapitel zwei der bekanntesten Technologien vorgestellt, die typischerweise Einsatz in Big Data-Anwendungen finden. An dieser Stelle ist jedoch hervorzuheben, dass es keine technische Universallösung für sämtliche Anwendungsbereiche gibt. Es gibt stattdessen nur relativ zu den verschiedenen Aufgabenstellungen geeignete Lösungen. Die vorgestellten Beispiele von Big Data-Technologien stellen somit keineswegs einen allgemeingültigen Ersatz für konventionelle Systeme mit relationalen Datenbanken dar, sondern sie sind nur dann und dort einzusetzen, wo sie wirksam einen Mehrwert bieten.

Abschließend betrachtet, können Big Data-Technologien allerdings als eine gute Möglichkeit angesehen werden, mit der die immer größer werdenden Mengen von unstrukturierten Daten für Analysen und Weiterverarbeitungen zugänglich gemacht werden. Besonders im industriellen Bereich haben sich diese Lösungen schon zum gegenwärtigen Zeitpunkt als sehr zweckmäßig erwiesen.

3.3. Plattformen zur Bereitstellung von Open Data

Die Mengen an Daten die heute in den verschiedensten Anwendungsdomänen erzeugt und verarbeitet werden, erfordern ein IT-gestütztes System zur Erfassung, Speicherung,

Verwaltung, Verarbeitung, Verteilung und Bereitstellung dieser Daten, das für diese Zwecke üblicherweise eine Datenbank bzw. Persistenzschicht und entsprechende Software vereint.

Für die Bewältigung solcher Datenmengen existiert bereits eine Reihe verschiedener bewährter Datenmanagementsysteme, die jedoch alle meistens nur auf einen bestimmten Anwendungsbereich zugeschnitten sind (Hagen u. a., 2005, S. 663). So gibt es etwa *Document Management Systeme*, *Enterprise Content Management Systeme* oder als bekanntestes Beispiel *Content Management Systeme*, die sich genauer mit der Erfassung und Verwaltung von *Content* befassen (Götzer u. a., 2004, S. 97-98), welcher von Ostheimer und Janz (2005, S. 24) als die Summe der Einzelinformationen über die Struktur, den Inhalt und die Ausgabeform beschrieben wird. Content Management Systeme werden typischerweise von Redaktionen als Werkzeug zur Erstellung und Veröffentlichung von journalistischen oder anderen Inhalten eingesetzt, oder etwa von Bloggern zur Führung eines digitalen Tagebuchs bzw. Journals genutzt.

Traditionelle Datenmanagementsysteme, wie die oben genannten, werden dem Open Data Bereich allerdings nicht vollständig gerecht, da sich die Anforderungen an die Dienstleistungen dieser Systeme teils grundlegend unterscheiden. Wie bereits in der übergreifenden Einleitung zu diesem Kapitel (Abschnitt 3.1) erwähnt wurde, stehen im Kontext der Bereitstellung von Open Data besonders die Zugangsmöglichkeiten zu den Daten im Mittelpunkt. Dies geht vor allem aus der Open Data Definition sowie den verschiedenen Aktivitäten hervor, die sich mit der Zeit um den Open Data Bereich gebildet haben und in einer Vielzahl von Anwendungsbereichen zu beobachten sind. Welche Anforderungen sich hierbei konkret stellen, wird im nachfolgenden Abschnitt diskutiert.

3.3.1. Erfassung der allgemeinen Anforderungen

Die Dienstleistungen, die im Augenblick vermutet werden, dass sie im Open Data Bereich aus Bereitsteller- und Nutzersicht benötigt werden, lassen sich genauer von der Open Data Definition (siehe Kapitel 2) und den vielseitigen externen Aktivitäten, die um den Open Data Bereich herum zu beobachten sind, ableiten.

So wird zunächst aus Sicht der Open Data Bereitsteller ein System benötigt, das erstrangig die Erfassung, Speicherung, Verwaltung und Veröffentlichung der Daten gewährt.

Angebote die auf überregionaler oder nationaler Ebene Open Data bereitstellen, müssen außerdem in der Lage sein, die Daten der untergeordneten lokalen und regionalen Angebote automatisiert in ihr System mit aufnehmen zu können. Es wird also eine Datenplattform benötigt, die die zu veröffentlichenden Datensätze zentral innerhalb eines digitalen Kataloges verwaltet, und diese schließlich über geeignete und einfach zugängliche Schnittstellen an Nutzer, Drittanbieter-Applikationen und andere Open Data Portale zur Verfügung stellt. Webbasierte Plattformen stellen für diesen Zweck eine geeignete Lösung dar, da über das World Wide Web ein einfacher Datenaustausch erfolgen kann, und zugleich auch eine unkomplizierte Verbreitung der Daten an einen großen Teil der Bevölkerung[1] erreicht wird, der aus verschiedensten Personen- und Interessengruppen besteht.

Aus Nutzersicht betrachtet, ergeben sich ebenfalls bestimmte Anforderungen, die berücksichtigt werden müssen. Einerseits sollen beispielsweise einfache Bürger die Möglichkeit haben, mit möglichst wenig Suchaufwand eine gewünschte Information aus dem Bestand eines Open Data Portals in Erfahrung bringen zu können. Zusätzlich müssen aber auch Personen aus den unterschiedlichsten Berufsgruppen (wie etwa Anwendungsentwickler oder Datenjournalisten), sowie auch Unternehmen und Einrichtungen generell, Zugang zu einer technischen Schnittstelle haben, sodass maschinell auf die veröffentlichten Daten zugegriffen werden kann. Hierzu ist es wichtig, dass die veröffentlichten Datensätze mit Metadaten (beschreibenden Attributen) versehen sind, sodass sie im Katalog anhand dieser Attribute einfach aufgefunden werden können.

Sollen die Bedürfnisse der verschiedenen beschriebenen Nutzergruppen berücksichtigt werden, ergeben sich somit vielseitige Anforderungen an potenzielle Open Data Plattformen, die es in der Praxis umzusetzen gilt.

3.3.2. Vorhandene Softwarelösungen für Open Data Portale

Zum Zeitpunkt der Verfassung dieser Arbeit sind vorrangig zwei große Softwarelösungen bekannt, die auf die Bereitstellung von Open Data ausgelegt und auf die zuvor beschriebenen Anforderungen zugeschnitten sind. Die weltweit verbreitetste Lösung ist das Comprehensive Knowledge Archive Network (CKAN)[2], das als Open Source-

[1] Die Anzahl der Internet-Nutzer im Jahr 2014 wird weltweit auf etwa 2,92 Milliarden geschätzt (vgl. ITU (2014)).
[2] http://www.ckan.org

Projekt von der Open Knowledge Foundation ins Leben gerufen wurde, und aktuell ca. 115 Open Data Portale zählt, die auf diese Software setzen (Open Knowledge Foundation, 2014d). Als zweite Lösung ist das Socrata Open Data Portal[3] zu nennen, das auf kommerzieller Basis von dem gleichnamigen Unternehmen Socrata angeboten wird, und u.a. als Plattform für die Open Data Portale der Weltbank sowie der Städte New York und San Francisco dient (Socrata, 2014).

Da sich die Softwarelösung CKAN größerer Bekanntheit erfreut, vor allem aber kostenlos unter der Open Source-Lizenz vertrieben wird, und weltweit unumstritten als *der* Standard für Open Data Portale angesehen wird, wird im nachfolgenden Abschnitt nur auf die Lösung näher eingegangen.

3.3.3. Comprehensive Knowledge Archive Network

Das Comprehensive Knowledge Archive Network (CKAN) ist eine der weltweit führenden Lösungen für Datenplattformen. Es ist eine komplette Out-of-the-box Softwarelösung die von der Open Knowledge Foundation entwickelt wurde und der Open Source Lizenz unterliegt (Open Knowledge Foundation, 2014b). CKAN wird primär von Organen des öffentlichen Sektors (Verwaltungen, Regierungen etc.), Unternehmen und Organisationen eingesetzt, die den Open Data Ansatz verfolgen und ihre Daten zugänglich machen möchten. So zählen zu den bekanntesten Betreibern von CKAN die Open Data Portale der Vereinigten Staaten von Amerika (data.gov) und von Großbritannien (data.gov.uk), die als die Vorreiter der globalen Transparenzbewegung gelten.

Funktional betrachtet, ähnelt die Datenplattform CKAN einem Content Management System, das jedoch anstelle von Seiten und Blogeinträgen, Sammlungen von Daten verwaltet und öffentlich verfügbar macht. Datensätze die über CKAN veröffentlicht werden, können von Nutzern über die integrierten Suchmöglichkeiten durchsucht und gefunden werden, und schließlich in einer Vorschau als Karten, Graphen oder Tabellen betrachtet werden. Eine standardisierte technische Schnittstelle (engl. *Application Programming Interface*, kurz: API) bietet ebenfalls die Möglichkeit, sämtliche Funktionen die über das Web Interface möglich sind, auch maschinell auszuführen.

[3] http://www.socrata.com/products/open-data-portal/

3.3.3.1. Architektur und Features von CKAN

Im Folgenden werden der Aufbau und zugleich einige der nennenswertesten Features von CKAN beschrieben. Die Informationen dazu basieren im Wesentlichen auf der offiziellen CKAN Dokumentation[4]. Um einen besseren Lesefluss der einzelnen Beschreibungen zu den Teilbereichen bzw. -funktionen zu gewährleisten, werden die verschiedenen herangezogenen Dokumentationsabschnitte in der Regel zum Ende eines jeden Abschnittes aufgeführt.

Der Kern von CKAN besteht aus folgenden drei Komponenten:

1. Dem Datenkatalog, der für die Verwaltung der Datensätze zuständig ist.
2. Der webbasierten Benutzerschnittstelle, die Bereitstellern eine einfache Administration der Daten ermöglichen soll, sowie Daten-Interessierten einen einheitlichen Zugangspunkt zu dem Datenkatalog und dessen Filter- und Vorschaufunktionen bieten soll.
3. Der API, die alle Kernfunktionen von CKAN an externe Client-Anwendungen herausgibt und somit auch einen automatisierten, maschinellen Zugriff auf die veröffentlichten Datensätze erlaubt.

Die Architektur von CKAN zeichnet sich durch ihre hohe Modularität aus: Komponenten und Erweiterungen erlauben es, das System an die unterschiedlichen Bedürfnisse der Betreiber von Datenportalen anzupassen (Open Knowledge Foundation, 2014b) (siehe auch Abbildung 3.1). Die Funktionalität von CKAN lässt sich auf einfache Weise erweitern und setzt ferner kein Verständnis des Gesamtsystems voraus.

3.3.3.1.1. Verwaltung von Daten in CKAN

Die Daten werden für die Zwecke von CKAN in Einheiten veröffentlicht, die sich *Datensätze* nennen. Ein Datensatz ist ein Paket bestehend aus mehreren Daten, und kann zum Beispiel die Kriminalitätsstatistik für eine Region oder die Auflistung der Ausgaben einer öffentlichen Behörde sein. Wenn Nutzer in einem CKAN-Portal nach Daten suchen, erhalten sie in den Suchergebnissen individuelle Datensätze zurück.

[4] http://docs.ckan.org/en/latest/

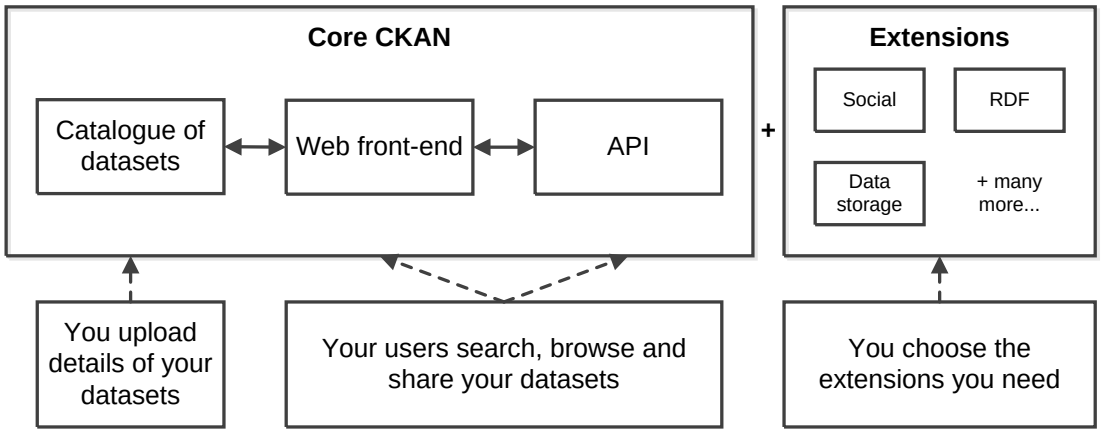

Abbildung 3.1.: Die Architektur von CKAN nach der Open Knowledge Foundation (2012a).

Die Datensätze liegen in CKAN in Form von Informationsobjekten vor, die mit Metadaten versehen sind und die rohen Daten als angehängte Ressourcen enthalten. Beispiele für Metadaten sind etwa der Titel und Name des Herausgebers des Datensatzes, das Veröffentlichungsdatum oder die unterliegende Lizenz. Ressourcen können wiederum beispielsweise Bilder, DOC- und PDF-Dokumente, CSV- und Excel-Tabellen, XML-Dateien etc. sein. CKAN kann die Ressourcen hierbei entweder intern speichern oder aber auch einfach nur als Link zu einer externen Quelle im Web speichern, die die Originalressource zur Verfügung stellt (Open Knowledge Foundation, 2014h). Um eine Verwaltung von Daten aus den unterschiedlichsten Domänen zu ermöglichen, ist das in CKAN vorgegebene Metadaten-Schema sehr einfach und generisch, doch es lässt sich beliebig erweitern um die Anforderungen der verschiedenen potenziellen Einsatzbereiche zu erfüllen (Open Knowledge Foundation, 2012b).

Die Speicherung der Ressourcen (bzw. eigentlichen Daten) erfolgt im FileStore und DataStore. Während der in CKAN bereits integrierte FileStore nur die Speicherung von ganzen Dateien anbietet, stellt die DataStore-Erweiterung eine *ad hoc* Datenbank für die Speicherung von strukturierten Daten aus CKAN Ressourcen zur Verfügung. Der DataStore ist hierbei eine Ergänzung zum FileStore, und ermöglicht analog zu einer Datenbank den Zugriff auf einzelne Datenelemente sowie des Weiteren auch die automatische Vorschau von strukturierten Daten innerhalb der Ressourcen-Seite (unter Verwendung der Data Explorer-Erweiterung). Ein Zugriff auf ein im FileStore gespeichertes Tabellendokument wäre zum Beispiel nur über einen Download der

gesamten Datei möglich. Liegt im Gegensatz das gleiche Dokument im DataStore vor, haben Nutzer die Möglichkeit, via der DataStore API sowohl auf einzelne Tabelleneinträge zuzugreifen, als auch Anfragen über den Inhalt des Tabellendokuments zu stellen. Die DataStore API erlaubt außerdem den Bereitstellern, gespeicherte Daten zu aktualisieren ohne diese erneut vollständig hochladen zu müssen. Dies kann insbesondere bei sehr umfangreichen Datensätzen von hohem Nutzen sein (Open Knowledge Foundation, 2014f).

3.3.3.1.2. Vorschau und Visualisierung von Daten

Daten die in CKAN gespeichert und verwaltet werden, können je nach Dateityp automatisch betrachtet und sogar teilweise interaktiv im Browser analysiert werden, ohne von vornherein heruntergeladen werden zu müssen. CKAN verwendet hierfür die Recline.js Bibliothek, die in eigener Hand von der Open Knowledge Foundation für Datenexplorations-Zwecke entwickelt wurde. Abhängig von der Art der Daten, können diese dann etwa direkt eingebettet oder in tabellarischer Form dargestellt werden, auf einer Zeitachse oder eingezeichnet auf einer Karte angezeigt werden. Mit Hilfe von CKAN Erweiterungen können zusätzlich benutzerdefinierte Vorschauen für andere Dateiarten erstellt werden, sodass sich diese Funktionalität von CKAN in beliebiger Weise ausbauen bzw. ergänzen lässt (Open Knowledge Foundation (2014e); Open Knowledge Foundation (2013)).

3.3.3.1.3. Import von Daten aus externen Datenportalen

CKAN eröffnet mit Hilfe der frei verfügbaren ckanext-harvest Extension die Möglichkeit, Datensätze aus mehrfachen externen CKAN-Instanzen automatisiert in ein einziges CKAN-Portal zu importieren. Der Import-Vorgang der in CKAN auch *Harvesting* (dt. „Ernten“, bzw. „Einsammeln“) genannt wird, greift dabei als Job in festen Intervallen auf die APIs der einzubindenden Liefersysteme, und fragt nach neuen verfügbaren Datensätzen, die importiert werden sollen. Pro neuen, zu importierenden Datensatz erstellt der Importer einen neuen Datensatz-Eintrag in CKAN, der auf die entsprechende Originalressource verlinkt. Eine wichtige Aufgabe des Importers ist hierbei, die korrekte Abbildung der fremden, einzubindenden Metadaten auf das eigene interne Metadatenmodell zu erstellen. Da zwischen den verschiedenen Liefersystemen oft

kein übergreifend einheitliches Metadatenmodell verwendet wird, muss für jedes dieser Systeme ein eigener Importer die spezifische Abbildung vornehmen.

Die Extension bietet zusätzlich ein Framework an, das zur Erstellung von benutzerdefinierten Importern dient, die Daten aus entfernten Portalen beziehen sollen, die nicht auf der Basis von CKAN betrieben werden (Open Knowledge Foundation, 2012c).

3.3.3.1.4. CKAN APIs

Als die vielleicht wichtigste Eigenschaft von CKAN ist die eigene RESTful API zu erwähnen, die auf JSON-Darstellungen basiert und einen Zugriff auf alle Funktionen des Kernsystems gewährt. Hierzu liefert CKAN ein umfangreiches Interface, dessen Funktionen über die API in größerem Umfang genutzt werden können, als über das grafische Web Interface, das den Benutzern zur Verfügung steht. Für das Hochladen und Modifizieren von Dateien im FileStore, sowie das Anlegen, Auslesen, Aktualisieren und Löschen von strukturierten Daten im DataStore, stehen ebenfalls entsprechende APIs zur Verfügung. Funktionen, die eine Authentifizierung erfordern, werden von der API ebenfalls unterstützt.

Durch die breite Bereitstellung von Programmierschnittstellen in CKAN, lässt sich das System beliebig erweitern oder in bestehende Systeme integrieren. Somit können zum Beispiel von außen Daten in den CKAN-Katalog eingespeist werden, oder Extensions erweiterte Features für die Vorschau und den Abruf von Datensätzen hinzufügen und in CKAN anbieten (Open Knowledge Foundation, 2014c).

3.3.3.2. Zugriff auf Datensätze

Der öffentliche Zugriff auf die in CKAN publizierten Datensätze kann nach der offiziellen CKAN Dokumentation der Open Knowledge Foundation (2014h) über zwei verschiedene Wege erfolgen:

Die erste Zugriffsmöglichkeit stellt das von CKAN bereitgestellte Web Interface dar, das den Benutzern auf einfache Weise ermöglicht, den Katalog gezielt nach Datensätzen zu durchsuchen, die anschließend heruntergeladen werden können. Mit Hilfe der eingebetteten Suchfunktion besteht die Möglichkeit, die Suchergebnisse anhand der vorliegenden Metadaten einzugrenzen und somit zu präzisieren. Der Vorteil hierbei ist,

dass das Web Interface vom Benutzer keine spezielle Kenntnis über die technische Schnittstelle von CKAN verlangt, und der Zugang über den Webbrowser erfolgt.

Als Alternative zum Web Interface kann, wie auch bereits in Abschnitt 3.3.3.1 diskutiert, über eine API maschinell auf den CKAN-Datenkatalog zugegriffen werden. Der Zugriff erfolgt hierbei über eine fest definierte REST-Schnittstelle, von der etwa Entwickler und Unternehmen Gebrauch machen können, um die verfügbaren Datensätze für ihre eigenen Anwendungen nutzen, und in einer beliebigen Weise weiterverarbeiten zu können. Mit der folgenden HTTP-Anfrage, die aus einem Beispiel der offiziellen CKAN API-Dokumentation stammt (Open Knowledge Foundation, 2014c), kann etwa die vollständige JSON-Repräsentation eines bestimmten Datensatzes aus CKAN (hier mit der ID *adur_district_spending*) abgerufen werden:

```
http://demo.ckan.org/api/3/action/package_show?id=adur_district_spending
```

Die API liefert auf die obige Anfrage ein entsprechendes JSON-Objekt zurück, das die Metadaten des Datensatzes und seiner zugehörigen Ressourcen (Rohdaten) enthält, und welches zu Beispielzwecken in Anhang A.1 (S. 76) dokumentiert wird. Bei der Nutzung der API müssen Entwickler jedoch beachten, dass jeder Datensatz und jede zugehörige Ressource (die ggf. unter einer externen Zieladresse gespeichert ist) zwar mit festen Katalog-internen Metadaten versehen ist, die Struktur der eigentlichen angehängten Rohdaten allerdings keinem standardisierten Modell folgt. Vor der eigentlichen Nutzung bzw. Auswertung müssen die Rohdaten somit auf ihre Struktur geprüft werden, da zum Beispiel verschiedene veröffentlichte Geodatensätze in Tabellen, nicht zwangsweise die gleichen Spaltenbezeichner, Einheiten, Wertebereiche, etc. verwenden.

3.3.3.3. Skalierbarkeit

Open Data Portale werden im Laufe der Zeit eine immer weiter wachsende Menge an Datensätzen aufnehmen und bewältigen müssen. Die in Kapitel 2 vorgestellten Beispiele von bestehenden Open Data Portalen verwalten zum Zeitpunkt der Verfassung dieser Arbeit Datensätze, die sich im 5- bis 6-stelligen Bereich bewegen. Wie kommen diese und andere CKAN-basierte Portale zukünftig mit Millionen von Datensätzen aus?

Von Seiten der Verwaltung von Millionen von Datensätzen liegen zum Zeitpunkt der Verfassung keine genauen Erfahrungswerte vor, da in der Praxis noch kein CKAN-Portal mit einem so großen Katalog bekannt ist bzw. existiert. Auch grenzenübergreifende Open Data Portale wie zum Beispiel das paneuropäische Portal publicdata.eu der Open Knowledge Foundation die auf Landesebene einen Single Point of Access zu allen Datensätzen der untergeordneten lokalen, regionalen und nationalen Portale anbieten (Open Knowledge Foundation, 2014a), sind von der Millionen-Marke noch weit entfernt. Nach aktuellem Stand setzt CKAN für die Verwaltung der Datensätze das relationale Datenbanksystem PostgreSQL ein (Open Knowledge Foundation, 2014b), das im Vergleich zu NoSQL-Datenbanksystemen (siehe Abschnitt 3.2.3.1) nicht optimal auf den Umgang mit großen unstrukturierten Datenbeständen ausgelegt ist. Die in CKAN eingebundene Enterprise Suchplattform Apache Solr (Open Knowledge Foundation, 2014b) stellt jedoch ausgleichend sicher, dass die Durchsuchung des Kataloges nach Datensätzen auch mit steigender Anzahl an Einträgen performant bleibt. Hierbei spielen die hohe Zuverlässigkeit, Skalierbarkeit und Fehlertoleranz von Apache Solr eine Rolle, die sich nach eigenen Angaben der Apache Software Foundation (2014) auf einigen der meist frequentiertesten Webseiten und Anwendungen der Welt erfolgreich unter Beweis gestellt haben.

Mit der Fragestellung wie weit CKAN mit dem Import einer Datenmenge in Millionenhöhe zurecht kommt, hat sich das Language Archive des Max Planck Institute for Psycholinguistics in Nijmegen, den Niederlanden im Rahmen des EUDAT[5] Projektes auseinandergesetzt. Dem hierzu veröffentlichten Bericht des Language Archive (The Language Archive, 2014) zufolge, erlauben das Design und die Standard-Konfiguration von CKAN primär den Import von nur einigen Tausenden von Datensätzen. Ein vom Language Archive durchgeführter Test-Import von 2 Millionen Datensätzen hätte ohne vorangegangene Performance-Optimierungen über ein Jahr gedauert. Erst durch Maßnahmen wie etwa der Veränderung der Tabellenstruktur der PostgreSQL-Datenbank, und Verzögerung der Indizierung der Solr-Suchplattform, gelang es dem Language Archive den besagten Import innerhalb von weniger als zwei Wochen durchzuführen.

[5] European Data Infrastructure, siehe auch EUDAT (2014).

3.3.4. Zwischenfazit

Datenplattformen sind für die Freigabe von Open Data eine wichtige Voraussetzung, da sie den Zugang zu Daten und deren Veröffentlichung vereinfachen, und weiterhin auch die Möglichkeit eröffnen, diese weiter zu verbreiten und zu verwenden. Welche Anforderungen sich hier, insbesondere aus Nutzersicht betrachtet, an diese Plattformen stellen, ist weitgehend geklärt. Bestehende Softwarelösungen für Open Data Plattformen wie etwa CKAN oder Socrata, setzen die in dieser Arbeit vermuteten und beschriebenen Anforderungen um, und sind weltweit bereits vielfach im Betrieb.

CKAN ist hier *de facto* die Standard-Software für die Veröffentlichung von Open Data. Im Rahmen dieses Unterkapitels wurden die Architektur und einige der nennenswertesten Funktionen dieser Software vorgestellt, um die vielseitigen Möglichkeiten aufzuzeigen, die Datenbereitstellern auf der einen, und Bürgern auf der anderen Seite, zur Verfügung stehen. Besonders Bürger profitieren hiervon sehr: Sie erhalten über ein Web Interface direkten Zugang zum gesamten Datenkatalog, und können darüber hinaus mit Hilfe einer technischen Schnittstelle (API) auch maschinell auf die veröffentlichten Rohdaten zugreifen. Letztere Zugriffsmöglichkeit ist insbesondere für Journalisten sowie Entwickler von Apps interessant, die üblicherweise für ihre Zwecke die Datensätze aggregieren, analysieren oder anderweitig weiterverarbeiten.

CKAN ist zu guter Letzt betrachtet, ein ausgereiftes, offenes Projekt, das ein solides Entwicklungsmodell besitzt, und eine wachsende Gemeinschaft von Entwicklern zu sich zählt. Innerhalb kürzester Zeit wurde CKAN auf internationaler Ebene von großen öffentlichen Projekten als technische Lösung zur Umsetzung von Open Government Data Portalen eingesetzt. Aus den oben genannten Gründen ist deswegen damit zu rechnen, dass auch in Zukunft immer mehr Bereitsteller von Open Data sowie Open Government Data auf diese Lösung setzen werden.

3.4. Verknüpfung von Datenbeständen

Die Förderung der Wiederverwendung oder Nachnutzung der Daten ist ein Aspekt, der neben der öffentlichen Bereitstellung ebenfalls von zentralem Interesse ist und verfolgt werden sollte. Damit aus Daten verwertbare und nützliche Informationen für Bürger

sowie Unternehmen gewonnen und letztlich konkrete Fragestellungen beantwortet werden können, müssen die Daten in Beziehung zueinander gestellt werden können.

Für den interoperablen Austausch und Wiederverwertungsprozess von Daten im Internet, hat sich hierbei zunehmend der Linked Data Ansatz bewährt (Barnickel und Klessmann, 2012, S. 142). In den folgenden Abschnitten wird dieser Ansatz näher erläutert.

3.4.1. Linked Data und Linked Open Data

Das Konzept *Linked Data* geht im Wesentlichen auf den Erfinder des World Wide Web Tim Berners-Lee zurück und beschreibt eine Methode und zugehörige Technologien, um strukturierte Daten über das Internet zu veröffentlichen und zu vernetzen (Berners-Lee, 2009). Bizer u. a. (2009) beschreiben genauer:

> „Linked data refers to data published on the Web in such a way that it is machine-readable, its meaning is explicitly defined, it is linked to other external data sets, and can in turn be linked to from external data sets."

Der Begriff *Linked Open Data* bezieht sich wiederum konkret auf die Bereitstellung von Open Data im Internet, bei der eine Vernetzung der Daten über Domänen und Organisationsgrenzen hinweg erreicht wird (Klessmann u. a., 2012, S. 433). In von Lucke und Geiger (2010, S. 4) wird dazu folgende Arbeitsdefinition gegeben:

> „Offene vernetzte Daten sind sämtliche Datenbestände, die im Interesse der Allgemeinheit der Gesellschaft ohne jedwede Einschränkung zur freien Nutzung, zur Weiterverbreitung und zur freien Weiterwendung frei zugänglich gemacht und über das World Wide Web miteinander vernetzt sind."

Durch die oben erwähnte Vernetzung und mit der Unterstützung von offenen Schnittstellen (APIs) können Anwendungen und Tools das Web wie eine strukturierte Datenbank für spezifische Anfragen oder gezieltes Browsen nutzen, und dadurch beispielsweise automatisiert Recherchen und Berichterstattung unterstützen. Der Ansatz beruht dabei im Wesentlichen auf der Idee des Semantic Web, das eine Erweiterung des World Wide Web darstellt und vorsieht, dass Informationen im Web von Anwendungen interpretierbar sein sollen und verarbeitet werden können (Berners-Lee u. a., 2001).

In den folgenden Abschnitten wird dazu das Datenmodell, basierend auf der Beschreibung von Barnickel und Klessmann (2012, S. 142-144), erklärt sowie die damit verbundenen Technologien übersichtsartig vorgestellt. Da Linked Data als Oberklasse auch den Begriff Linked Open Data mit einschließt, wird in dieser Arbeit der Begriff Linked Data bevorzugt verwendet.

3.4.2. Das Datenmodell

Das zugrunde liegende Datenmodell basiert auf der Resource Description Framework (RDF) Spezifikation des World Wide Web Consortium (W3C) und zeichnet sich durch seine Flexibilität und Plattformunabhängigkeit aus. Anders als im relationalen Datenmodell (das mit Zeilen und Spalten in Datenbanktabellen arbeitet) oder im hierarchischen Datenmodell (das an eine Baumstruktur wie beispielsweise in XML-Dokumenten gebunden ist) besteht das Datenmodell des Linked Data Ansatzes aus gerichteten Graphen, bei dem insbesondere die Beziehungen zwischen den Daten mit einer höheren Flexibilität definiert werden können. Abbildung 3.2 soll den Ansatz beispielhaft skizzieren.

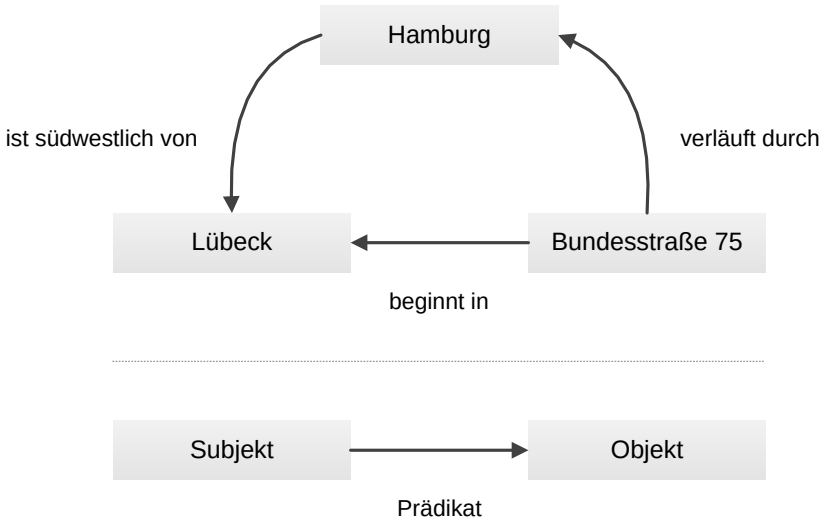

Abbildung 3.2.: Graph-basiertes Datenmodell des Linked Data Ansatzes in Anlehnung an Barnickel und Klessmann (2012), S. 143.

Die Idee des Modells sieht dabei vor, Datensätze und ihre Fakten als Tripel, bestehend aus Subjekt, Prädikat und Objekt, darzustellen, wobei alle drei Elemente des Tripels über eine global eindeutige Adresse im Internet auffindbar sind – den Uniform Resource

Identifier (URI) des Hypertext Transfer Protocols. Durch die Verwendung von URIs wird somit nicht nur die Austauschbarkeit von Daten über das Web gefördert, sondern es wird dadurch auch möglich, Datensätze aus den unterschiedlichsten Quellen (und zugrundeliegenden IT-Systemen) miteinander zu vernetzen und kombinierbar zum Abruf bereit zu stellen, um ein *Web of Data* herzustellen. Die Verknüpfung von Datensätzen trägt außerdem zur Entstehung von Informationsmehrwerten bei; durch die definierten Beziehungen zwischen den Daten wird der Kontext eines Datensatzes dargelegt und abrufbar gemacht. Mit Hilfe dieser Kontextbeschreibungen ist es dann Nutzern und Anwendungen möglich, die Daten zu verarbeiten und einzuordnen, um daraus anschließend Informationen zu generieren, die zur Beantwortung einer konkreten Fragestellung dienen können.

3.4.3. Technologien und Standards

Die Schlüsseltechnologien auf denen Linked Data basiert, sind Standards der Internet Engineering Task Force (IETF) sowie der Initiativen für Semantic Web, die durch das World Wide Web Consortium (W3C) errichtet wurden. In Anlehnung an die Beschreibungen von Barnickel und Klessmann (2012, S. 145-147) soll folgend ein kurzer Überblick zu diesen Standards gegeben werden. Eine umfassende Vorstellung dieser ist nicht Bestandteil dieser Arbeit. Zur Vertiefung empfehlen sich die offiziellen Spezifikationen der jeweiligen Standards (IETF (1994), IETF (1999), W3C (2014a), W3C (2014b), W3C (2004) und W3C (2008)).

3.4.3.1. Uniform Resource Identifier (URI)

URIs werden als eindeutige Bezeichner für Ressourcen verwendet, die eine Dateneinheit repräsentieren. Anhand dieser können andere Bezug auf diese Ressourcen nehmen. Als bekanntestes Beispiel hierfür gilt der Einsatz von URIs als Webadressen: Beim Aufruf einer solchen Adresse in einem Webbrowser, greift dieser auf die entsprechende Dateneinheit zu.

3.4.3.2. Hypertext Transfer Protocol (HTTP)

HTTP ist das Kommunikationsprotokoll, das für den Aufruf von URIs und den Transport der Dateneinheiten über das Internet verwendet wird. Analog wie beim Aufruf von Webseiten über einen Webbrowser, wird eine gewünschte URI mit dem HTTP-Befehl GET aufgerufen, den der adressierte Server daraufhin erhält. Dieser bearbeitet die Anfrage und liefert ebenfalls unter Verwendung des HTTP-Protokolls die angeforderte Ressource als Antwort zurück.

3.4.3.3. Resource Description Framework (RDF)

Der Resource Description Framework Standard definiert das Datenmodell, das für die Beschreibung von Aussagen über Ressourcen verwendet wird. Für die Beschreibung der Ressourcen werden Tripel definiert, für dessen Elemente (Subjekt, Prädikat und Objekt) jeweils eine eindeutige URI festgelegt ist (vgl. Abschnitt 3.4.2).

3.4.3.4. RDF Schema (RDFS) und Web Ontology Language (OWL)

Das RDF-Datenmodell ermöglicht es, einem Subjekt über beliebige Prädikate beliebige Objekte zuzuordnen. Diese Beliebigkeit kann mit den Ontologie-Sprachen RDFS und OWL eingeschränkt werden, indem für eine Anwendung festgelegt wird, welche Klassen von Dateneinheiten im jeweiligen Kontext als sinnvoll gelten und welche Beziehungen sie untereinander haben können. Bei dem Beispiel in Abbildung 3.2 wäre es z.B. sinnvoll nur Subjekte die aus einer zu definierenden Klasse „Straße“ stammen, dem Prädikat „verläuft durch“ zuzuweisen.

3.4.3.5. SPARQL Protocol and RDF Query Language

SPARQL ist eine Abfragesprache, die es erlaubt, komplexe Abfragen über Linked Data-Quellen zu machen. Die Abfrage erfolgt dabei analog einer Datenbank-Abfrage mit Hilfe von Variablen und eingrenzenden Attributen, sodass sich innerhalb des Linked Data-Netzes optimierte Recherchemöglichkeiten über die Daten anbieten.

3.4.4. Linked Data in der Praxis

Für die praktische Umsetzung von Linked Data im Open Data Bereich, hat Tim Berners-Lee den Vorschlag für ein 5-Sterne Bewertungssystem gemacht (Berners-Lee, 2009). Dieses Bewertungssystem soll Datenbereitstellern als „Fahrplan“ für die stufenbasierte Umsetzung dienen, und beschreibt die aufeinander aufbauenden Schritte für die Bereitstellung von Open Data, bei der im fünften und letzten Schritt auf eine Verlinkung der Daten abgezielt wird (siehe Abbildung 3.3). Im Rahmen des Vorschlages von Berners-Lee wird außerdem ein zusätzliches Kriterium genannt, das eine Anreicherung der bereitgestellten Daten mit Metadaten vorsieht. Diese Metadaten können zum Beispiel den zeitlichen Bezug oder den Ursprung der Datenquellen beschreiben und zugehörige Schlagwörter enthalten.

Die Daten sind über das Web in einem beliebigen Format verfügbar und besitzen eine offene Nutzungslizenz.

Die Daten sind strukturiert und in einem maschinenlesbaren Format verfügbar.

Die Daten sind in einem nicht proprietären Format verfügbar (z.B. CSV anstatt Excel).

Die Daten werden unter Verwendung offener Standards des W3C (RDF und SPARQL) bereitgestellt.

Die Daten werden mit anderen verlinkt, um ihren Kontext explizit verfügbar zu machen.

Abbildung 3.3.: Das 5-Sterne Bewertungssystem für Open Data nach Berners-Lee (2009).

Im Open Data und auch speziell Open Government Data Bereich gibt es bereits erste Praxisbeispiele, die den Linked Data Ansatz verfolgen und zunehmend mehr einsetzen. Hierzu zählt als eines der bekanntesten Beispiele das Portal data.gov.uk von Großbritannien, bei dessen Umsetzung Tim Berners-Lee mitgewirkt hat (data.gov.uk, 2014). Als weiteres Beispiel kann die Deutsche Nationalbibliothek genannt werden, die auf ihrer Webseite sämtliche nationalbibliografische Daten über einen Linked Data-Service zur Verfügung stellt (Deutsche Nationalbibliothek, 2014). Erwähnenswert ist auch, dass durch den Linked Data Ansatz ein Netzwerk von miteinander verlinkten

offenen Daten entstanden ist, das nach Dietrich (2011b) bereits im Jahr 2011 aus über 25 Milliarden Fakten bestand, die über ca. 400 Millionen Links miteinander vernetzt sind. Diese Datenwolke wird auch als *Linked Open Data Cloud* bezeichnet, und verbindet Datensätze aus öffentlichen und privaten Datenbeständen des Semantic Web, wie zum Beispiel geografische Daten aus Open Street Map[6] und Linked GeoData[7] oder enzyklopädische Daten aus DBpedia[8] (Dietrich, 2011b).

3.4.5. Zwischenfazit

Durch die Etablierung einheitlicher und semantisch reichhaltiger Zugänge zu Open Data steigert sich der Mehrwert für verschiedene Zielgruppen. Datenbereitsteller können unter Verwendung des Linked Data Ansatzes ihre Datenbestände mit den verschiedensten heterogenen Datenquellen interoperabel in Relation stellen (sowohl national als auch international) und damit neue Potenziale eröffnen. Die zugrunde liegenden praxistauglichen Standards stellen hierfür die nötige Grundlage dar und bieten Mechanismen zur Datenabfrage an, auf dessen Basis beispielsweise Entwickler Anwendungen erstellen oder Journalisten Recherchen betreiben können. Das maschinenlesbare Format von Linked Data sorgt hier schließlich für eine leichte Weiterverarbeitung und Nachnutzung durch externe Akteure.

Bei der praktischen Umsetzung des Linked Data Ansatzes spielt aber im ersten Schritt die Wahl der Technologie noch keine wichtige Rolle, sondern es geht vorrangig um die grundsätzliche Öffnung der Daten zur freien Nachnutzung, wie ferner auch in der Publikation von Barnickel und Klessmann (2012, S. 147) betont wird. Hier dient das von Tim Berners-Lee vorgeschlagene 5-Sterne Bewertungssystem als Orientierung, das Bereitstellern von Open Data letztendlich bei der Umsetzung behilflich sein soll. Im Bereich des öffentlichen Sektors ist diesbezüglich jedoch zu beobachten, dass die Umsetzung von Linked Data noch nicht als essenzielle Aufgabe bei der Bereitstellung von Open Data aufgefasst wurde. Zum gegenwärtigen Zeitpunkt werden, verhältnismäßig betrachtet, nur sehr wenige Datensätze als Linked Data zur Verfügung gestellt. Die Anreicherung von Open Government Data mit semantischen Annotationen gemäß dem Linked Data Prinzip, sollte somit in Zukunft intensiver von den Bereitstellern verfolgt werden.

[6] http://www.openstreetmap.org/
[7] http://www.linkedgeodata.org/
[8] http://www.dbpedia.org/

3.5. Fazit

Für Open Data und auch den Teilbereich Open Government Data gibt es mittlerweile eine gewisse Standardisierung in Bezug auf die Sichtweise und Dienstleistungen, die im Zusammenhang mit der technischen Bereitstellung dieser Daten angeboten werden. Die entscheidenden benötigten Elemente für die Umsetzung einer Infrastruktur zur Bereitstellung von Open Data sind heute gegeben und bereits zahlreich im Einsatz.

Die im Unterkapitel 3.2 erwähnten Herausforderungen die sich bei der Speicherung, Verwaltung und Verarbeitung der weltweit immer weiter wachsenden Datenmenge ergeben, betreffen auch den Open Data Bereich. So sind diese Herausforderungen auch speziell für den Open Government Data Bereich von hoher Relevanz, da die Behörden des öffentlichen Sektors täglich große Mengen von Daten erzeugen und verarbeiten, und sofern zutreffend, auch veröffentlichen. Die Big Data-Technologien die in diesem Zusammenhang für die Bewältigung von großen Datenbeständen vorgeschlagen wurden, sind inzwischen gut erprobt und somit praxistauglich. Unternehmen die ihre Kernaktivitäten auf der Basis von Big Data betreiben, setzen heute erfolgreich auf diese Technologien. In Anbetracht an die stetig wachsende Anzahl von erzeugten Daten und die Öffnung von zukünftig immer mehr Informationen der öffentlichen Behörden, sollte untersucht werden, ob sich der Einsatz moderner Big Data-Technologien in Open Government Data Portalen nutzbringend auswirkt. Hier ist vor allem interessant zu prüfen, wie die Skalierbarkeit in Bezug auf die Verwaltung der Daten sowie den Zugriff auf diese beeinflusst wird. Der Weiterverarbeitung und Analyse von Open Data durch Dritte können diese Technologien letztendlich ebenfalls zugutekommen.

Im weiteren Verlauf des Kapitels wurden Datenplattformen vorgestellt, die auf die Veröffentlichung von Open Data ausgelegt sind, und mit ihren Features die Beteiligung sowohl von öffentlichen Behörden als auch von Bürgern, Unternehmen und sonstigen Organisationen ermöglichen. In diesem Bereich haben sich vor allem die Softwarelösungen CKAN und Socrata bewährt, die für die Bereitsteller von Daten eine geeignete Plattform zur Verwaltung und Veröffentlichung bieten, und für die Endnutzer zugleich geeignete Zugriffsmöglichkeiten auf die Daten zur Verfügung stellen. Für den maschinellen Zugriff auf die Daten bieten beide Lösungen standardisierte APIs an, die es Entwicklern ermöglichen, die Daten in neuen Anwendungen weiter zu verwenden. Für die Bereitstellung von Open Data über das Web hat sich zuletzt vor allem die Open Source Softwarelösung CKAN als geeignete Plattform erwiesen.

Das zum Ende des Kapitels vorgestellte Linked Data Konzept ist als Forschungsgebiet unter der Bezeichnung *Semantic Web* inzwischen gut überprüft und etabliert. In der Praxis stellt die Technologie allerdings noch eine weniger wichtige Rolle dar. Open Data wird aber zu einem ersten möglichen Erfolg in der Umsetzung des Semantic Web beitragen, besonders da in Zukunft mit einer erhöhten Bereitstellung von Open Data als Linked Data zu rechnen ist. Hierzu müssen die Bereitsteller an die Vernetzung der Daten von Beginn an denken, da eine nachträgliche Anreicherung einen erhöhten Aufwand bedeutet. Linked Data werden noch eine gewisse Zeit lang eine Ausnahme im Bereich von Open Government Data sein; doch sie sollten laut Klessmann u. a. (2012, S. 434) als eine gute, da derzeit interoperabelste Beschreibung von Daten anerkannt werden. Wird Open Data nach dem Linked Data Prinzip veröffentlicht und vernetzt, entsteht eine Möglichkeit, diese Daten als wertvolle Informationsquelle nutzen und weiterverarbeiten zu können.

4. Interpretation von Datenbeständen

4.1. Einleitung

Das eigentliche Potenzial der Vielfalt an Daten, die in den verschiedensten Anwendungsbereichen zur Verfügung stehen – Open Data mit eingeschlossen – liegt nicht in dessen Rohform, sondern in den oftmals versteckten Informationen, die diese enthalten. Im Bereich des Journalismus zum Beispiel, nutzen Redakteure immer häufiger auch Open Data als unterstützendes Material für ihre Berichterstattung oder sogar als wesentlichen Auslöser und Grundlage für die Erzählung einer Geschichte. Dazu ist es jedoch erforderlich, die vorliegenden Daten erst einmal auszuwerten und auf interessante Strukturen zu durchsuchen, die einen Informationsgehalt aufweisen, der für den Anwender von Nutzen sein könnte.

Zur Auswertung der Datensätze, die teils einen großen Umfang besitzen, reicht es heute jedoch meist weder, einen kurzen Blick auf diese zu werfen, noch klassische Werkzeuge aus der Statistik auf ihnen anzuwenden. Vielmehr werden andere Techniken benötigt, die auf die Analyse von Datenbeständen ausgerichtet sind. Diese Techniken sollen etwa nach Mustern, Auffälligkeiten oder Zusammenhängen suchen, die dem Anwender eine neuwertige bzw. brauchbare Information liefern.

Die Suche nach Mustern oder Zusammenhängen in Daten wird auch mit dem Begriff *Data Mining* zusammengefasst. Für die dazugehörige Disziplin der Extraktion von Wissen aus großen Datenbeständen hat sich in der Fachliteratur auch der Begriff *Knowledge Discovery in Databases* (KDD) etabliert. Das vorliegende Kapitel befasst sich mit diesen Begriffen näher, und soll ein grundlegendes Verständnis zu dem Bereich der Interpretation von Datenbeständen vermitteln.

Nachdem zunächst in Abschnitt 4.2 eine allgemeine Definition der Begriffe gegeben wird, beschreibt Abschnitt 4.3 den Ablauf einer Datenanalyse. In Abschnitt 4.4 werden

anschließend die bekanntesten Verfahren zur Extraktion von Wissen aus Datenbeständen aufgeführt und anhand von kurzen Beispielen veranschaulicht. Abschnitt 4.5 liefert abschließend ein Fazit zu den behandelten Inhalten. Diese Arbeit behandelt dabei jedoch nicht die mathematischen Aspekte des Data Mining.

4.2. Knowledge Discovery in Databases und Data Mining

Für den Begriff Wissensentdeckung in Datenbanken (engl. *Knowledge Discovery in Databases*, kurz: KDD) wird in der Fachliteratur oft folgende Definition von Fayyad u. a. (1996) zitiert:

> Wissensentdeckung in Datenbanken ist der nichttriviale Prozess der Identifikation gültiger, neuer, potenziell nützlicher und schlussendlich verständlicher Muster in (großen) Datenbeständen.

Der Begriff *Data Mining* bezeichnet hierbei einen Teilschritt des KDD-Prozesses, in dem die eigentliche Analyse anhand der Suche und Bewertung von Hypothesen stattfindet (Wrobel u. a., 2014, S. 409). Data Mining wird allerdings auch oft als Synonym für die Begriffe „Knowledge Discovery in Databases“ und „Datenmustererkennung“ verwendet (Alpar und Niedereichholz, 2000, S. 3). In der vorliegenden Arbeit werden die Begriffe Data Mining und Knowledge Discovery in Databases synonym verwendet, bevorzugt wird jedoch der Begriff Data Mining.

Data Mining als solches, ist ein interdisziplinäres Thema, das Methoden aus den wissenschaftlichen Gebieten Statistik, Künstliche Intelligenz, Visualisierung bzw. Computergrafik und Datenbanken vereint (Cleve und Lämmel, 2014, S. 11-12). Im Vergleich zu klassischen Verfahren der Statistik, wird Data Mining jedoch meist für die Analyse sehr großer Datenbestände eingesetzt, weswegen die Data Mining-Algorithmen spezielle Anforderungen in Bezug auf ihre Laufzeit erfüllen müssen. Zusätzlich dazu sollen die Data Mining-Verfahren auch semi-automatisch durchgeführt werden können, ohne vom Anwender zu verlangen, dass dieser Fachstatistik-Wissen besitzt (Müller, 2013, S. 75).

Als Untergebiet der Künstlichen Intelligenz, findet sich Data Mining heutzutage in vielen praktischen Anwendungen wieder. Im Finanzsektor zum Beispiel, nutzen Banken die Möglichkeiten des Data Mining, um Bonitätsprüfungen für Kunden vorzunehmen

und somit den Vorgang der Kreditvergabe zu unterstützen. Des Weiteren wenden auch Onlineshops Data Mining-Verfahren an, um das Kaufverhalten ihrer Nutzer zu analysieren, und auf dessen Basis gezielte Werbung zu schalten oder Empfehlungen für andere Produkte zu geben.

4.3. Ablauf einer Datenanalyse

Vor Beginn der Datenanalyse bzw. des Data Mining-Prozesses müssen zunächst die Ziele klar sein, die damit verfolgt werden. So muss als allererstes ein Verständnis des Anwendungsbereichs und des bereits bekannten Anwendungswissens aufgebaut werden. Basierend auf dieser Grundlage wird das Ziel des Data Mining definiert, das zur Extraktion von neuem, nützlichen Wissen führen soll (Fayyad u. a., 1996). Für den anschließenden Ablauf einer Datenanalyse wird in der Fachliteratur überwiegend auf das Modell von Fayyad zurückgegriffen, das sich in mehrere Phasen aufteilt, und in Abbildung 4.1 aufgezeigt wird.

In der Fachliteratur wird noch ein zweites Modell für Data Mining-Prozesse erwähnt, das CRISP-Modell das für *Cross Industry Standard Process for Data Mining* steht. Dieses Modell spiegelt die Sicht der Industrie auf Data Mining-Projekte wider, während sich im Vergleich das Fayyad-Modell auf die eigentliche Bereitstellung der Daten und dessen Analyse fokussiert (Cleve und Lämmel (2014, S. 6-8); Müller (2013, S. 76-78)). Da der Literatur nach jedoch bestimmte Phasen des CRISP-Modells sehr stark vom jeweiligen Projekt abhängig sind, orientiert sich die vorliegende Arbeit an dem Modell von Fayyad.

Basierend auf den Ausführungen von Cleve und Lämmel (2014, S. 9-11), werden hierzu im Folgenden die verschiedenen Phasen dieses Modells beschrieben.

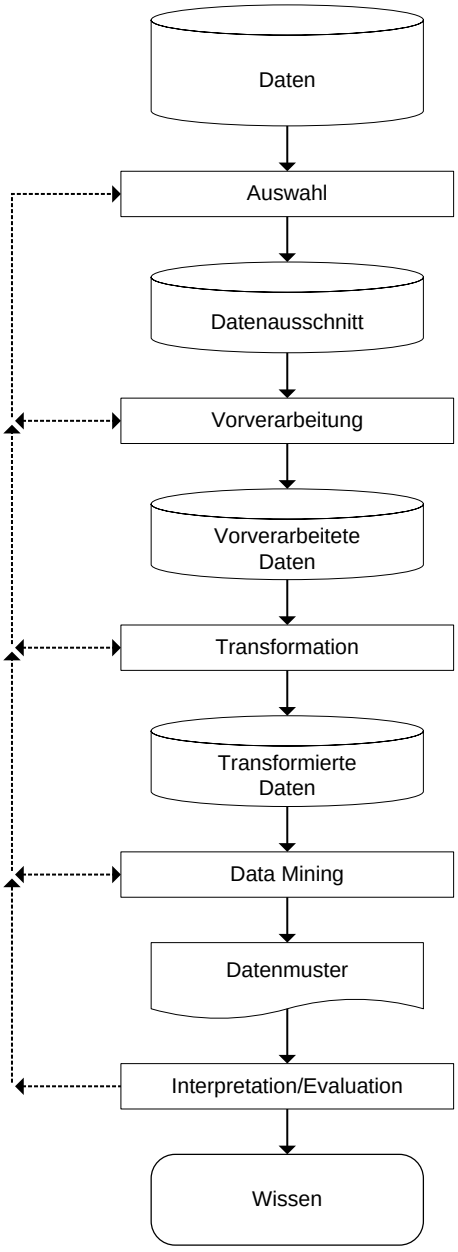

Abbildung 4.1.: Der Data Mining-Prozess nach Fayyad u. a. (1996).

4.3.1. Selektion

In diesem ersten Schritt wird bestimmt, aus welchen Daten neues Wissen generiert werden soll, bzw. welche Daten für die Analyse als geeignet erscheinen und einbezogen werden sollen. Hierzu müssen die benötigten Daten aus ihren Quellen exportiert werden und in einen Zieldatenbestand überführt werden. Kann ein Export der ausgewählten Daten wegen technischen oder rechtlichen Einschränkungen nicht stattfinden, muss die Selektion der Daten entsprechend neu durchgeführt werden. Einschränkungen können beispielsweise fehlende Zugriffsrechte auf einen gewünschten Datenbestand sein, oder etwa Kapazitäts- und Datentyp-Beschränkungen, die im Zielsystem vorliegen. Diese Restriktionen können in manchen Fällen umgangen werden, indem der ursprünglich einzubeziehende Gesamtdatenbestand auf eine repräsentative Teilmenge von Daten eingeschränkt wird, die für die Analyse verwendet wird.

4.3.2. Datenvorverarbeitung

Die selektierten Daten, die zuvor aus ihren Datenquellen extrahiert wurden, können aufgrund technischer oder menschlicher Fehler fehlerhafte Elemente enthalten, die sich negativ auf die Qualität der Untersuchungsergebnisse auswirken können. Datenbestände, die falsche Angaben aufweisen, können zu verfälschten Ergebnissen führen, von denen die Benutzer der Data Mining-Verfahren möglicherweise keine Kenntnis nehmen. Ziel der Datenvorverarbeitung ist es, den zu analysierenden Datenbestand auf seine Zuverlässigkeit und Korrektheit zu prüfen, und sofern erforderlich, mit Hilfe geeigneter Verfahren zu bereinigen und zu verbessern.

Ungewöhnliche Werte in einem Datenbestand die stark vom Durchschnitt abweichen (deswegen auch *Ausreißer* genannt), können zum Beispiel einen potenziellen Fehler darstellen. Diese Wertausprägungen können korrekt erfasste Daten sein, die für die Analyse zu berücksichtigen sind; es kann sich hierbei aber auch um fehlerhafte Daten handeln, die es durch Weglassen aus dem betroffenen Datenbestand auszuschließen gilt. Für die Behandlung fehlender Werte, die eine der häufigsten Fehlerart darstellen und sich leicht identifizieren lassen, gibt es ebenfalls unterschiedliche Techniken, die angewendet werden können.

4.3.3. Transformation

Die zu analysierenden Datenbestände weisen in ihrer ursprünglichen Form oftmals ein inadäquates Format auf, das sich nicht für die Data Mining-Verfahren eignet. Abhängig von dem Verfahren das angewendet werden soll, ist deswegen häufig eine Umwandlung der Daten in ein entsprechend geeignetes Datenbankschema erforderlich, um den jeweils spezifischen Anforderungen an die Datenstruktur der Eingangsdaten gerecht zu werden. Dies geschieht in der Phase der Datentransformation, bei der die Erzeugung neuer Attribute oder Datensätze erfolgt bzw. die Transformation existierender Attribute vorgenommen wird. Hierbei werden beispielsweise textuelle Informationen in eindeutige Schlüssel oder Codierungen umgewandelt, um konstante Datendarstellungsformen sicherzustellen. Oder es werden Wertebereiche eingeschränkt, um etwa die Menge der möglichen Ausprägungen zu reduzieren (Dimensionsreduktion). So eine Dimensionsreduktion kann zum Beispiel durch die Verwendung von Taxonomien (Klassifikationsschemata) oder erzeugten Wertintervallen erfolgen. Ein fiktiver Wert 18.296 der etwa für die Einwohnerzahl einer Stadt steht, könnte zum Beispiel durch den Wert < 25.000 ersetzt werden. Eine derartige Transformation ändert demzufolge nicht nur die Granularität der Daten, sondern bringt nach Alpar und Niedereichholz (2000, S. 6-7) auch Informationsverluste mit sich, welche es zu berücksichtigen gilt.

4.3.4. Data Mining

In dieser Phase erfolgt die eigentliche Durchführung der Analyse. Hierzu wird zunächst ein geeignetes Data Mining-Verfahren ausgewählt, das anschließend auf dem vorbereiteten Zieldatenbestand angewendet wird. Dazu wird als Erstes bestimmt, welche grundlegende Anwendungsklasse (z.B. Clusterbildung oder Assoziationsanalyse, siehe auch Abschnitt 4.4) vorliegt. Anschließend folgt die Wahl eines konkreten Data Mining-Verfahrens, das sich für die gegebene Problemstellung eignet. Bevor die Analyse mit Hilfe des gewählten Verfahrens stattfinden kann, muss dieses noch konfiguriert werden. Hierzu werden bestimmte methodenspezifische Parameter eingestellt, welche beispielsweise die minimalen relativen Häufigkeiten festhalten, die für einen Interessantheitsfilter verwendet werden sollen, oder welche die Gewichtungsfaktoren für einzelne Eingabevariablen oder Attribute bestimmen, die bei der Suche nach Mustern berücksichtigt werden sollen. Wurde eine geeignete Konfiguration gefunden, kann das gewählte Data

Mining-Verfahren durchgeführt werden, welches schließlich ein Modell generiert, das für die Phase der Interpretation und Evaluation zugrunde gelegt wird.

4.3.5. Interpretation und Evaluation

Der letzte Schritt befasst sich mit der Interpretation und Auswertung der aufgedeckten Muster und Zusammenhänge. Entsprechen diese Muster dem Kriterium der Interessantheit, d.h. den Anforderungen der Gültigkeit, Neuartigkeit, Nützlichkeit und Verständlichkeit, so lassen sich aus diesen neue Erkenntnisse ableiten und darstellen. Die Data Mining-Ergebnisse müssen also zunächst anhand der Dimensionen der Interessantheit bewertet werden:

- Die Dimension der *Gültigkeit* misst, wie zuverlässig ein entdecktes Muster auch für neue Daten Geltung besitzt.
- Die *Neuartigkeit* zeigt auf, inwiefern ein gefundenes Muster das bereits bekannte Wissen erweitert oder eine Gegensätzlichkeit beweist.
- Das Kriterium der *Nützlichkeit* ermisst den praktischen Nutzwert des Musters für den Anwender.
- Die *Verständlichkeit* erfasst, wie weit die Bedeutung eines Musters vom Anwender nachvollzogen wird.

Die anschließende Schaffung von konkreten Handlungsmaßnahmen zur Umsetzung der gewonnenen Erkenntnisse setzt an dieser Stelle eine Interpretation der zugrunde liegenden Muster voraus. Die zuvor genannten Aspekte der Interessanheit treffen jedoch bei weitem nicht immer auf die gefundenen Muster zu, sodass eine Interpretation in diesen Fällen überflüssig ist. Häufig handelt es sich bei den Mustern um bereits bekanntes, irrelevantes oder sogar nicht nachvollziehbares Wissen, das für den Anwender keinen Nutzwert trägt. Erfüllen die Muster wiederum die erforderlichen Kriterien, so wird für dessen korrekte Interpretation und Bewertung ein umfangreiches Wissen über die Anwendungsdomäne benötigt, damit abschließend die gewonnenen Erkenntnisse das Domänenwissen des Anwenders effektiv bereichern. In dieser Phase werden die Erkenntnisse meist in einer geeigneten Form aufbereitet bzw. visualisiert, die es dem Anwender ermöglicht, das extrahierte Wissen besser nachzuvollziehen und umzusetzen.

4.3.5.1. Bedeutung von Visualisierung

Die Ergebnisse, die am Ende aus einem Data Mining-Projekt erzielt werden, sind meist in einer geeigneten Form darzustellen, sodass zum einen die gefundenen Muster ersichtlich bzw. greifbar gemacht werden, und zum anderen auch das Verständnis des Anwenders über die Problemstellung und die Daten unterstützt wird (Cleve und Lämmel, 2014, S. 235). „Eine gute Visualisierung ist für den Erfolg eines Data-Mining-Projekts unerlässlich“, so Cleve und Lämmel. William Cleveland (Cleveland, 1993) schreibt ebenfalls in diesem Zusammenhang:

> „Visualization is critical to data analysis. It provides a front line of attack, revealing intricate structure in data that cannot be absorbed in any other way. We discover unimagined effects, and we challenge imagined ones.“

Eine verständliche Veranschaulichung der Ergebnisse belegt außerdem die Gültigkeit des extrahierten Wissens und hilft, dieses auch anderen Personen zugänglich zu machen und letztere von dessen Korrektheit und Nützlichkeit zu überzeugen (Cleve und Lämmel, 2014, S. 235-244).

Die letztendliche Bewertung liegt beim Anwender. Für die Auswertung der visualisierten Ergebnisse wird somit, wie auch schon im vorhergehenden Abschnitt angedeutet, ein umfangreiches Domänenwissen des Anwenders vorausgesetzt. Die Repräsentation der Data Mining-Ergebnisse spielt gleichzeitig eine ebenfalls wichtige Rolle, weswegen die Visualisierung ein für Menschen adäquates Format aufweisen muss. Hier treten in diesem Zusammenhang auch Begriffe wie *Visual Analytics* (Bertini und Lalanne, 2010) oder *explorative Datenvisualisierung* in den Vordergrund, die jedoch nicht weiterer Bestandteil der Untersuchung dieser Arbeit sind.

An dieser Stelle sei angemerkt, dass der Data Mining-Prozess ein iteratives Verfahren ist, bei dem häufig eine Wiederholung der vorhergehenden Phasen erforderlich ist, um die gewünschten Ziele bzw. Ergebnisse zu erreichen. So kann es beispielsweise nötig sein, die Parameter der gewählten Data Mining-Technik anzupassen, oder den verwendeten Datenbestand neu zu bestimmen und vorzubereiten, um zu den erwarteten Ergebnissen zu gelangen. Cleve und Lämmel schätzen zuletzt, dass alleine die Phase der Datenvorbereitung ganze 80% des gesamten Data Mining-Prozesses einnimmt.

Für die Durchführung der Analysen können verschiedene Werkzeuge eingesetzt werden. So sind nach Cleve und Lämmel (2014, S. 16-17) zum einen Tabellenkalkulationsprogramme ein hilfreiches Werkzeug, da mit ihnen die zu analysierenden Daten einer ersten Untersuchung unterzogen werden können, und beispielsweise Beziehungen zwischen Attributen identifiziert werden können. Zum anderen bietet sich auch spezielle Data Mining-Software an, wie zum Beispiel Rapid Miner[1] oder der IBM SPSS Modeler[2].

4.4. Anwendungsklassen

Data Mining teilt sich in verschiedene Anwendungsklassen auf, entsprechend den Zielen, mit denen die Analyse eines Datenbestandes betrieben wird. So gibt es beispielsweise den typischen Anwendungsfall, bei dem eine große Datenmenge einer Datenanalyse unterzogen wird, um diese Menge auf explorative Weise nach potenziellen nützlichen Informationen zu durchsuchen. Hier ist also zunächst noch nicht bekannt, wonach genau in den Daten gesucht werden soll. Eine Möglichkeit ist dann, aus den Daten Gruppen von ähnlichen Objekten zu bilden (auch bekannt unter dem Begriff *Clustering* der in Abschnitt 4.4.2 näher erläutert wird). Wurden solche Gruppen, auch Cluster genannt, gefunden, so werden aus diesen Klassen gebildet, indem ihnen Namen vergeben werden. Sind Daten bereits in klassifizierter Form vorhanden – zum Beispiel Stadtteil-gebundene Daten mit einer Einordnung in niedrige, mittlere oder hohe Kriminalitätsrate – so wird mit Hilfe des *Klassifikationsverfahrens* (Abschnitt 4.4.3) nach Regeln gesucht, die bisher unerfasste Stadtteile in eine dieser Klassen einordnen können.

Ein weiterer Anwendungsfall kann sich wiederum mit der Frage beschäftigen, ob bestimmte Zusammenhänge bzw. Abhängigkeiten zwischen Attributen eines Datenbestandes existieren. Mit dieser Fragestellung befasst sich das Verfahren der *Assoziationsanalyse*, das in Abschnitt 4.4.4 genauer beschrieben wird.

Neben den bereits genannten Verfahren gibt es weitere, die konkret zur Analyse von Texten und Webseiten verwendet werden. Diese sind zum Beispiel für die Auswertung

[1] http://www.rapidminer.com
[2] http://www.ibm.com/software/analytics/spss/products/modeler/

der vielfachen Berichte, Protokolle, Verordnungen und sonstigen Dokumente interessant, die im öffentlichen Sektor anfallen und auf den entsprechenden Open Data Portalen meist in Form von PDF- oder DOC-Dateien veröffentlicht werden. Die konkreten Verfahren für die Auswertung solcher Daten werden in den Abschnitten 4.4.5 und 4.4.6 kurz vorgestellt.

Für die folgenden Erläuterungen der verschiedenen Anwendungsklassen wird primär die Publikation von Cleve und Lämmel (2014, S. 57-67) als Grundlage verwendet. Zu der Clusteranalyse, der Klassifikation und der Assoziationsanalyse werden neben einer kurzen Beschreibung auch jeweils die bekanntesten Anwendungsverfahren aufgelistet. Eine detaillierte Vorstellung dieser ist jedoch nicht Bestandteil dieser Arbeit.

4.4.1. Bedeutung von Abstands- und Ähnlichkeitsmaßen

Das Auffinden von Mustern und Beziehungen erfolgt in vielen Anwendungen im Data Mining durch den Vergleich von Datensätzen. So wird beispielsweise bei der Clusteranalyse versucht, ähnliche Objekte zu Gruppen zusammenzufassen (siehe auch Abschnitt 4.4.2). Um eine solche Zusammenfassung vornehmen zu können, muss jedoch die Ähnlichkeit von zwei Datensätzen quantifizierbar sein. Hierzu werden meist sogenannte *Abstandsmaße* eingesetzt, welche nach der Definition von Cleve und Lämmel (2014, S. 43) die Distanz zwischen zwei Datensätzen messen und dessen „Unähnlichkeit“ bestimmen: Je größer der Abstand ist, desto geringer ist die Ähnlichkeit. Cleve und Lämmel sprechen an dieser Stelle von Abstandsfunktionen, oder auch sogenannten *Distanzfunktionen*, welche jedoch in der vorliegenden Arbeit nicht genauer behandelt werden. Damit sich mehrere Datensätze miteinander vergleichen lassen können, ist es zuletzt erforderlich, dass diese ein geeignetes Format aufweisen, das eine Anwendung von Distanzfunktionen erlaubt. Unstrukturierte Daten, wie zum Beispiel eingescannte Textdokumente, die in PDF-Dateien eingebettet sind, eignen sich etwa nur bedingt, und lassen sich kaum oder nur eingeschränkt mit Distanzfunktionen vergleichen.

4.4.2. Cluster-Analyse

Die *Clusteranalyse* (kurz auch *Clustering* oder *Clusterbildung* genannt) stellt eine bekannte Data Mining-Anwendung dar, bei der eine gegebene Instanzenmenge *E*

eines Instanzenraumes X in Teilmengen (*Cluster*) aufgeteilt wird. Objekte die dem selben Cluster zugehören, sollen dabei möglichst ähnlich sein (homogen), während hingegen Objekte verschiedener Cluster möglichst unähnlich zueinander sein sollen. Für die Durchführung der Clusteranalyse wird somit nicht nur die Instanzenmenge benötigt, sondern insbesondere auch eine Distanzfunktion, die zur Bestimmung der Ähnlichkeit zweier Datensätze erforderlich ist, sowie eine Qualitätsfunktion, die einen Vergleich von Clusterbildungen ermöglicht.

Konkrete Verfahren, die zur Durchführung einer Clusteranalyse angewendet werden können, sind etwa:

- Der k-Means-Algorithmus
- Das k-Medoid-Verfahren
- Das Verfahren der Erwartungsmaximierung
- Die agglomerative Cluster-Bildung
- Die dichtebasierte Cluster-Bildung
- Einige Architekturen künstlicher neuronaler Netze, wie z.B. selbstorganisierende Karten, neuronale Gase sowie die adaptive Resonanz Theorie (ART-Netze)

Die Clusteranalyse findet in der Praxis zum Beispiel im Rahmen von Marketingmaßnahmen Anwendung, um homogene Gruppen von Kunden zu suchen, für die jeweils, angepasst auf den Kundentyp, gezielte Angebote erstellt werden können. Da hierbei nach bisher unbekannten Klassen von ähnlichen Objekten gesucht wird, wird die Clusteranalyse auch oft als *Klassenbildung* bezeichnet. Als weiteres Beispiel kann die Ermittlung von Gebieten mit ähnlicher Entwicklungstendenz genannt werden, die 2003 von der Stadt Berlin im Rahmen einer Untersuchung zur sozialen Stadtentwicklung durchgeführt wurde (Senatsverwaltung für Stadtentwicklung Berlin, 2003). Hier wurden mit Hilfe des Clusteranalyse-Verfahrens Gebiete auf der Basis von Werten für verschiedene einzelne Indikatoren typisiert und zu Gruppen zusammengefasst, die eine vergleichbar ähnliche Entwicklungstendenz aufweisen.

4.4.3. Klassifikation

Die *Klassifikation* beschäftigt sich mit der Einteilung von Daten in Klassen. Bei diesem Verfahren wird eine Menge von Trainingsdaten festgelegt, die bereits in klassifizierter Form vorliegen und die Grundlage für die spätere Klassenzuordnung von neuen Daten bilden sollen. Dazu wird zunächst innerhalb eines Lernprozesses ein Klassifikationsmodell angelernt, welches beschreibt, wie die Eigenschaften der vorliegenden Trainingsdaten die Einordnung in bestimmte Klassen beeinflussen. Sobald ein gutes Klassifikationsmodell vorliegt, kann die Einordnung neuer Datensätze in die vorgegebenen Klassen erfolgen.

Bekannte Verfahren, die für eine Klassifikationsaufgabe eingesetzt werden können, sind die folgenden:

- Das instanzenbasierte k-Nearest-Neighbour-Verfahren
- Die Erzeugung von Entscheidungsbäumen
- Der Naive-Bayes-Algorithmus, der sich auf Wahrscheinlichkeiten stützt
- Die vorwärtsgerichteten neuronalen Netze als Teilbereich der künstlichen neuronalen Netze
- Das Verfahren der Support Vector Machines

Ein denkbares Anwendungsbeispiel wäre etwa die Vorhersage über den Erfolg eines Straßenfestes, das jedes Jahr wiederholt stattfindet. Bei diesem Beispiel müssen für die Analyse die verschiedenen Faktoren berücksichtigt bzw. einbezogen werden, die maßgeblich über den Erfolg mitbestimmen. Solche Faktoren können etwa die Wetterbedingungen oder Sozialatlas-Daten der umliegenden Nachbarschaftsgebiete sein. Es muss vor allem aber auch im Vorfeld definiert werden, wann ein Straßenfest überhaupt „erfolgreich“ ist. Erst wenn diese Vorbedingungen erfüllt sind, kann die eigentliche Durchführung des Klassifikations-Verfahrens erfolgen.

4.4.4. Assoziationsanalyse

Die *Assoziationsanalyse* befasst sich mit dem Auffinden von Beziehungen zwischen den vorhandenen Attributen eines Datensatzes und zählt aktuell nach Wrobel u. a.

(2014, S. 442) zu den beliebtesten Analyseverfahren im Data Mining. Das in der Literatur wohl bekannteste Beispiel bei dem dieses Verfahren zum Einsatz kommt, ist das sogenannte *Warenkorbanalyseproblem*, das sich mit der Fragestellung beschäftigt, unter welchen Voraussetzungen bestimmte Artikel innerhalb einer Transaktion im Supermarkt zusammen gekauft werden. So können für dieses Beispiel im Rahmen der Assoziationsanalyse Regeln der Form „Wer A kauft, kauft häufig auch B“ hergestellt werden, die nicht nur mit Hilfe von Wahrscheinlichkeiten bestimmte Regelmäßigkeiten im Kundenverhalten beschreiben, sondern auch dabei helfen, das Verhalten neuer Datensätze zu prognostizieren.

Bei der Assoziationsanalyse sind primär zwei Anwendungsverfahren zu nennen:

- Das A-Priori-Verfahren
- Frequent Pattern Growth

Neben der Analyse des Kundenverhaltens in einem Supermarkt oder Onlineshop, kann die Assoziationsanalyse beispielsweise auch in der Versicherungsbranche praktische Anwendung finden, um bei der Abschätzung bestimmter Risiken zu helfen.

4.4.5. Text Mining

Text Mining ist ein Analyseverfahren, das auf Texten angewendet wird, um Informationen aus diesen zu extrahieren, die für den Anwender von potenziellem Nutzen sein können. Als Anwendungsfälle wären beispielsweise eine Zusammenfassung von Texten nach Ähnlichkeit oder eine Einordnung von Dokumenten nach Themengebiet denkbar, bei der aus den Dokumenten relevante Begriffe extrahiert werden, um mit Hilfe letzterer die Klassifizierung vornehmen zu können.

Die Vorgehensweise beim Text Mining ist dabei analog zu dem Ablauf des Data Mining-Prozesses, der in Abschnitt 4.3 beschrieben wird. Nach der Begriffsdefinition in Abschnitt 2.2.1 fallen Texte in die Kategorie der unstrukturierten Daten, weswegen im ersten Schritt die Extraktion der relevanten bzw. interessanten Informationen aus den Textdokumenten stattfinden muss. Relevante Informationen können zum Beispiel Schlüsselwörter, die Verteilung von Wörtern nach der Anzahl ihres Vorkommens im Text, oder eine erste Einteilung nach Kategorien oder hierarchischen Gruppen sein. Begriffe, die nicht von Relevanz sind, müssen wiederum herausgefiltert werden; im Text verwendete Abkürzungen müssen entsprechend identifiziert werden. Im Rahmen

der Vorverarbeitung des Textes können je nach Anwendungsfall weitere Maßnahmen erforderlich sein. Die eigentliche Datenanalyse kann zuletzt erst stattfinden, sobald eine reduzierte Menge von Wörtern vorliegt, die Abstandsmaße erlaubt.

4.4.6. Web Mining

Unter dem Begriff *Web Mining* wird die Anwendung von Techniken des Data Mining verstanden, die zur Mustererkennung in Daten eingesetzt werden, die dem World Wide Web entstammen. Je nachdem, ob sich die Analyse des World Wide Web mit dem Inhalt oder der Nutzung befasst, wird spezifischer von den Teilgebieten *Web Content Mining* und *Web Usage Mining* gesprochen.

Web Content Mining ist der Prozess der Extraktion nützlicher Informationen aus den Inhalten von im Web befindlichen Dokumenten. Inhaltsdaten entsprechen dabei der Sammlung von Fakten, die eine Webseite enthält und vermitteln soll. Beispiele hierfür sind etwa Texte, Bilder, Audios, Videos, strukturierte Informationen wie etwa Tabellen sowie Hyperlinks zu anderen Webseiten.

Das Teilgebiet des *Web Usage Mining* befasst sich wiederum mit der Entdeckung von interessanten Verhaltensmustern bei der Nutzung des Internets. Hierbei werden die Protokolldateien von Webservern einer Data Mining-Analyse unterzogen, um etwa Informationen über das Verhalten oder die Interessen der Besucher dieser Server zu erhalten.

4.5. Fazit

Die vorliegende Arbeit bietet nur einen groben Überblick zu dem Bereich der Interpretation von Datenbeständen. Mithilfe moderner Data Mining-Lösungen, die im Rahmen dieses Kapitels behandelt wurden, können sowohl einzelne Dokumente aus dem Internet als auch komplette Datenbank-Bestände analysiert werden. Die Anwendung dieser Lösungen ermöglicht das Auffinden interessanter Muster oder Zusammenhänge innerhalb dieser Bestände, die schließlich als neues gewonnenes Wissen Nutzen für die verschiedensten Anwendungsbereiche bringen können.

Ein großer Nachteil des Data Mining liegt jedoch in dessen Komplexität. Das Gebiet der Wissensentdeckung in Datenbanken ist von großem Umfang, und die Anwendung der vorgestellten Data Mining-Verfahren setzt umfassende Kenntnisse über die Domäne und die zu analysierenden Daten voraus. Eine Beurteilung und Interpretation der extrahierten Muster kann ohne dieses Wissen nicht stattfinden. Anwendern, die mit der Verfahrensweise der Algorithmen unvertraut sind und keine fundierten Kenntnisse in den Bereichen Statistik, Datenbanken und Informatik mitbringen, bleibt letztlich ein sinnvoller Einsatz der Möglichkeiten des Data Mining vorenthalten.

Außerdem ist eine sorgfältige Vorbereitung der Daten unerlässlich, da dieser Schritt ausschlaggebend die Qualität der gewonnenen Erkenntnisse beeinflusst. Unvorbereitete bzw. nicht ausreichend bereinigte Daten können zu Verfälschungen in den Ergebnissen führen, und damit eine Auswertung schnell unbrauchbar oder fehlerhaft machen. Letztendlich ist für die Bewertung sowie den Erfolg der Datenanalyse auch die Visualisierung der Ergebnisse maßgeblich entscheidend, da sie diese grafisch darstellt und darüber hinaus auch als Technik im Data Mining eingesetzt wird, um interessante Muster oder Beziehungen zwischen Attributen zu identifizieren.

Abschließend sei darauf hingewiesen, dass im Rahmen der Analyse und Interpretation von Datenbeständen keine Erfolgsgarantie für die Erfüllung der gesetzten Ziele gegeben ist. Es kann vorkommen, dass eine Analyse zu bestimmten Erkenntnissen führt, die im Voraus gar nicht erwartet wurden. Andersherum kann auch der Fall eintreten, dass eine umfangreiche und zeitintensive Analyse gar keine Resultate liefert. Deswegen ist oftmals eine wiederholte Anwendung der Phasen des Data Mining-Prozesses notwendig, in der etwaige Anpassungen in Bezug auf den Datenbestand, das eingesetzte Verfahren oder dessen Parameter vorgenommen werden müssen.

Zum gegenwärtigen Zeitpunkt bietet die Wissenschaft der Informatik und speziell das Gebiet der Künstlichen Intelligenz eine Reihe erprobter Techniken, die zur Extraktion von Wissen aus beliebig großen Datenbeständen in bereits vielen Bereichen erfolgreich eingesetzt werden. Diese Techniken des Data Mining werden auch in Zukunft weiterhin von Bedeutung sein, da die damit gewonnenen Erkenntnisse neue Potenziale für die verschiedensten Anwendungsbereiche versprechen.

5. Daten-gestützte narrative Strukturen

5.1. Einleitung

In dem vorherigen Kapitel wurde beschrieben, wie mit Hilfe des Gebietes des Data Mining, neues, nützliches Wissen aus Datenbeständen und somit auch aus Open Data gewonnen werden kann. Vielfache Anwendungen aus den verschiedensten Bereichen können dann im Weiteren mit diesen gewonnenen Erkenntnissen arbeiten und sie vervielfältigen. Im Bereich des Journalismus hat sich in diesem Zusammenhang in den letzten Jahren der Begriff *Data-driven journalism* oder auch zu dt. *Datenjournalismus* als ein neues Genre der Berichterstattung geformt, der Gebrauch von Open Data macht und diese als Ausgangsbasis und Gegenstand zur Erzählung einer Geschichte einsetzt. Immer mehr Medienhäuser versuchen heute an diese Daten anzuknüpfen, um mit ihnen neue, interessante Fragestellungen zu beantworten, dessen Antworten möglicherweise versteckt in den Datensätzen stecken und erstmal extrahiert werden müssen. Im Rahmen eines journalistischen Beitrags werden diese Erkenntnisse dann mit neuen Erzählformen an den Leser gebracht, der sich schließlich in einer interaktiven Umgebung mit der Geschichte, aber auch dem zugrundeliegenden Datensatz auseinandersetzen kann. Simon Rogers, ehemaliger Datenjournalist beim britischen Guardian, beschreibt die Rolle der Journalisten auch „als Brücke zwischen denen, die die Daten haben (und es nicht hinkriegen, sie verständlich zu machen), und der Öffentlichkeit, die nach Antworten fragt, die Daten einsehen und verstehen möchte, dabei aber Unterstützung benötigt.“ (Rogers, 2012). Der Datenjournalismus macht Open Data somit für die breite Masse greifbar.

Dieses Kapitel befasst sich mit dieser Thematik näher, und soll dem Leser hierzu einen ersten Überblick über den Bereich des Datenjournalismus verschaffen. Dazu wird zunächst einleitend in den Abschnitten 5.2 und 5.3 der Hintergrund erwähnt, der bei der Entstehung des Datenjournalismus eine Rolle gespielt hat. Im Anschluss befasst sich Abschnitt 5.4 mit dem noch relativ neuen Genre des Datengetriebenen Journalismus.

Im Rahmen dieses Unterkapitels wird zunächst eine Definition des Begriffes gegeben sowie die Bedeutung der Datenanalyse und -aufbereitung für dieses neue Genre vor Augen geführt. Anschließend werden die verschiedenen Erzählformen und Methoden, die dem Datenjournalismus zugeordnet werden können, beschrieben und anhand von Praxisbeispielen näher gebracht. Im weiteren Verlauf wird in Abschnitt 5.5 die Rolle der IT in diesem Feld diskutiert, und schließlich in Abschnitt 5.6 auch die, im Zuge des Medienwandels, veränderte Rolle des Journalisten beschrieben. Abschnitt 5.7 liefert zum Schluss ein Fazit zu den behandelten Inhalten.

5.2. Die Medien im Wandel

Die Digitalisierung der Medien und der daraus entstandenen Informationsgesellschaft hat Veränderungen in vielen Bereichen herbeigeführt. Während in den vergangenen Jahren die Nutzung von gedruckten Medien immer weiter nachgelassen hat, und infolgedessen sinkende Verkaufszahlen in den Verlagshäusern ausgelöst hat, ist dagegen der Konsum von digitalen Angeboten auf mobilen Endgeräten zunehmend gestiegen. Dies bestätigen auch Ergebnisse der Onlinestudie, die jährlich von den öffentlich-rechtlichen Rundfunkanstalten ARD und ZDF zur Nutzung der Medien erhoben wird: Knapp 80 Prozent der deutschen Bevölkerung waren im Jahr 2014 online, und davon griff bereits jeder Zweite auch von unterwegs auf Onlineangebote zu (van Eimeren und Frees, 2014). Dies ist besonders der schnellen Verbreitung von mobilen Endgeräten wie Smartphones, Tablets und Notebooks zuzuschreiben, die als weitere Plattformen den unmittelbaren Zugang zu Netzinhalten möglich machen.

Auch der heutige Umgang mit den medialen Inhalten hat sich geändert: Die Bevölkerung greift schnell auf Informationen im Internet zu, während sie die Zeitung durchblättert, oder sie nutzt soziale Netzwerke, um Meinungen über das aktuell ausgestrahlte Fernsehprogramm auszutauschen. Neue Medien haben ihre alten Vorreiter somit nicht ersetzt, sondern sie bestehen parallel zueinander, und werden ebenso parallel genutzt (Bunz, 2012, S. 104). Für diese Parallelnutzung wird in der Branche auch der Begriff *Second Screen* verwendet, der zu den weiteren Veränderungen zählt, auf die sich die Industrie einstellen muss (Busemann und Tippelt, 2014).

Neben den genannten Transformationen dürfen die Anbieter jedoch letztendlich nicht die Tatsache aus den Augen verlieren, dass ihre Inhalte unter der schieren Masse

von Angeboten dem Nutzer einer allgemeinen Wahlfreiheit unterliegen. Deswegen ist es für die Medienhäuser umso unerlässlicher, ihre Inhalte als Marken zu etablieren, die geräte- und plattformübergreifend angeboten werden (van Eimeren und Frees, 2014). Diese wichtige Botschaft ist heute bereits bei vielen Anbietern von medialen Inhalten angekommen: Eine Vielzahl von Fernsehanstalten stellt zum Beispiel ihr einst exklusiv im Fernsehen übertragenes Programm nun auch als Videostream in Online-Mediatheken zur Verfügung, und erlaubt somit einen nachträglichen Abruf der Inhalte. Hier ist also durchaus eine Tendenz zur crossmedialen sowie zeit- und ortsunabhängigen Bereitstellung der medialen Inhalte zu beobachten.

5.3. Der Journalismus im Wandel

Auch der Journalismus als ein Format mit langer Tradition muss sich auf die veränderte Landschaft im Bereich der Medien einstellen. Hier hat sich besonders in den vergangenen Jahren ein entscheidender Wandel vollzogen, der sich auf verschiedenen Ebenen bemerkbar macht. Dieser fängt bereits bei den Verlagshäusern an, die ihre Zeitungen vormals nur in gedruckter Form angeboten haben und sich heutzutage neu erfinden müssen, um aus ökonomischer Sicht weiterhin schwarze Zahlen zu schreiben. Seit der Digitalisierung haben die Zeitungen ihre Leserschaft sowie Einnahmen aus geschalteten Anzeigen zu wesentlichen Teilen an das Internet verloren (Bunz, 2012, S. 105-106). Um diese Verluste größtmöglich abzufedern, haben nach Bunz die Verlagshäuser begonnen, neue Einnahmequellen zu erschließen, die etwa Nutzen aus neuen technischen Möglichkeiten ziehen, und dabei helfen, Material unterschiedlich zu arrangieren, aufzumachen, und mehrfach zu verwerten. So lässt sich feststellen, dass heute immer mehr Zeitungen ihre Artikel auch online veröffentlichen, oder als E-Paper zum kostenpflichtigen Download anbieten, wie es beispielsweise das Hamburger Abendblatt[1] tut.

Bei dieser Veränderung hat auch die rege Verbreitung von mobilen Endgeräten eine wesentliche Rolle gespielt. Simon Sturm spricht hierzu in seiner Publikation „*Digitales Storytelling*“ von einem Wandel in der Produktion und Rezeption geistiger Inhalte, der sich durch die Übertragung dieser Inhalte auf die neuen digitalen und mobilen Plattformen ergibt (Sturm, 2013, S. 3-4). Sturm zufolge, erleben die klassischen Medienarten

[1] http://epaper.apps.abendblatt.de/

dadurch eine Änderung in ihrer Gestalt, die besonders auch von der äußerlichen Darstellung sowie technischen Grundlage bestimmt wird. Für den Journalismus bedeutet diese Entwicklung konkret, dass die Formen des sogenannten *Storytelling*, also die Möglichkeiten zur Erzählung von journalistischen Inhalten, sich an die veränderte Landschaft in der digitalisierten Welt anpassen müssen (Sturm, 2013, S. 4). Hier geht es also auch vor allem darum, die Inhalte auf den neuen mobilen Endgeräten in einer anderen, geeigneten Weise zu erzählen und an den Leser zu überbringen. Aus diesem Grund wird an dieser Stelle auch der Begriff *Digital Storytelling* verwendet.

5.4. Datengetriebener Journalismus

Der klassische und zugleich qualitative Journalismus, der vor allem aus dem Medium Print vertraut ist, benötigt, wie im Vorfeld angedeutet, neue Darstellungs- und Verbreitungsformen, um mit der fortschreitenden technischen und gesellschaftlichen Entwicklung mithalten zu können. Ein neues Genre des Journalismus, das in diesem Zusammenhang entstanden ist und angetrieben durch die Open Data-Bewegung zunehmend in den Vordergrund tritt, ist der *Datenjournalismus* oder auch *Datengetriebene Journalismus*, der in den Onlineangeboten von Zeitungen ein neues Zuhause gefunden hat. Die folgenden Abschnitte befassen sich mit dieser neuen Journalismusform näher, und sollen dem Leser einen ersten Überblick über das Gebiet verschaffen.

5.4.1. Definition von Datenjournalismus

Der Begriff *Datenjournalismus* steht nach der Beschreibung von Lorenz Matzat für ein neues Genre des Onlinejournalismus, bei dem sich die Berichterstattung auf Informationen stützt, die aus Datensätzen bezogen werden (Matzat, 2011). Der korrespondierende englische Begriff *Data-driven journalism* (kurz: DDJ oder auch zu dt. „Datengetriebener Journalismus“) gibt diese Bedeutung besser wieder.

Die Methoden, die bei diesem Genre angewendet werden, bauen auf der seit über 50 Jahren bekannten computergestützten Recherche (engl. *Computer-assisted reporting*) auf, die laut Matzat und weiterer Literatur (Open Data Network (2010); Elmer und Wormer (2014)) als direkter Vorfahre des Datenjournalismus gilt. Diese Form der Recherche wird, Matzat zufolge, im englischsprachigen Raum schon seit Jahrzehnten angewendet

und auch an Journalismusschulen vermittelt. Die Hauptaufgabe der computergestützten Recherche ist dabei, auffällige Muster oder Unregelmäßigkeiten in großen Datensätzen aufzudecken, die als Grundlage für eine weitergehende Recherche dienen könnten (Matzat, 2011). Während hier jedoch die Daten nur als Beweismittel oder Unterfütterung für einen journalistischen Beitrag dienen, geht der Datenjournalismus laut Matzat einen Schritt weiter: Die Daten werden zum „zentralen Gegenstand der Geschichte und deren Präsentation“.

Beim Datenjournalismus geht es, der Beschreibung von Christina Elmer und Holger Wormer nach (Elmer und Wormer, 2014), um noch mehr:

> „Guter Datenjournalismus kann tiefe Einblicke, exklusive und relevante Geschichten liefern, die sich nicht ohnehin schon viral verbreiten. Und im Idealfall lassen sich die recherchierten Daten und ihre Geschichte gleich interaktiv an die Lebenswirklichkeit des Publikums anbinden.“

Die „tiefen Einblicke“, wie sie Elmer und Wormer beschreiben, werden dabei im Datenjournalismus erst durch die Aufbereitung und Darstellung komplexer Zusammenhänge ermöglicht, die in den Datensätzen stecken (Matzat, 2011). Durch die anschließende Bereitstellung einer interaktiven Rechercheumgebung, können sich, Matzat zufolge, die Leser dann idealerweise mit der Thematik der Geschichte genauer auseinandersetzen. Im Bestfall werden die zugrundeliegenden Rohdaten in maschinenlesbaren, offenen Formaten angeboten (Open Data Network, 2010). Damit werden die Leser ermächtigt, die Quelle und Grundlage des Berichts einzusehen, aber auch weiterzuverwenden.

Eine weitere Form des Journalismus, die ihre Berichterstattung auf der Basis von Daten macht, ist der *Roboterjournalismus* (engl. Robot journalism), der sich jedoch vom zuvor erwähnten Datenjournalismus abgrenzt. Bei dieser Form führen programmierte Algorithmen die Auswertung der Datensätze aus, und übernehmen das Verfassen von Texten, und somit auch das Erzählen einer Geschichte (Bunz, 2012, S. 14). Eine weitere Auseinandersetzung mit diesem Gebiet ist jedoch nicht Bestandteil dieser Arbeit.

5.4.2. Bedeutung von Datenanalyse und -aufbereitung

Die Analyse der aggregierten Daten steht im Mittelpunkt von datenjournalistischen Projekten. Journalisten entdecken mit der eingehenden Auseinandersetzung von Daten neues Wissen und Geschichten in ihnen, allerdings auch „auf die Gefahr hin, dass sie

zwar viel Zeit investieren, aber nichts entdecken“ (Leßmöllmann, 2012). Die Analyse und Aufbereitung der Daten stellt somit eine der wichtigsten Aufgaben im Datenjournalismus dar, sie nimmt allerdings auch einen großen Anteil an Arbeitszeit in Anspruch. So sind nach eigenen Erfahrungswerten des Datenjournalisten Lorenz Matzat „mindestens 50 Prozent der Arbeit damit verbunden, die Daten zu erheben und so aufzubereiten, dass sie sich überhaupt sinnvoll erschließen lassen“ (Matzat, 2014). Nach Schätzungen des ehemaligen Guardian-Datenjournalisten Simon Rogers gehen bei den meisten Projekten sogar 70 Prozent der gesamten Arbeitszeit auf die Beschaffung, Prüfung und Bearbeitung der Daten, bevor die Visualisierung überhaupt vorgenommen werden kann (Lorenz, 2013).

Das große Potenzial, das in der Analyse von Daten liegt, haben heute jedoch bereits viele Medienhäuser erkannt. Im Ausland sind hier etwa der Guardian und die New York Times für ihre datenjournalistische Arbeit bekannt. In Deutschland befassen sich wiederum vor allem die ZEIT ONLINE, der Spiegel und die Berliner Tageszeitung taz mit der Analyse von Daten für redaktionelle Beiträge. Neben diesen bieten inzwischen auch externe Unternehmen Dienstleistungen im Bereich des Datenjournalismus an. Ein Beispiel hierfür ist etwa die Datenjournalismus-Agentur OpenDataCity[2], die sich mit der Recherche und Bearbeitung großer Datenmengen aus journalistischer Perspektive beschäftigt. Die gewonnenen Informationen werden hier meist in webbasierten Applikationen aufbereitet, und können des Weiteren als datengetriebene Recherche-Umgebung dienen (OpenDataCity, 2015).

5.4.3. Erzählformen und Methoden

Im Bereich des Datenjournalismus existieren verschiedene Erzählformen, die sich nach Lorenz Matzat in sechs verschiedene Formen aufteilen: *Datastorytelling*, *Echtzeitdaten*, *Datensätze*, *Crowdsourcing*, *Hyperlokal* und *Newsgames* (Matzat, 2011). Im Folgenden werden die verschiedenen Formen basierend auf den Beschreibungen von Matzat übersichtsartig vorgestellt. Die entsprechenden Praxisbeispiele, die dabei zu Veranschaulichungszwecken vorgebracht werden, verwenden nicht nur zwangsweise Open Data aus dem Bereich des öffentlichen Sektors, sondern bedienen sich zusätzlich auch anderer Datensätze.

[2] https://opendatacity.de/

An dieser Stelle wird darauf hingewiesen, dass in der vorliegenden Arbeit nur annähernde Definitionen bzw. Beschreibungen zu den folgenden Erzählformen und Methoden gegeben werden können, da es laut Woytewicz (2013, S. 27-38) teilweise Mischformen der verschiedenen Formen gibt, und „die Grenzen zwischen ihnen fließend verlaufen". Damit ist konkret gemeint, dass beispielsweise ein und derselbe Datensatz als Auslöser für ein Datastorytelling-, Datensatz- und Hyperlokal-Werk dienen kann.

5.4.3.1. Datastorytelling

Datastorytelling dreht sich im Wesentlichen um einen Datensatz, mit dem sich der Leser auseinandersetzen kann, um sich Zusammenhänge zu verdeutlichen. Laut Matzat meint hier Storytelling, „dass eine Geschichte erzählt wird indem der Nutzer sich in dem gesetzten Rahmen nach eigenem Gusto über Detail- und Hintergründe eines Vorgangs erkundigen kann" (Matzat, 2011). Der Nutzer kann sich also selbständig und auf interaktive Weise mit einem Beitrag befassen: er kann sich etwa online durch die Datenbank bzw. oftmals visualisierte Karte bewegen, „sich Ausschnitte betrachten oder Zusatzinformationen abrufen" (Matzat, 2011). Der Auslöser der Berichterstattung ist dabei meistens der Datensatz selbst, so Matzat. Ein inzwischen „ikonisches" Beispiel hierfür, wie Leßmöllmann (2012) beschreibt, ist die Veröffentlichung der Bewegungen und Aktivitäten des Grünen-Abgeordneten Malte Spitz, die von der ZEIT ONLINE-Redaktion aus den gesammelten Mobilfunk-Vorratsdaten des Politikers rekonstruiert, und auf einer interaktiven Karte visualisiert wurden (siehe auch Abbildung 5.1) (ZEIT ONLINE, 2011).

5.4.3.2. Echtzeitdaten

Bei *Echtzeitdaten*-Werken handelt es sich der Definition von Matzat nach um das gleichzeitige Sammeln von Daten und deren gleichzeitige Aufbereitung (Matzat, 2011). Die Berichterstattung erfolgt somit noch meist während des Ereignisses, welches die Daten beschreiben, und erlaubt den Lesern auf diese Weise, in Echtzeit, aber auch im Nachhinein, Geschehnisse zu verfolgen und nachzuvollziehen. Als ein bekanntes Beispiel für Echtzeitdaten-Werke gilt das Projekt „Zugmonitor" der Süddeutschen Zeitung, das zwischen 2011 und 2013 den Fernverkehr der Deutschen Bahn anhand der Angaben zur Pünktlichkeit der einzelnen Züge erfasst hat, die auf der Bahn-Website

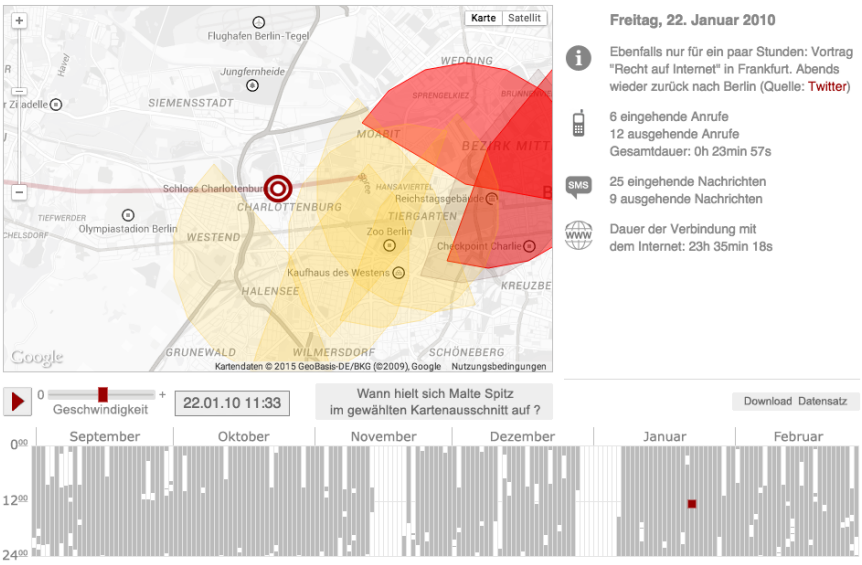

Abbildung 5.1.: Visualisierung „Verräterisches Handy“ auf zeit.de auf Basis der Mobilfunk-Vorratsdaten des Politikers Malte Spitz (ZEIT ONLINE, 2011).

abgerufen werden können. Die Verspätungen der Züge wurden schließlich in Echtzeit auf einer Karte visualisiert, die stets aktualisiert wurde (Süddeutsche Zeitung, 2015).

5.4.3.3. Datensätze

Als weitere Methode des Datenjournalismus kann nach Matzat auch die schlichte Bereitstellung von *Datensätzen* angesehen werden (Matzat, 2011). Die Bereitstellung als solche, bekommt dadurch einen Dienstleistungs-Charakter, allerdings steckt hier meist eine journalistische Rechercheleistung dahinter, die nach Matzat, im Zusammentragen und der Aufbereitung der strukturierten Informationen liegt. Als Beispiel für diese datenjournalistische Methode kann der Datablog[3] des britischen Guardian genannt werden, der zu seinen Berichten die vollständigen, zugrundeliegenden Rohdaten zur Verfügung stellt (Rogers, 2015).

[3] http://www.theguardian.com/data

5.4.3.4. Crowdsourcing

Das Verfahren des *Crowdsourcing* beschreibt im Datenjournalismus den Einbezug von Lesern in die Recherchearbeit und Auswertung von Quellen (Bunz, 2012, S. 109). Hier wird dann, Matzat zufolge, „auf die Effizienz der Menge gesetzt“ (Matzat, 2011). Crowdsourcing ist etwa dann sinnvoll, wenn eine maschinelle Auswertung von Datensätzen nicht möglich ist, zum Beispiel, wenn Dokumente aus dem öffentlichen Sektor in eingescannter und damit zugleich nicht-maschinenlesbarer Form vorliegen (was in der Praxis aktuell noch häufig der Fall ist). Als ein Beispiel, bei dem Crowdsourcing groß zum Einsatz kam, kann etwa der Guardian genannt werden, der im Jahr 2009 seine Leser unter dem Motto „Überprüfen Sie die Ausgaben Ihres Abgeordneten“ aufgefordert hat, die teils handschriftlich verfassten Spesenabrechnungen ihrer Parlamentsabgeordneten auf Auffälligkeiten zu prüfen, um so den britischen Parlamentsspesenskandal aufzudecken (Bunz, 2012, S. 109-110).

5.4.3.5. Hyperlokal

Unter der Erzählform *Hyperlokal* wird das Zusammenführen von Informationen aus den verschiedensten Bereichen verstanden, die anschließend auf Mikroebene von Stadtvierteln oder Straßen dargestellt werden (Matzat, 2011). Das nach Matzat beste Beispiel in Deutschland, ist hierzu die Plattform Frankfurt Gestalten[4], die straßengenau über lokalpolitische Ereignisse berichtet, und nach eigenen Angaben ein „Radar der Stadtentwicklung“ ist (Frankfurt Gestalten, 2015). Ein weiteres Beispiel für hyperlokale Berichterstattung sind die Eimsbütteler Nachrichten[5], eine Online-Zeitung, die Themen des Hamburger Stadtviertels Eimsbüttel aufgreift, welche laut selbiger „buchstäblich vor der Haustür liegen“ (Eimsbütteler Nachrichten, 2015).

5.4.3.6. Newsgames

Bei der Erzählform *Newsgames* werden laut Matzat (2011) die Nachrichten mit Hilfe des *Gamification*-Prinzips, also durch Verwendung spielerischer Elemente, an den Leser gebracht. Hierbei handelt es sich nicht um eine geschlossene Form, die ausschließlich

[4] http://www.frankfurt-gestalten.de/
[5] http://www.eimsbuetteler-nachrichten.de/

im Datenjournalismus wiederzufinden ist; es gibt aber, Matzat zufolge, „Möglichkeiten mit Hilfe von Datenbanken Spielprinzipien zu nutzen“, um zuletzt „den Leser mehr zu engagieren und zu involvieren“. Matzat erwähnt hier als ein konkretes Beispiel aus der Praxis die New York Times, die 2010 ihre Leserschaft einlud, eigene Vorschläge für Kürzungen des Staatshaushaltes abzugeben, indem diese mit einem Datensatz „spielen“ konnte und sich auf diese Weise auch direkt mit dem Thema auseinandersetzen konnte.

5.5. Die Rolle von IT in der Redaktion

Durch den Einzug des Datenjournalismus in die Redaktionen und dem damit einhergehenden Auftakt von neuen Erzählformen, treten verstärkt auch Prozesse der Softwareentwicklung in diese ein. Dies bestätigt auch der Datenjournalist Lorenz Matzat in seinem Fachartikel über Datenjournalismus (Matzat, 2014). Die Methoden des Datenjournalismus erfordern, neben der klassischen Arbeit mit Content Management Systemen zur Ausspielung der Inhalte, auch die Auseinandersetzung mit verschiedenen anderen Werkzeugen und Techniken der Informatik, um etwa die Schritte der Vorverarbeitung und Analyse der Daten zu unterstützen, die später als Grundlage für eine neue Geschichte dienen sollen. Hier werden zum Beispiel Tabellenkalkulationsprogramme oder die in Kapitel 4 beschriebenen Data Mining-Verfahren eingesetzt, welche vom Anwender jedoch umfangreiche Kenntnisse aus dem Bereich der Wissensentdeckung in Datenbanken voraussetzen (siehe hierzu auch Kapitel 4). Neben der Datenanalyse besteht für die Datenjournalisten aber auch die Aufgabe, die Daten bzw. aus ihnen gewonnene Informationen in einer verständlichen Weise zu visualisieren und interaktiv zu machen. Speziell für diese Aufgabe gibt es heute bereits eine Reihe von Tools, wie zum Beispiel DataWrapper[6], CartoDB[7], Tableau[8] und Google Docs/Fusion Tables[9]. Diese Werkzeuge bieten nach der Beurteilung von Matzat jedoch nur einen eingeschränkten Funktionsumfang. Sie erlauben wenig Spielraum für eigene Anpassungen (etwa hinsichtlich der Gestaltung), weswegen sie sich nur für einen ersten Einstieg in den Bereich der (interaktiven) Datenvisualisierung eignen (Matzat, 2014).

[6] https://www.datawrapper.de/
[7] http://www.cartodb.com/
[8] http://www.tableau.com/
[9] https://support.google.com/fusiontables/answer/2571232

Redaktionen bzw. Teams, die sich gründlicher mit der Datenvisualisierung aber auch den übrigen Teilaufgaben des Datenjournalismus befassen möchten, können demnach nicht nur auf die fertigen Tools setzen. Matzat betont aus diesem Grund die Notwendigkeit von Statistik- und Programmierkompetenzen innerhalb des Teams (Matzat, 2014). Die Teams benötigen demnach nicht nur eine passende technische Infrastruktur, sondern sie müssen idealerweise auch Personal beschäftigen, das ein vertieftes Know-how in den Fachbereichen der Statistik und Programmierung besitzt. Die Arbeit des Datenjournalismus verbindet somit die Bereiche IT und Journalismus in hohem Maße, und setzt eine enge Zusammenarbeit zwischen Programmierern und Journalisten voraus. Auch Interaction oder User Experience Designer müssen in diesen Prozess integriert werden, da sie maßgeblich an der Aufmachung und Visualisierung der datenjournalistischen Werke beteiligt sind.

5.6. Die Rolle des Journalisten

Der Einzug der neuen Erzählformen und Methoden in die Welt des Journalismus definiert in gewissem Maße die Nachrichtenindustrie neu. Aus diesem Grund ist es wichtig, dass angehende Journalisten auf diesen Umschwung vorbereitet werden, und das erforderliche technische und crossmediale Verständnis vermittelt bekommen. Was die Ausbildung neuer Journalisten betrifft, ist in Deutschland durchaus eine zukunftsgerichtete Haltung zu beobachten. Universitäten und Journalismus-Schulen reagieren bereits mit entsprechenden Veränderungen in ihren Rahmenbedingungen, und richten teils Labore ein, die eine bestmögliche crossmediale Ausbildung der Studenten ermöglichen sollen (Weichler, 2012), siehe beispielsweise im Leuphana Centre for Digital Cultures (2015) und in der Akademie für Publizistik Hamburg (2015). In diesem Zusammenhang wird in der Veröffentlichung von Kurt Weichler (Weichler, 2012) Ulrich Brenner, Leiter der Deutschen Journalistenschule in München zitiert:

> „Multimediales, crossmediales Denken zu beherrschen, ist unabdingbar für eine erfolgreiche journalistische Laufbahn – heute schon, aber erst recht in der Zukunft.“ Außerdem, so Ulrich Brenner: „Wenn wir ein Thema erarbeiten, kann die Frage nicht mehr sein: Was ist für dieses Thema die optimale Darstellungsform? Oder: Wie machen wir daraus eine hervorragende Reportage für die Seite 3 oder einen spannenden Radiobeitrag fürs

> Mittagsmagazin? Die Frage lautet jetzt: Was machen wir wie für welches Medium?“

Die Rolle der Journalisten beschränkt sich demnach nicht mehr ausschließlich auf die reine publizistische Tätigkeit, sondern sie befasst sich auch mit der Aufgabe, die Inhalte für die verschiedenen modernen Ausspielkanäle geeignet aufzubereiten. Im Feld des Datenjournalismus kommen zusätzlich die Aufgaben der Recherche und eingehenden Auseinandersetzung mit Datensätzen sowie dessen Analyse und Interpretation ins Spiel. Die journalistische Ausbildung muss sich deswegen neu ausrichten und vor allem die nötigen Kenntnisse zum Aggregieren, Analysieren, Filtern und Darstellen der Informationen vermitteln. Leßmöllmann (2012) beschreibt das Arbeitsfeld als „äußerst komplex, weswegen Datenjournalisten immer im Team arbeiten sollten“. Ein grundlegendes technisches Verständnis für die gemeinsamen Werkzeuge, die bei der Arbeit im Team zum Einsatz kommen, ist deswegen ebenfalls erforderlich, um die Zusammenarbeit mit Programmierern (und Designern) zu unterstützen.

5.7. Fazit

Immer mehr Redaktionen versuchen mit ihrer Berichterstattung auf die veränderten Nutzerinteressen und neuen technischen Möglichkeiten zu reagieren. Das Genre des Datenjournalismus, das Gebrauch von Open Data macht, um versteckte Zusammenhänge in Datensätzen aufzudecken, die dann meist auf interaktive Weise an den Nutzer herangebracht werden, ist eine Antwort auf diese veränderten Bedingungen.

Die Analyse und Interpretation der Daten ist dabei eine zentrale Aufgabe im Datenjournalismus. Demzufolge müssen Techniken und Werkzeuge der IT angewendet werden, um interessante Muster in den zugrundeliegenden Datensätzen zu entdecken, die als neues gewonnenes Wissen für die Berichterstattung verwendet werden können. Hier können neben Tabellenkalkulationsprogrammen auch Techniken des Data Mining zum Einsatz kommen, welche vom Anwender jedoch umfangreiche Kenntnis über das Gebiet der Wissensentdeckung in Datenbanken voraussetzen. Neben der Extraktion von Wissen aus Daten, besteht für die Datenjournalisten aber auch die Aufgabe, das gewonnene Wissen in einer verständlichen Weise zu visualisieren und idealerweise auch interaktiv aufzubereiten, sodass die Leser sich mit dem behandelten Thema des Berichts näher auseinandersetzen können. Die verschiedenen Tools, die für diese

Aufgabe bereits existieren und von Datenjournalisten eingesetzt werden, eignen sich allerdings wegen ihrer begrenzten Anpassungsmöglichkeiten nur eingeschränkt, und sind noch dazu in der Regel nicht integraler Bestandteil der Content Management Systeme, die typischerweise für die Ausspielung der redaktionellen Inhalte zuständig sind. Die Teams oder Redaktionen, die sich, vor allem langfristig gesehen, umfassend mit der Erstellung datenjournalistischer Inhalte befassen möchten, benötigen somit neben den eigentlichen Journalisten auch Personen mit vertieften Kenntnissen aus den Bereichen Statistik und Programmierung, sowie Designer, die sich mit dem gestalterischen und nutzungsorientierten Aspekt des Datenjournalismus beschäftigen.

Durch die Interaktionsmöglichkeiten, die dem Publikum heutzutage auf den verschiedenen Endgeräten und Ausspielkanälen zur Verfügung stehen, hat sich der Journalismus letztlich zu einem interaktiven Prozess geformt, der von multimedialer Berichterstattung geprägt ist. Die Experimente der Redaktionen mit den neuen interaktiven Erzählformen sind aktuell jedoch noch sehr kostspielig in Bezug auf die Arbeit und Zeit, die hierfür investiert wird. Die Formate mit denen Geschichten heute erzählt werden, benötigen somit aus technischer Sicht Lösungen, die sich gut in den Alltag der Redaktionen integrieren lassen und schließlich den Workflow verbessern. Zentral für den Erfolg des Datenjournalismus ist aber letztendlich auch, dass die Journalisten das nötige technische und crossmediale Verständnis besitzen. Das Handwerk und der Beruf des Journalisten befindet sich somit in einem Wandel, mit dem sich die Medienlandschaft auch in Zukunft immer mehr auseinandersetzen muss.

6. Schlussbetrachtung

6.1. Zusammenfassung

In dieser Arbeit wurde der gegenwärtige Stand der Entwicklung des Open Data-Konzeptes im öffentlichen Sektor sowie des Datenjournalismus untersucht. Dabei wurden aus der Sicht der Informationstechnologie die verschiedenen Teilaspekte betrachtet, die in dem gesamten Kontext zum Tragen kommen, angefangen bei der Bereitstellung von Open Data durch öffentliche Einrichtungen, dessen Beschaffung, Analyse und Interpretation bis hin zu der Verwendung der extrahierten Informationen in datenjournalistischen Werken.

Im Rahmen dieser Arbeit wurde zunächst das Konzept von Open Data nahegebracht, das für die heutige Informationsgesellschaft einen hohen Mehrwert bietet. Die Öffnung von Daten, die speziell im Bereich des öffentlichen Sektors erzeugt werden, ist nicht nur ein Schritt in Richtung mehr Transparenz, sondern sie verspricht auch vielfache Potenziale, die vor allem in der Wirtschaft zu positiven Effekten führen. Neben einer Erläuterung der Potenziale wurde auch ein Überblick zum Stand von Open Data auf nationaler sowie internationaler Ebene gegeben, aus dem zu erkennen ist, dass weltweit immer mehr Einrichtungen des öffentlichen Sektors ihre Daten als Open Data zur freien Weiterverwendung zur Verfügung stellen. Die am Ende des Kapitels zusammengefassten Herausforderungen bei der Bereitstellung von Open Data im öffentlichen Bereich haben jedoch gezeigt, dass das Open Data Konzept zum jetzigen Zeitpunkt noch nicht optimal umgesetzt wurde, und infolgedessen die Nutzung der Daten durch externe Akteure erschwert wird.

Im weiteren Verlauf der Arbeit wurde die technische Infrastruktur diskutiert, die benötigt wird, um Open Data geeignet bereitzustellen, sodass Journalisten aber auch sonstige externe Akteure auf die Daten zugreifen, und sie in beliebiger Weise für neue Informationsprodukte und -dienstleistungen weiterverwenden können. Dazu wurde

als Erstes die Rolle von Big Data für IT-gestützte Projekte sowie die Bedeutung der Wahl geeigneter Technologien zur effizienten Verwaltung und Verarbeitung von großen Datenbeständen vor Augen geführt, die zuletzt auch im Bereich von Open Government Data vermehrt anfallen. Die Technologien, die in diesem Zusammenhang erläutert wurden, können sich sowohl bei der Bereitstellung der Daten als auch bei dessen Aggregation und Analyse nutzbringend auswirken. Im Anschluss wurden die vermuteten Anforderungen skizziert, die Datenplattformen zur Bereitstellung von Open Data erfüllen müssen. Die bestehenden Softwarelösungen, die hierzu vorgestellt wurden und bereits länderübergreifend im Einsatz sind, erfüllen diese Anforderungen und bieten geeignete Zugriffsmöglichkeiten auf die veröffentlichten Daten. So existieren für den maschinellen Zugriff auf Open Data auch entsprechende standardisierte APIs, von denen Entwickler Gebrauch machen können, um die Daten in neuen Anwendungen weiter zu verwenden. Im Rahmen dieses Kapitels wurde zum Schluss auch das Linked Data Konzept vorgestellt, das sich für den interoperablen Austausch und Wiederverwertungsprozess von Daten bewährt hat. Die zugrunde liegenden Standards haben sich als praxistauglich erwiesen, und bieten zusätzlich Mechanismen zur Datenabfrage an, die bei der Nutzung durch externe Akteure zweckdienlich sind. Eine semantische Anreicherung und Vernetzung von Open Government Data steht in der Praxis jedoch noch vielfach aus, weswegen gegenwärtig nur sehr wenige Datensätze als Linked Data existieren. Abschließend kann aber zusammengefasst werden, dass es für Open Data und auch den Teilbereich Open Government Data mittlerweile eine gewisse Standardisierung in Bezug auf die Sichtweise und Dienstleistungen gibt, die im Zusammenhang mit der technischen Bereitstellung dieser Daten angeboten werden.

In dem darauf folgenden Kapitel wurde anschließend die Analyse und Interpretation von Datenbeständen mittels Verfahren des Data Mining behandelt, die für eine Extraktion neuer, nützlicher Informationen aus Open Data essenziell ist und besonders im Bereich des Datenjournalismus im Mittelpunkt steht. Zu Beginn des Kapitels wurde der Ablauf einer Datenanalyse beschrieben, aus dem die wesentliche Rolle der einzelnen Teilphasen hervorgeht. Die Schritte, die innerhalb dieser Phasen vorgenommen werden, entscheiden maßgeblich über die Ergebnisse der Datenanalyse und somit auch über die Qualität des gewonnenen Wissens. Hier besitzt besonders der Schritt der Vorverarbeitung einen hohen Stellenwert, da ein nicht ausreichend bereinigter Datensatz zu inkorrekten Ergebnissen führen kann, und damit das gewonnene Wissen unbrauchbar oder fehlerhaft macht. Die Data Mining Verfahren, die im Weiteren konkret vorgestellt wurden, haben sich in der Praxis in vielen Anwendungsbereichen als bewährt erwiesen,

ihre Anwendung setzt allerdings umfassende Kenntnisse in den Bereichen Statistik, Datenbanken und Informatik voraus, weswegen im Feld des Datenjournalismus die Durchführung des Data Mining-Prozesses in der Regel von Personen erfolgen sollte, die die erforderlichen Kompetenzen besitzen.

Im letzten Kapitel wurde zum Schluss ein Einblick in den Bereich des Datenjournalismus gegeben, der Open Data, aber auch sonstige Daten als Grundlage für die Erzählung einer Geschichte verwendet. Die Geschichten beruhen dabei grundlegend auf den Zusammenhängen bzw. Informationen, die aus der Analyse und Interpretation der Daten hervorgehen. Aus den verschiedenen beschriebenen Erzählformen und Methoden, die im Datenjournalismus angewendet werden, wird weiterhin erkennbar, dass die Rolle der IT eine nicht ungeachtete Bedeutung trägt. Nicht nur die Schritte der Datenanalyse und -auswertung, sondern auch die Aufgaben der Visualisierung und interaktiven Aufbereitung der gewonnenen Informationen erfordern die Anwendung passender technischer Werkzeuge. Die hierfür verfügbaren Werkzeuge eignen sich jedoch nur eingeschränkt, weswegen neben geeignetem Personal, das Design-, Statistik- und Programmierkompetenzen mitbringt, auch Lösungen benötigt werden, die hochgradig anpassbar sind, und den Workflow innerhalb der Redaktionen optimieren. Zuletzt ist auch ein technisches Verständnis der Journalisten für eine erfolgreiche Arbeit erforderlich.

Zusammenfassend lässt sich festhalten, dass zum gegenwärtigen Zeitpunkt sowohl das Open Data-Konzept als auch der Datenjournalismus eine gewisse ausgereifte Stufe erreicht haben, trotz dessen gibt es noch vereinzelt technische Herausforderungen in beiden Bereichen, die verbesserungswürdig sind. Ungeachtet der ausstehenden Herausforderungen, gibt es heute aber bereits einige Beispiele für guten und erfolgreichen Datenjournalismus, der mit Hilfe von Open Data versteckte Informationen aufdeckt und für die Öffentlichkeit greifbar macht.

6.2. Ausblick

Die vorliegende Arbeit hat sich mit der Verwendung von Open Data, die ausschließlich aus dem Bereich des öffentlichen Sektors stammen, als Ausgangsbasis für die Erzählung eines journalistischen Beitrags auseinandergesetzt. An dieser Stelle wäre es möglich, im Rahmen einer möglichen Arbeit zu untersuchen, wie sich Open Data sowohl aus dem öffentlichen als auch aus dem privatwirtschaftlichen Bereich überlagern

bzw. aggregieren lassen, um Fragestellungen zu beantworten, die im Rahmen eines datenjournalistischen Werkes an den Leser näher gebracht werden. Ferner gibt es auch Analysepotenziale, was den Umgang des Nutzers mit datenjournalistischen Infografiken und ihrer narrativen Kraft anbetrifft. Wie kann hier die Informationstechnologie die Aufgabe der Aufmachung und Visualisierung von Daten unterstützen, und zusätzlich etwa den persönlichen Kontext des Nutzers (aktuelle Positionsdaten, Tageszeit etc.) in diese mit einbeziehen?

Eine weitere mögliche Untersuchung könnte auch darin bestehen, wie die primären technischen Arbeitswerkzeuge der Journalisten sich in den Rahmen der datenjournalistischen Arbeit integrieren. Wie können moderne Content Management Systeme den Workflow der Redaktionen vereinfachen, etwa wenn es darum geht, Open Data oder sonstige Datensätze für einen Beitrag zu aggregieren, zu analysieren und weiterhin aufzubereiten und geeignet auszuspielen? Wie kann hier die Zusammenarbeit der am gesamten Prozess beteiligten Personen wie Programmierern, (Interaction bzw. User Experience) Designern und Datenjournalisten mit Hilfe technischer Mittel gefördert werden?

Im Rahmen der vorliegenden Arbeit wurden zu guter Letzt die Herausforderungen zusammengeführt, die im Augenblick bei der Bereitstellung von Open Data des öffentlichen Sektors und dessen anschließende Verwendung im Datenjournalismus ausstehen. Der Fokus der Untersuchung lag dabei auf der technischen Infrastruktur, die hier zum Einsatz kommt. Basierend auf den gewonnenen Erkenntnissen können nun verschiedenste weiterführende Arbeiten entstehen.

A. Anhang

A.1. API-Beispiel für den Abruf eines Datensatzes in CKAN


```
{
   "help":"Return the metadata of a dataset (package) and its
      resources.\n\n    :param id: the id or name of the dataset\
      n    :type id: string\n    :param use_default_schema: use
      default package schema instead of\n        a custom schema
      defined with an IDatasetForm plugin (default: False)\n    :
      type use_default_schema: bool\n\n    :rtype: dictionary\n\n
          ",
   "success":true,
   "result":{
      "license_title":"License Not Specified",
      "maintainer":"",
      "relationships_as_object":[

      ],
      "private":false,
      "maintainer_email":"",
      "revision_timestamp":"2014-04-21T10:40:23.916113",
      "id":"015e0233-1f01-4439-98b5-cc9c5170282e",
      "metadata_created":"2014-04-21T08:01:41.503118",
      "owner_org":null,
      "metadata_modified":"2014-10-09T07:18:40.174476",
      "author":"Lucy Chambers",
      "author_email":"",
      "state":"active",
      "version":"",
      "license_id":"notspecified",
      "type":"dataset",
      "resources":[
```

```
        {
            "resource_group_id":"f444e121-136e-4bdd-b900-3
               dca465dfa0e",
            "cache_last_updated":null,
            "revision_timestamp":"2014-05-08T10:09:58.946797",
            "webstore_last_updated":"2014-04-21T08:01:55.219712",
            "id":"04127ad5-77e5-4a08-9f40-12d3c383e460",
            "size":null,
            "state":"active",
            "hash":"1868d87eb6e3c26238a0d42e46f63f06",
            "description":"Revised CSV for import",
            "format":"CSV",
            "tracking_summary":{
               "total":0,
               "recent":0
            },
            "mimetype_inner":"",
            "url_type":null,
            "openspending_hint":"data",
            "mimetype":"",
            "cache_url":null,
            "name":"",
            "created":"2014-04-21T08:01:41.579322",
            "url":"http://mk.ucant.org/info/data/adur.csv",
            "webstore_url":"active",
            "last_modified":null,
            "position":0,
            "revision_id":"bf3d8ca4-30c6-410d-960d-9fdf56fc93d0",
            "resource_type":"file"
        },
        {
            "resource_group_id":"f444e121-136e-4bdd-b900-3
               dca465dfa0e",
            "cache_last_updated":null,
            "revision_timestamp":"2014-05-08T10:09:58.946797",
            "webstore_last_updated":"2014-05-06T03:10:48.602339",
            "id":"281dffa6-ea9b-4446-be41-05dced06591f",
            "size":null,
            "state":"active",
            "hash":"b19731f6a08c98079adfaeffadb719f5",
```

```
            "description":"Adur District Council April 2009",
            "format":"CSV",
            "tracking_summary":{
               "total":0,
               "recent":0
            },
            "mimetype_inner":"",
            "url_type":null,
            "mimetype":"",
            "cache_url":null,
            "name":"",
            "created":"2014-04-21T08:01:41.579275",
            "url":"http://ckan.net/storage/f/file/3ffdcd42-5c63
               -4089-84dd-c23876259973",
            "webstore_url":"active",
            "last_modified":null,
            "position":1,
            "revision_id":"bf3d8ca4-30c6-410d-960d-9fdf56fc93d0",
            "resource_type":"file"
         },
         {
            "resource_group_id":"f444e121-136e-4bdd-b900-3
               dca465dfa0e",
            "cache_last_updated":null,
            "revision_timestamp":"2014-05-08T10:09:58.946797",
            "webstore_last_updated":null,
            "id":"6cce3936-b169-4d12-82ba-65fcb79734a0",
            "size":null,
            "state":"active",
            "hash":"",
            "description":"Mapping Metadata for Adur",
            "format":"JSON",
            "tracking_summary":{
               "total":0,
               "recent":0
            },
            "mimetype_inner":"",
            "url_type":null,
            "openspending_hint":"model",
            "mimetype":"",
```

```
            "cache_url":null,
            "name":"",
            "created":"2014-04-21T08:01:41.579310",
            "url":"http://ckan.net/storage/f/file/c8ce520c-c2e6
               -463a-99a3-ad24b023ccb4",
            "webstore_url":null,
            "last_modified":null,
            "position":2,
            "revision_id":"bf3d8ca4-30c6-410d-960d-9fdf56fc93d0",
            "resource_type":"file"
         },
         {
            "resource_group_id":"f444e121-136e-4bdd-b900-3
               dca465dfa0e",
            "cache_last_updated":null,
            "revision_timestamp":"2014-07-28T20:09:04.864795",
            "webstore_last_updated":null,
            "id":"f1d75078-e024-4540-84e2-e30bdc4b8be0",
            "size":null,
            "state":"active",
            "hash":"",
            "description":"description",
            "format":"HTML",
            "tracking_summary":{
               "total":0,
               "recent":0
            },
            "mimetype_inner":null,
            "url_type":null,
            "mimetype":null,
            "cache_url":null,
            "name":"example",
            "created":"2014-07-28T20:09:04.923894",
            "url":"http://dot.com",
            "webstore_url":null,
            "last_modified":null,
            "position":3,
            "revision_id":"f09776c5-28ca-43a2-ba3a-f754141c1e70",
            "resource_type":null
         }
```

```
        ],
        "num_resources":4,
        "tags":[
            {
                "vocabulary_id":null,
                "display_name":"country-uk",
                "name":"country-uk",
                "revision_timestamp":"2014-04-21T08:01:41.503118",
                "state":"active",
                "id":"00b03026-dfbd-4d5c-87b8-ee7a087052cb"
            },
            {
                "vocabulary_id":null,
                "display_name":"date-2009",
                "name":"date-2009",
                "revision_timestamp":"2014-04-21T08:01:41.503118",
                "state":"active",
                "id":"1e5709f3-8d4b-40ef-987f-b4f292dce0cb"
            },
            {
                "vocabulary_id":null,
                "display_name":"openspending",
                "name":"openspending",
                "revision_timestamp":"2014-04-21T08:01:41.503118",
                "state":"active",
                "id":"b880e952-fb9f-486f-bfbc-286b6162ffd9"
            },
            {
                "vocabulary_id":null,
                "display_name":"regional",
                "name":"regional",
                "revision_timestamp":"2014-04-21T08:01:41.503118",
                "state":"active",
                "id":"39594190-41eb-49f2-b804-e42894823c9b"
            }
        ],
        "tracking_summary":{
            "total":0,
            "recent":0
        },
```

```
      "groups":[
         {
            "display_name":"Data Explorer Examples",
            "description":"This group contains various real
               datasets that show CKAN's data previewer in action
               . The previewer shows a configurable grid view of
               tabular data, plots columns of data on a graph,
               and shows geo-coded data on an interactive map. It
                can also preview image files and web pages.",
            "title":"Data Explorer Examples",
            "image_display_url":"http://farm8.staticflickr.com
               /7129/7041988029_411d985015_c.jpg",
            "id":"be9ff477-9547-4dec-85b3-ced05068775e",
            "name":"data-explorer"
         }
      ],
      "creator_user_id":"ba4c807a-1402-4117-a7cc-b308b01774af",
      "relationships_as_subject":[

      ],
      "num_tags":4,
      "name":"adur_district_spending",
      "isopen":false,
      "url":"http://www.spotlightonspend.org.uk/Downloads/1038",
      "notes":"From Spikes Cavell, Spotlight on Spend. For Ardur,
          records from April 2009-March 2010 are currently
         available (2011-008-04) ",
      "title":"UK: Adur District Council Spending Data",
      "extras":[
         {
            "value":"{ \"type\": \"Polygon\", \"coordinates\": [
               [ [-0.3715, 50.8168],[-0.3715, 50.8747], [-0.2155,
                50.8747], [-0.2155, 50.8168], [-0.3715, 50.8168]
               ] ] }",
            "key":"spatial"
         },
         {
            "value":"Adur, West Sussex, South East England,
               England, United Kingdom",
            "key":"spatial-text"
```

```
206             }
207         ],
208         "organization":null,
209         "revision_id":"62467531-c0c5-4620-b841-cc631babf382"
210     }
211 }
```

Listing A.1: JSON-Repräsentation eines Datensatzes aus CKAN (siehe hierzu auch Abschnitt 3.3.3.2).

Abbildungsverzeichnis

2.1. Wissenstreppe nach North (2011, S. 36). 8

3.1. Die Architektur von CKAN nach der Open Knowledge Foundation (2012a). 30

3.2. Graph-basiertes Datenmodell des Linked Data Ansatzes in Anlehnung an Barnickel und Klessmann (2012), S. 143. 37

3.3. Das 5-Sterne Bewertungssystem für Open Data nach Berners-Lee (2009). 40

4.1. Der Data Mining-Prozess nach Fayyad u. a. (1996). 47

5.1. Visualisierung „Verräterisches Handy“ auf zeit.de auf Basis der Mobilfunk-Vorratsdaten des Politikers Malte Spitz (ZEIT ONLINE, 2011). 66

Literaturverzeichnis

[Akademie für Publizistik Hamburg 2015] AKADEMIE FÜR PUBLIZISTIK HAMBURG: *Visuelle Publizistik | Visual Journalism*. http://www.visuelle-publizistik.de/. 2015. – abgerufen am 06.01.2015

[Alpar und Niedereichholz 2000] ALPAR, Paul ; NIEDEREICHHOLZ, Joachim: Einführung zu Data Mining. In: ALPAR, Paul (Hrsg.) ; NIEDEREICHHOLZ, Joachim (Hrsg.): *Data Mining im praktischen Einsatz*. Vieweg+Teubner Verlag, 2000 (Business Computing), S. 1–27. – URL http://dx.doi.org/10.1007/978-3-322-89950-7_1. – ISBN 978-3-528-05748-0

[Apache Software Foundation 2014] APACHE SOFTWARE FOUNDATION: *Apache Solr*. http://lucene.apache.org/solr/. 2014. – abgerufen am 17.11.2014

[Barnickel und Klessmann 2012] BARNICKEL, Nils ; KLESSMANN, Jens: Open Data – Am Beispiel von Informationen des öffentlichen Sektors. In: HERB, Ulrich (Hrsg.): *Open Initiatives: Offenheit in der digitalen Welt und Wissenschaft*. universaar, 2012, S. 127–158

[Bauer und Kaltenböck 2012] BAUER, Florian ; KALTENBÖCK, Martin: *Linked Open Data: The Essentials*. Vienna, Austria : Edition Mono/Monochrom, 2012

[Behörde für Justiz und Gleichstellung der Freien und Hansestadt Hamburg 2012] BEHÖRDE FÜR JUSTIZ UND GLEICHSTELLUNG DER FREIEN UND HANSESTADT HAMBURG: Hamburgisches Gesetz- und Verordnungsblatt Nr. 29 vom 6. Juli 2012. (2012). – URL http://www.hamburg.de/contentblob/3625198/data/hmbgtg.pdf. – abgerufen am 12.01.2015

[Berners-Lee 2009] BERNERS-LEE, Tim: *Linked Data - Design Issues*. http://www.w3.org/DesignIssues/LinkedData.html. 2009. – abgerufen am 17.11.2014

[Berners-Lee u. a. 2001] BERNERS-LEE, Tim ; HENDLER, James ; LASSILA, Ora: The Semantic Web. In: *Scientific American* 284 (2001), Nr. 5, S. 34–43

[Bertini und Lalanne 2010] BERTINI, Enrico ; LALANNE, Denis: Investigating and Reflecting on the Integration of Automatic Data Analysis and Visualization in Knowledge Discovery. In: *SIGKDD Explor. Newsl.* 11 (2010), Mai, Nr. 2, S. 9–18. – URL http://doi.acm.org/10.1145/1809400.1809404. – ISSN 1931-0145

[Bizer u. a. 2009] BIZER, Christian ; HEATH, Tom ; BERNERS-LEE, Tim: Linked Data - the story so far. In: *International Journal on Semantic Web and Information Systems* 5 (2009), Nr. 3, S. 1–22

[Blumauer u. a. 2011] BLUMAUER, Andreas ; HÖCHTL, Johann ; KALTENBÖCK, Martin ; KRABINA, Bernhard ; PARYCEK, Peter ; PELLEGRINI, Tassilo ; SCHOSSBÖCK, Judith ; THURNER, Thomas ; KALTENBÖCK, Martin (Hrsg.) ; THURNER, Thomas (Hrsg.): *Open Government Data Weißbuch (Österreich)*. Ed. Donau-Universität Krems, 2011

[Borthakur u. a. 2011] BORTHAKUR, Dhruba ; GRAY, Jonathan ; SARMA, Joydeep S. ; MUTHUKKARUPPAN, Kannan ; SPIEGELBERG, Nicolas ; KUANG, Hairong ; RANGANATHAN, Karthik ; MOLKOV, Dmytro ; MENON, Aravind ; RASH, Samuel ; SCHMIDT, Rodrigo ; AIYER, Amitanand: Apache Hadoop Goes Realtime at Facebook. In: *Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data*. New York, NY, USA : ACM, 2011 (SIGMOD '11), S. 1071–1080. – URL http://doi.acm.org/10.1145/1989323.1989438. – ISBN 978-1-4503-0661-4

[Bunz 2012] BUNZ, Mercedes: *Die stille Revolution: wie Algorithmen Wissen, Arbeit, Öffentlichkeit und Politik verändern, ohne dabei viel Lärm zu machen*. Suhrkamp, 2012

[Busemann und Tippelt 2014] BUSEMANN, Katrin ; TIPPELT, Florian: Second Screen: Parallelnutzung von Fernsehen und Internet. Ergebnisse der ARD/ZDF-Onlinestudie 2014. In: *Media Perspektiven 7-8/2014* (2014), S. 408–416

[Chen u. a. 2014] CHEN, Min ; MAO, Shiwen ; LIU, Yunhao: Big Data: A Survey. In: *Mobile Networks and Applications* 19 (2014), Nr. 2, S. 171–209. – URL http://dx.doi.org/10.1007/s11036-013-0489-0. – ISSN 1383-469X

[Cleve und Lämmel 2014] CLEVE, Jürgen ; LÄMMEL, Uwe: *Data Mining*. De Gruyter Oldenbourg, 2014

[Cleveland 1993] CLEVELAND, William S.: *Visualizing Data*. Hobart Press, 1993. – ISBN 0963488406

[CTIC 2015] CTIC: *Open Data @ CTIC » Sandbox » - Public Dataset Catalogs Faceted Browser*. http://datos.fundacionctic.org/sandbox/catalog/faceted/. 2015. – abgerufen am 02.02.2015

[data.gov.uk 2014] DATA.GOV.UK: *About | data.gov.uk*. http://data.gov.uk/about. 2014. – abgerufen am 01.12.2014

[daten.bremen.de 2015] DATEN.BREMEN.DE: *Daten.Bremen - Offene Daten*. http://www.daten.bremen.de/sixcms/detail.php?gsid=bremen02.c.734.de. 2015. – abgerufen am 12.01.2015

[Dekkers u. a. 2006] DEKKERS, Makx ; POLMAN, Femke ; VELDE, Robbin te ; VRIES, Marc de ; EUROPÄISCHE KOMMISSION (Hrsg.): *MEPSIR – Measuring European Public Sector Information Resources, Final Report of Study on Exploitation of public sector information – benchmarking of EU framework conditions*. 2006

[Deutsche Nationalbibliothek 2014] DEUTSCHE NATIONALBIBLIOTHEK: *Deutsche Nationalbibliothek - Linked Data Service - Linked Data Service der Deutschen Nationalbibliothek*. http://www.dnb.de/DE/Service/DigitaleDienste/LinkedData/linkeddata.html. 2014. – abgerufen am 25.11.2014

[Dietrich 2011a] DIETRICH, Daniel: *bpb.de - Open Data - Offene Daten In Deutschland*. http://www.bpb.de/gesellschaft/medien/opendata/64061/offene-daten-in-deutschland. 2011. – abgerufen am 06.01.2015

[Dietrich 2011b] DIETRICH, Daniel: *bpb.de - Open Data - Was sind offene Daten?* http://www.bpb.de/gesellschaft/medien/opendata/64055/was-sind-offene-daten. 2011. – abgerufen am 08.08.2014

[Dumbill 2012] 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

[van Eimeren und Frees 2014] EIMEREN, Birgit van ; FREES, Beate: 79 Prozent der Deutschen online – Zuwachs bei mobiler Internetnutzung und Bewegtbild. Ergebnisse der ARD/ZDF-Onlinestudie 2014. In: *Media Perspektiven 7-8/2014* (2014), S. 378–396

[Eimsbütteler Nachrichten 2015] EIMSBÜTTELER NACHRICHTEN: *Unsere Redaktion | Eimsbütteler Nachrichten*. http://www.eimsbuetteler-nachrichten.de/unsere-redaktion/. 2015. – abgerufen am 18.01.2015

[Elmer und Wormer 2014] ELMER, Christina ; WORMER, Holger: *Datenjournalismus – was ist das? | blog14*. http://netzwerkrecherche.org/wordpress/blog14/datenjournalismus-was-ist-das/. 2014. – abgerufen am 16.01.2015

[EUDAT 2014] EUDAT: *What is EUDAT? | EUDAT*. http://www.eudat.eu/what-eudat. 2014. – abgerufen am 07.11.2014

[Europäische Kommission 2011] EUROPÄISCHE KOMMISSION: *Mitteilung der Kommission an das Europäische Parlament, den Rat, den Europäischen Wirtschafts- und Sozialausschuss und den Ausschuss der Regionen – Offene Daten: Ein Motor für Innovation, Wachstum und transparente Verwaltung*. 2011

[Fayyad u. a. 1996] FAYYAD, Usama M. ; PIATETSKY-SHAPIRO, Gregory ; SMYTH, Padhraic: Advances in Knowledge Discovery and Data Mining. Menlo Park, CA, USA : American Association for Artificial Intelligence, 1996, Kap. From Data Mining to Knowledge Discovery: An Overview, S. 1–34. – URL http://dl.acm.org/citation.cfm?id=257938.257942. – ISBN 0-262-56097-6

[Fornefeld u. a. 2009] FORNEFELD, Martin ; BOELE-KEIMER, Gaby ; RECHER, Stephan ; FANNING, Michael ; MICUS MANAGEMENT CONSULTING GMBH (Hrsg.): *Assessment of the Re-use of Public Sector Information (PSI) in the Geographical Information, Meteorological Information and Legal Information Sectors*. 2009

[Forsterleitner und Gegenhuber 2011] FORSTERLEITNER, Christian ; GEGENHUBER, Thomas: Lasst die Daten frei! Open Government als kommunale Herausforderung und Chance. In: DOBUSCH, Leonhard (Hrsg.) ; FORSTERLEITNER, Christian (Hrsg.) ; HIESMAIR, Manuela (Hrsg.): *Freiheit vor Ort: Handbuch kommunale Netzpolitik*. Open Source Press, 2011, S. 233–266

[Frankfurt Gestalten 2015] FRANKFURT GESTALTEN: *Frankfurt gestalten - Bürger machen Stadt*. http://www.frankfurt-gestalten.de/. 2015. – abgerufen am 18.01.2015

[Gartner 2011] GARTNER: *Gartner Says Solving 'Big Data' Challenge Involves More Than Just Managing Volumes of Data*. http://www.gartner.com/newsroom/id/1731916. 2011. – abgerufen am 13.12.2014

[Gens 2011] GENS, Frank: IDC Predictions 2012: Competing for 2020, IDC, 2011. – URL http://cdn.idc.com/research/Predictions12/Main/downloads/IDCTOP10Predictions2012.pdf

[Götzer u. a. 2004] GÖTZER, Klaus ; SCHNEIDERATH, Udo ; MAIER, Berthold ; KOMKE, Torsten: *Dokumenten-Management*. 3., vollständig überarb. und erw. Auflage. Heidelberg: dpunkt, 2004

[Hagen u. a. 2005] HAGEN, H ; STEINEBACH, G ; MÜNCHHOFEN, M ; RUBY, M ; SCHELER, I ; WADLÉ, M ; MICHEL, F: Datenmanagementsystem für die Stadtplanung. In: *CORP 2005, Wien* (2005), S. 663–669

[IETF 1994] IETF: *Universal Resource Identifiers in WWW: A Unifying Syntax for the Expression of Names and Addresses of Objects on the Network as used in the World-Wide Web, RFC 1630*. https://datatracker.ietf.org/doc/rfc1630/. 1994. – abgerufen am 25.11.2014

[IETF 1999] IETF: *Hypertext Transfer Protocol – HTTP/1.1, RFC 2616*. https://datatracker.ietf.org/doc/rfc2616/. 1999. – abgerufen am 25.11.2014

[Intel IT Center 2013] INTEL IT CENTER: *Planning Guide: Getting Started with Big Data: Steps IT Managers Can Take to Move Forward with Apache Hadoop Software*. Intel Corporation, 2013

[ITU 2014] ITU: *Number of worldwide internet users from 2000 to 2014 (in millions)*. http://www.statista.com/statistics/273018/number-of-internet-users-worldwide/. 2014. – abgerufen am 20.10.2014

[Klein u. a. 2013] KLEIN, Dominik ; TRAN-GIA, Phuoc ; HARTMANN, Matthias: Big Data. In: *Informatik-Spektrum* 36 (2013), Nr. 3, S. 319–323. – URL http://dx.doi.org/10.1007/s00287-013-0702-3. – ISSN 0170-6012

[Klessmann u. a. 2012] KLESSMANN, Jens ; DENKER, Philipp ; SCHIEFERDECKER, Ina ; SCHULZ, Sönke E: *Open Government Data Deutschland – Eine Studie zu Open Government in Deutschland im Auftrag des Bundesministerium des Innern*. Bundesministerium des Innern, 2012

[Kubicek 2008] KUBICEK, Herbert: Next Generation FoI Between Information Management and Web 2.0. In: *Proceedings of the 2008 International Conference on Digital Government Research*, Digital Government Society of North America, 2008

(dg.o '08), S. 9–16. – URL http://dl.acm.org/citation.cfm?id=1367832.1367838. – ISBN 978-1-60558-099-9

[Laney 2001] LANEY, Doug: *3D Data Management: Controlling Data Volume, Velocity, and Variety*. http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf. 2001. – abgerufen am 13.12.2014

[Leßmöllmann 2012] LESSMÖLLMANN, Annette: *Datenjournalismus: Chance für den Journalismus von morgen | Journalistik Journal*. http://journalistik-journal.lookingintomedia.com/?p=843. 2012. – abgerufen am 06.01.2015

[Leuphana Centre for Digital Cultures 2015] LEUPHANA CENTRE FOR DIGITAL CULTURES: *CDC: Gamification Lab*. http://cdc.leuphana.com/structure/gamification-lab/. 2015. – abgerufen am 06.01.2015

[Lorenz 2013] LORENZ, Mirko: *Was ist Datenjournalismus? (with image) · mirkolorenz · Storify*. https://storify.com/mirkolorenz/datenjournalismus. 2013. – abgerufen am 16.01.2015

[von Lucke 2011] 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, S. 261–272

[von Lucke und Geiger 2010] 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) / Deutsche Telekom Institute for Connected Cities, Zeppelin University Friedrichshafen. 2010. – Forschungsbericht

[Matzat 2011] MATZAT, Lorenz: *bpb.de - Open Data - Datenjournalismus*. http://www.bpb.de/gesellschaft/medien/opendata/64069/datenjournalismus. 2011. – abgerufen am 06.01.2015

[Matzat 2014] MATZAT, Lorenz: *Datenjournalismus: Methoden, Prozesse und Kompetenzen - Fachjournalist*. http://www.fachjournalist.de/datenjournalismus-methoden-prozesse-und-kompetenzen/. 2014. – abgerufen am 06.01.2015

[Mills u. a. 2012] MILLS, Steve ; LUCAS, Steve ; IRAKLIOTIS, Leo ; RAPPA, Michael ; CARLSON, Teresa ; PERLOWITZ, Bill: Demystifying Big Data: A practical guide to transforming the business of government / Tech America Foundation. 2012. – Forschungsbericht

[Müller 2013] MÜLLER, Roland M. ; LENZ, Hans-Joachim (Hrsg.): *Business Intelligence*. Berlin, Heidelberg ;s.l. : Springer Berlin Heidelberg, 2013. – Online–Ressource (XXII, 306 S. 198 Abb) S

[Niedersächsisches Ministerium für Umwelt, Energie und Klimaschutz 2015] NIEDERSÄCHSISCHES MINISTERIUM FÜR UMWELT, ENERGIE UND KLIMASCHUTZ: *NUMIS - Alle Funktionen auf einen Blick*. http://numis.niedersachsen.de/funktionen. 2015. – abgerufen am 12.01.2015

[North 2011] NORTH, Klaus: *Wissensorientierte Unternehmensführung: Wertschöpfung durch Wissen*. 5. Auflage. Gabler, 2011

[Obama 2009] OBAMA, Barack: *Memorandum for the Heads of Executive Departments and Agencies, Subject: Transparency and Open Government*. http://www.whitehouse.gov/the_press_office/TransparencyandOpenGovernment/. 2009. – abgerufen am 27.10.2014

[Open Data Network 2010] OPEN DATA NETWORK: *Data Driven Journalism: Versuch einer Definition*. http://opendata-network.org/2010/04/data-driven-journalism-versuch-einer-definition/. 2010. – abgerufen am 16.01.2015

[Open Knowledge Foundation 2012a] OPEN KNOWLEDGE FOUNDATION: *CKAN Data Hub*. https://commondatastorage.googleapis.com/ckannet-storage/2012-02-13T201110/CKAN_Overview.pdf. 2012. – abgerufen am 17.11.2014

[Open Knowledge Foundation 2012b] OPEN KNOWLEDGE FOUNDATION: *Dataset — CKAN Data Management System Documentation 1.8 documentation*. http://docs.ckan.org/en/ckan-1.8/domain-model-dataset.html. 2012. – abgerufen am 17.11.2014

[Open Knowledge Foundation 2012c] OPEN KNOWLEDGE FOUNDATION: *Import ("Harvest") Data from Other Sites — CKAN Data Management System Documentation 1.7.3 documentation*. http://docs.ckan.org/en/ckan-1.7.3/harvesting.html. 2012. – abgerufen am 17.11.2014

[Open Knowledge Foundation 2013] OPEN KNOWLEDGE FOUNDATION: *Adding custom previews to CKAN | ckan - The open source data portal software*. http://ckan.org/2013/03/13/custom-previews/. 2013. – abgerufen am 17.11.2014

[Open Knowledge Foundation 2014a] OPEN KNOWLEDGE FOUNDATION: *About - PublicData.eu*. http://publicdata.eu/about. 2014. – abgerufen am 17.11.2014

[Open Knowledge Foundation 2014b] OPEN KNOWLEDGE FOUNDATION: *About CKAN | ckan - The open source data portal software*. http://ckan.org/developers/about-ckan/. 2014. – abgerufen am 17.11.2014

[Open Knowledge Foundation 2014c] OPEN KNOWLEDGE FOUNDATION: *API guide — CKAN 2.3a documentation*. http://docs.ckan.org/en/latest/api/index.html. 2014. – abgerufen am 17.11.2014

[Open Knowledge Foundation 2014d] OPEN KNOWLEDGE FOUNDATION: *CKAN instances around the world | ckan - The open source data portal software*. http://ckan.org/instances/. 2014. – abgerufen am 17.11.2014

[Open Knowledge Foundation 2014e] OPEN KNOWLEDGE FOUNDATION: *Data Viewer — CKAN 2.3a documentation*. http://docs.ckan.org/en/latest/maintaining/data-viewer.html. 2014. – abgerufen am 17.11.2014

[Open Knowledge Foundation 2014f] OPEN KNOWLEDGE FOUNDATION: *DataStore extension — CKAN 2.3a documentation*. http://docs.ckan.org/en/latest/maintaining/datastore.html. 2014. – abgerufen am 17.11.2014

[Open Knowledge Foundation 2014g] OPEN KNOWLEDGE FOUNDATION: *Definition: Offenes Wissen - Open Definition - Defining Open in Open Data, Open Content and Open Knowledge*. http://opendefinition.org/od/1.1/de/. 2014. – abgerufen am 07.11.2014

[Open Knowledge Foundation 2014h] OPEN KNOWLEDGE FOUNDATION: *User guide — CKAN 2.3a documentation*. http://docs.ckan.org/en/latest/user-guide.html. 2014. – abgerufen am 17.11.2014

[Open Knowledge Foundation Deutschland 2014] OPEN KNOWLEDGE FOUNDATION DEUTSCHLAND: *Offene Daten | Open Knowledge Foundation Deutschland - Förderung von offenem Wissen im digitalen Zeitalter*. http://okfn.de/opendata/. 2014. – abgerufen am 07.11.2014

[Open Source Initiative 2015] OPEN SOURCE INITIATIVE: *Welcome to The Open Source Initiative | Open Source Initiative*. http://opensource.org/. 2015. – abgerufen am 06.01.2015

[OpenDataCity 2015] OPENDATACITY: *OpenDataCity Wir machen Daten greifbar - OpenDataCity*. https://opendatacity.de/. 2015. – abgerufen am 16.01.2015

[Ostheimer und Janz 2005] OSTHEIMER, Bernhard ; JANZ, Wolfhard: *Dokumenten-Management-Systeme : Abgrenzung, Wirtschaftlichkeit, rechtliche Aspekte*. 2005. – URL http://geb.uni-giessen.de/geb/volltexte/2005/2430

[Pollock 2009] POLLOCK, Rufus: The Economics of Public Sector Information / Faculty of Economics, University of Cambridge. URL http://EconPapers.repec.org/RePEc:cam:camdae:0920, 2009. – Cambridge Working Papers in Economics

[Probst u. a. 2012] PROBST, Gilbert ; RAUB, Steffen ; ROMHARDT, Kai: *Wissen managen: Wie Unternehmen ihre wertvollste Ressource optimal nutzen*. 7. Auflage. Gabler, 2012

[Right2Info.org 2013] RIGHT2INFO.ORG: *Access to Information Laws: Overview and Statutory Goals — Right2Info.org*. http://right2info.org/access-to-information-laws. 2013. – abgerufen am 20.12.2014

[Rogers 2012] ROGERS, Simon: *Change – Wie Daten den Journalismus verändern | Diskurs @ Deutschland*. http://diskurs.dradio.de/2012/02/22/change-wie-daten-den-journalismus-verandern/. 2012. – abgerufen am 18.01.2015

[Rogers 2015] ROGERS, Simon: *Behind the Scenes at the Guardian Datablog - The Data Journalism Handbook*. http://datajournalismhandbook.org/1.0/en/in_the_newsroom_3.html. 2015. – abgerufen am 18.01.2015

[Senat der Freien und Hansestadt Hamburg 2012] SENAT DER FREIEN UND HANSESTADT HAMBURG: Stellungnahme des Senats zu dem Ersuchen der Bürgerschaft vom 13. Juni 2012 „Erlass eines Hamburgischen Transparenzgesetzes“ (Ziffer 2 der Drucksache 20/4466). (2012). – URL http://www.hamburg.de/contentblob/3906986/data/berichterstattung-hmbtg-2013-04.pdf. – abgerufen am 12.01.2015

[Senatsverwaltung für Stadtentwicklung Berlin 2003] SENATSVERWALTUNG FÜR STADTENTWICKLUNG BERLIN: *Monitoring Soziale Stadtentwicklung 2004*. http://www.stadtentwicklung.berlin.de/planen/basisdaten_stadtentwicklung/monitoring/download/2004/endbericht_moni2004.pdf. 2003. – abgerufen am 20.12.2014

[Socrata 2014] SOCRATA: *Customer Spotlights World's Most Innovative Open Data Users*. http://www.socrata.com/customer-spotlight/. 2014. – abgerufen am 17.11.2014

[Statistisches Bundesamt 2015] STATISTISCHES BUNDESAMT: *Unsere Aufgaben - Statistisches Bundesamt (Destatis)*. https://www.destatis.de/DE/UeberUns/UnsereAufgaben/Aufgaben.html. 2015. – abgerufen am 12.01.2015

[Sturm 2013] STURM, Simon: *Digitales Storytelling Eine Einführung in neue Formen des Qualitätsjournalismus*. Springer, 2013

[Süddeutsche Zeitung 2015] SÜDDEUTSCHE ZEITUNG: *Bahn-Verspätungen - Wie unpünktlich der deutsche Fernverkehr ist - mit Zugmonitor - Süddeutsche.de*. http://www.sueddeutsche.de/thema/Bahn-Versp%C3%A4tungen. 2015. – abgerufen am 18.01.2015

[The Language Archive 2014] THE LANGUAGE ARCHIVE: *CKAN tested on 2 million datasets - The Language Archive*. https://tla.mpi.nl/tla-news/ckan-tested-2-million-datasets/. 2014. – abgerufen am 07.11.2014

[Twitter, Inc. 2012] TWITTER, INC.: *Visualizing Hadoop with HDFS-DU | Twitter Blogs*. https://blog.twitter.com/2012/visualizing-hadoop-with-hdfs-du. 2012. – abgerufen am 12.01.2015

[W3C 2004] W3C: *OWL Web Ontology Language Overview*. http://www.w3.org/TR/owl-features/. 2004. – abgerufen am 25.11.2014

[W3C 2008] W3C: *SPARQL Query Language for RDF*. http://www.w3.org/TR/rdf-sparql-query/. 2008. – abgerufen am 25.11.2014

[W3C 2014a] W3C: *RDF Current Status - W3C*. http://www.w3.org/standards/techs/rdf/. 2014. – abgerufen am 25.11.2014

[W3C 2014b] W3C: *RDF Schema 1.1*. http://www.w3.org/TR/rdf-schema/. 2014. – abgerufen am 25.11.2014

[Weichler 2012] WEICHLER, Kurt: Warum der Journalismus derzeit an Wert verliert. In: *Signsbook – Zeichen setzen in der Kommunikation*. Wiesbaden : Anda, Béla and Endrös, Stefan and Kalka, Jochen and Lobo, Sascha, 2012, S. 13–21

[Woytewicz 2013] WOYTEWICZ, Daniela: *Mit Daten Geschichten erzählen: Das Potential von Datenjournalismus im World Wide Web*. 2013

[Wrobel u. a. 2014] WROBEL, Stefan ; JOACHIMS, Thorsten ; MORIK, Katharina: Maschinelles Lernen und Data Mining. In: *Handbuch der Künstlichen Intelligenz*. 5. Auflage. Görz, Günther and Schneeberger, Josef and Schmid, Ute, 2014, S. 405–471

[ZEIT ONLINE 2011] ZEIT ONLINE: *Vorratsdatenspeicherung - interaktive Grafik | Datenschutz | Digital | ZEIT ONLINE*. `http://www.zeit.de/datenschutz/malte-spitz-vorratsdaten`. 2011. – abgerufen am 12.01.2015

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

Hamburg, 2. März 2015 Vasilis Ikonomou