Investigatives Arbeiten mit Daten

Journalistische Innenansicht zu den Möglichkeiten, Anforderungen und ethischen Implikationen im Bereich CAR. *Von Marvin Oppong*

Als ich das erste Mal von dem Begriff „CAR" hörte, dachte ich zuerst an Autos. Das „Netzwerk Recherche" lud zu seiner Jahrestagung in Hamburg und bot im „CAR-room" mehrere Veranstaltungen zu Computer-Assisted Reporting an. Bei der Konferenz lernte ich auch zwei Kollegen kennen, mit denen ich später eine Liste aller Unternehmen, die an Parteien gespendet hatten, abglich mit einer Liste aller Unternehmen, in denen Bundestagsabgeordnete eine Nebentätigkeit ausübten. Zwar gaben die Ergebnisse bei diesem ersten Mal nicht genug für eine Veröffentlichung her, aber ich hatte Feuer für Computer-Assisted Reporting gefangen. Sieben Jahre später habe ich eine Vielzahl von Recherchen, die auf CAR basieren, durchgeführt und vermittle als Dozent meine Kenntnisse über dieses spannende und junge, noch in der Entwicklung befindliche Thema.

Computer-Assisted Reporting nahm seinen Anfang in den USA der 1980er Jahre. Als damals elektronische Datenverarbeitung langsam aber sicher Einzug in Redaktionen hielt, wurde es auch für Journalisten möglich, größere Datenmengen gezielt auszuwerten. Computer-Assisted Reporting war geboren. So glich Elliot Jaspin, Reporter beim „Providence Journal", 1986 Datenbanken ab, um Schulbusfahrer mit Verkehrssünden und ausgedehntem Vorstrafenregister ausfindig zu machen. Auf diese Weise fand er heraus, dass mehr als jeder vierte Fahrer in der Vergangenheit einen Verkehrsverstoß begangen hatte und einige Fahrer zuvor mit Drogen gehandelt hatten.

Marvin Oppong ist freier Journalist und Dozent für Rechercheteчniken. Zu seinen Spezialfeldern gehören das Informationsfreiheitsgesetz, Datenjournalismus und Werkzeuge für Internet-Recherchen.

Computer-Assisted Reporting lässt sich kurz und bündig definieren als „investigative Arbeit mit Daten". Beim CAR werden Computer eingesetzt, um Daten zu analysieren, die aus Datenbanken oder dem Datenbestand von Behörden stammen können. Meist liegt der Fokus auf dem Entdecken von Mustern, Häufungen oder Überschneidungen, dem Feststellen von Unterschieden und Abweichungen oder dem Überprüfen von Daten auf Validität. Genutzt werden Tabellenkalkulations-, Da-

tenbank- und Statistikprogramme wie Excel, Access, SPSS, aber auch Anwendungen wie „Google Fusion Tables“, die ich persönlich neben Excel viel einsetze. In einer Liste von Empfängern staatlicher Zahlungen zum Beispiel kann man dann schauen, welche Empfänger besonders viele Zahlungen erhalten oder in welche Regionen besonders viel Geld fließt.

Computer-Assisted Reporting ist eng mit dem neueren Begriff Datenjournalismus verbunden. Beide Richtungen des Journalismus lassen sich dadurch abgrenzen, dass es beim Datenjournalismus nicht nur allein um die Analyse von Daten geht, sondern auch um die spätere Darstellung für Leser, Zuhörer oder Zuschauer. Eine wichtige Komponente von Datenjournalismus ist deshalb auch die Visualisierung von Daten. Was die genaue Definition des Begriffs Datenjournalismus betrifft, scheiden sich unter Datenjournalisten die Geister.

Eine meiner ersten CAR-Geschichten hatte ihren Ursprung im Förderkatalog des Bundes (vgl. Förderkatalog). In dieser öffentlichen Datenbank dokumentieren Bundesministerien tausende Vorhaben der Projektförderung des Bundes. Ich hatte durch Explorieren der Datenbank herausgefunden, dass man, anstatt einen Suchbegriff in die Suchmaske einzugeben, auch gar nichts eingeben und dann eine gewaltige Menge an Daten in Excel exportieren kann. Als ich dann die Spalte „Zuwendungsempfänger“ des Datensatzes händisch durchsah, stieß ich auf einen Eintrag, der es in sich hatte: Das Bundesumweltministerium ließ sich bei Verfahren im Energie- und Umweltbereich ausgerechnet von solchen Kanzleien beraten, die Atomkonzerne zu ihren Kunden zählen. Im Nachrichtenmagazin „Der Spiegel“ berichtete ich später darüber, dass unter anderem die Großkanzlei „White & Case“ mehr als 460.000 Euro für „Juristische Unterstützung“ im Rahmen einer Klimaschutz-Richtlinie erhielt (Der Spiegel vom 14.3., S. 17). Die Kanzlei war in der Vergangenheit mehrmals für den Energiekonzern „Vattenfall“ tätig und bietet auf ihrer Webseite auch heute noch „Lobbying“ an. In einer Stellungnahme schloss das Bundesumweltministerium damals eine „wie auch immer geartete politische Einflussnahme“ aus.

Beim Datenjournalismus geht es nicht nur um die Datenanalyse sondern auch um deren Visualisierung.

Das Beispiel zeigt, dass auf dem Gebiet von CAR mit einfachen Mitteln viel möglich ist. Die Daten aus dem Förderkatalog waren für jedermann verfügbar, ein Tabellenkalkulationsprogramm ist intuitiv und nahezu kinderleicht zu bedienen – von „Raketenwissenschaft“ weit entfernt. Auch wenn ich in dem Bei-

spiel zugegebenermaßen den Förderkatalog als Quelle überhaupt erst einmal kennen musste, den Willen und die Fähigkeit haben musste, diese entsprechend auszuwerten, es dann beim Durchsehen der Datenreihen beim Namen der Kanzlei „klick" machen und die anschließende Vervollständigungsrecherche und Konfrontation handwerklich sauber durchgeführt werden musste.

CAR und Datenjournalismus sind ein großer Segen für den Journalismus. Bei der Vielzahl von Daten, die es heutzutage in vielen gesellschaftlichen Subsystemen gibt, eröffnet CAR Journalisten, aber auch Wissenschaftlern neue Türen. Mit wenig technischen, finanziellen und personellen Ressourcen können jede Journalistin und jeder Journalist große Datengeschichten produzieren. Die Basics des CAR kann man sich, dem Internet und Tutorials sei Dank, leicht anlesen.

Eine in der Branche viel diskutierte Frage war in jüngerer Zeit, ob man als Datenjournalist programmieren können muss. Sicherlich bedarf ein erfolgreicher Datenjournalismus überdurchschnittlicher Kompetenz auf technischem Gebiet, doch genauso wie ein Zeitungsredakteur nicht wissen muss, wie man Papier herstellt, muss in Zeiten moderner Arbeitsteilung auch ein Datenjournalist kein Programmierer sein.

In den letzten Jahren sind im Internet immer mehr Datenportale entstanden. Zu nennen wäre hier vor allem OpenSpending[1], ein Open-Data-Portal, in dem man komplette Haushalte einzelner Kommunen, Bundesländer und Staaten genauso findet wie einen Auszug aus der Datenbank der Empfänger des EU-Agrarfonds oder das EU-Budget bis 2020. Auf GovData[2], dem offiziellen Datenportal für Deutschland, sind Daten aller Verwaltungsebenen zugänglich. Auch die Europäische Union betreibt ein offenes Datenportal.[3] Die US-Journalistenorganisation „Investigative Reporters & Editors" bietet sogar einen Shop an, in dem man als Journalist Daten findet, wie zum Beispiel eine 13 Megabyte große Datei zu Schiffsunfällen von 1969 bis 2012.[4] Im „Hamburgischen Transparenzportal"[5], das im

1 *https://openspending.org/.*
2 *https://www.govdata.de/.*
3 *Offenes Datenportal der Europäischen Union, https://open-data.europa.eu/de/data.*
4 *Data Library des National Institute for Computer-Assisted Reporting von Investigative Reporters & Editors, https://www.ire.org/nicar/database-library/. Vgl. zu Open-Data-Portalen in den USA auch Oppong (2014).*
5 *http://transparenz.hamburg.de/.*

Oktober 2014 online ging, sind aktuell allein 2780 Datensätze hinterlegt.

Die auf diese Weise generierbaren Themen liegen also quasi „auf der Straße". Mehr technische Möglichkeiten gehen aber auch mit erhöhten Anforderungen an die Eignung, Befähigung und Medienkompetenz des einzelnen Journalisten einher: Im Datenjournalismus ist der Journalist nicht nur in der Rolle des Beschaffers, sondern muss auch komplexe Zusammenhänge erklären. Er muss nicht nur schreiben, sondern auch präsentieren. Wenn Daten für den Leser verfügbar gemacht werden, wird aus Sicht des Journalisten nicht behalten, sondern geteilt. Dies kann insbesondere für freie Journalisten zu moralisch schwierigen Fragen führen wie „Tue ich etwas für die Information der Öffentlichkeit und stelle meinen Datensatz schon jetzt mit der Geschichte komplett online oder veröffentliche ich nur einen Teil der Daten und mache danach eine Zweitverwertung, die mir Luft für weitere Geschichten zum Wohle der Öffentlichkeit gibt?". Das Verfügbarmachen von Daten beim Datenjournalismus führt zu mehr Transparenz, aber auch zu mehr Überprüfbarkeit von journalistischen Veröffentlichungen, was in einem Text getroffenen Aussagen mehr Legitimation verschafft, am Ende aber auch Kritikfähigkeit und das Vorhandensein einer gelebten Fehlerkultur erfordert.

Datenjournalismus führt zu mehr Transparenz, aber auch zu mehr Überprüfbarkeit von journalistischen Veröffentlichungen.

Computer-Assisted Reporting birgt viele neue Chancen, aber auch Risiken. Die größte Gefahr und Verlockung von CAR liegt darin, dass sich die journalistisch-handwerkliche Arbeit darauf beschränkt, sich in einen Datensatz zu vertiefen und einzelne Aspekte nicht mit Hilfe klassischer Recherchemittel wie Telefonrecherche, Recherche vor Ort oder durch Anfragen bei Behörden und Institutionen ergänzt werden. So kann schnell eine einseitige Sicht auf Dinge entstehen, die das große Ganze außer Acht und Reflexion, Interpretation und Einordnung vermissen lässt. Die empirisch gesehen immer kürzer werdende Zeit, die Journalisten für Recherche aufwenden, fördert diesen Effekt. Nur weil Daten für etwas sprechen, muss sich ein bestimmter Sachverhalt in der Praxis nicht in gleicher Weise darstellen. Ein Beispiel: Im EU-Lobbyregister, in dem Firmen und Lobbyberater freiwillig Angaben zu ihren Lobbyingaktivitäten bei EU-Institutionen machen können und das von dem NGO-Zusammenschluss „ALTER-EU" ausgewertet wurde (ALTER-EU vom 25.6.2012), gab es einen selbstständigen Lobbyberater, Ge-

orgios Stilianou. Dieser gab an, nur einen Kunden zu haben, in einem Jahr aber 100 Millionen Euro für Lobbying ausgegeben zu haben – ein in höchstem Maße unrealistischer Wert. Was es mit der Zahl 100 Millionen genau auf sich hat, bleibt im Dunkeln. Hätte man nun daraus eine Geschichte mit der Schlagzeile „Einzelner Lobbyberater gab mehr für Lobbying aus als Siemens und Shell zusammen" gemacht, wäre diese das Papier, auf dem sie gedruckt wäre, nicht wert, auch wenn sie nach den vorliegenden Daten formal korrekt gewesen wäre und es sich beim EU-Lobbyregister sogar um ein staatliches Register handelt und man als Journalist nach dem sogenannten Behördenprivileg presserechtlich gesehen auf Aussagen von Behörden vertrauen darf.

Man kann also beim CAR Daten nie isoliert betrachten, sondern muss ihnen auf den Grund gehen. Eine Datenanalyse kann immer nur der Ausgangspunkt, nie aber das Ende einer Recherche sein, auch wenn man unter Umständen im Laufe des Rechercheprozesses zwischendurch immer wieder zu den Daten zurückkehrt. Eine CAR-Geschichte darf und muss aus Zahlen sprechen, aber bloße Zahlen nie zur alleinigen Substanz haben.

Als Dozent habe ich Kursteilnehmer_innen schon häufig die EU-Ausschreibungsdatenbank „Tenders Electronic Daily"[6] näher gebracht. Einmal sagte ich den Kursteilnehmer_innen, dass der Umstand, dass zu einer Behörde oder einem Auftragnehmer keine Ausschreibung verzeichnet ist, nicht zwangsläufig bedeuten muss, dass dazu keine Ausschreibung existiert. Es kommt auch vor, dass eine Behörde unter Verstoß gegen geltendes Recht Aufträge nicht öffentlich ausschreibt, obwohl sie dies müsste. In einem solchen Fall sind die vorhandenen Daten nur ein Ausschnitt einer unbekannten Größe, deren tatsächlicher Umfang nur unter großen Schwierigkeiten, wenn überhaupt, feststellbar ist. Für einige der Kursteilnehmer_innen war das ein neuer Gedanke, den sie überhaupt nicht in Betracht gezogen hatten.

Neben der Bandbreite von Daten ist auch deren Qualität von Bedeutung. Es gibt etwa Fehler bei der Erhebung oder beim Einpflegen von Daten durch verschiedene Schreibweisen und Bearbeiter. Angenommen in einer Liste von Parteispendern hat der Beamte, der die Dateneingabe getätigt hat, ein Mal „BMW AG" eingegeben, bei anderen Einträgen „Bayerische Motoren Werke Aktiengesellschaft". Addierte man jetzt bei der Suche

6 *http://ted.europa.eu/TED/misc/chooseLanguage.do.*

nach dem größten Einzelspender in dem Datensatz einfach alle Einzelwerte zum Spender „BMW AG“, hätte man nur die halbe Wahrheit erfasst. In einem Datensatz mit hunderten oder gar tausenden von Einträgen ist es mitunter aber gar nicht so leicht, derartige Ungenauigkeiten zu entdecken. Bei der Geocodierung von Daten wiederum können Fehler entstehen, wenn verschiedene Städte in verschiedenen Ländern denselben Namen tragen. Hier sind Erfahrung und technische Kompetenz gefragt.

Ein seltenerer, aber auch denkbarer Fall ist die bewusste Verfälschung von Daten. Machte man etwa die offizielle Arbeitslosenstatistik zur Grundlage einer datenjournalistischen Auswertung, ohne in Betracht zu ziehen, dass die Berechnungsweise der Arbeitslosigkeit über die Jahre mehrfach geändert wurde, Arbeitslose, die von privaten Vermittlern betreut werden, und Trainings- und Eingliederungsmaßnahmen aus der Statistik genommen wurden und sich die Bundesagentur für Arbeit sogar Manipulationsvorwürfen ausgesetzt sah, wäre das Ergebnis nicht besonders aussagekräftig. Machte man einen europaweiten Vergleich von Arbeitslosenquoten, müsste man wiederum in Rechnung stellen, dass die Arbeitslosenquote in jedem Land anders berechnet wird.

„Je mehr Datenjournalismus betrieben wird, desto größer ist der Anreiz, Daten zu frisieren.“ (Big-Data-Experte Hendrik Stange)

Zahlen geben immer auch Möglichkeiten zur Manipulation. Beim CAR vergrößert die Zahlenlastigkeit auch die Möglichkeit zur Manipulation. Der Big-Data-Experte Hendrik Stange vom Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme sagte bei der von mir veranstalteten Reihe „Recherche-Lab“ im Februar 2014: „Je mehr Datenjournalismus betrieben wird, desto größer ist der Anreiz, Daten zu frisieren.“ Auch die PR-Branche und Unternehmen haben die vermeintliche Neutralität von Daten entdeckt, um Datensätze für ihre Zwecke kostenlos Journalisten zur Verfügung zu stellen und die Medien so als Überbringer zu instrumentalisieren (vgl. Harris 2015). Beim CAR können vorhandene Zahlen so „frisiert“ werden, dass bestimmte Aussagen untermauert oder gar erst erzeugt werden oder ganz wegfallen. Dies gilt auch für die anschließende Visualisierung, etwa wenn Balken in Diagrammen so dargestellt werden, dass ein besonderer Umstand unter- oder überbetont wird.

All dies stellt besondere Anforderungen an die Integrität und die Moral von Datenjournalisten. Aber auch Institutionen, die Journalismus lehren, sind gefragt, entsprechende Medi-

enkompetenz zu vermitteln. Was das Thema CAR betrifft, so fällt auf, dass es in Deutschland keinen einzigen Lehrstuhl für dieses Gebiet gibt. Wie weit es mit der technischen Kompetenz bei manchen Lehrkräften bestellt ist, habe ich erfahren, als ich einmal einen Workshop für Studierende gegeben habe, die mir erzählten, dass die Vermittlung von Recherchekompetenz in einem anderen ihrer Kurse sich auf bloßes Googeln zu einem Thema beschränke.

Bei aller Integrität von Journalisten bedarf es aber auch kritischer Medienkonsumenten, die den Medienmachern auf die Finger schauen und damit als weitere Filterinstanz dafür sorgen, dass das Potential von CAR auf positivste Weise ausgeschöpft wird.

Literatur

Alliance for Lobbying Transparency and Ethics Regulation (ALTER-EU) (2012) (Hg.): Dodgy Data: Time to fix the EU's Transparency Register vom 25.6. http://www.alter-eu.org/documents/2012/06/dodgy-data (zuletzt aufgerufen am 16.2.2015).

Die Bundesregierung (o.J.): Förderkatalog. http://foerderportal.bund.de/foekat/jsp/StartAction.do?actionMode=list (zuletzt aufgerufen am 16.2.2015). [zitiert: Förderkatalog]

Harris, Jacob: Predictions for Journalism 2015. A wave of P.R. data. Predictions for Journalism 2015. NiemanLab. http://www.niemanlab.org/2014/12/a-wave-of-p-r-data/ (zuletzt aufgerufen am 16.2.2015).

(o.V.) (2011): Heikle Beratung. In: Der Spiegel vom 14.3., S. 17. http://magazin.spiegel.de/EpubDelivery/spiegel/pdf/77435162 (zuletzt aufgerufen am 16.2.2015).

Oppong, Marvin (2014): Die Open-Data-Portale in den USA. In: ZDF Hyperland vom 7. 1. 2014. http://blog.zdf.de/hyperland/2014/01/die-open-data-portale-in-den-usa/ (zuletzt aufgerufen am 16.2.2015).

Offene Datenportale

OpenSpending: Haushalte von Kommunen, Bundesländern und Staaten. https://openspending.org/

Datenportal für Deutschland mit Daten aller Verwaltungsebenen. https://www.govdata.de/

Datenportal der Europäischen Union. https://open-data.europa.eu/de/data

Data Library des National Institute for Computer-Assisted Reporting von "Investigative Reporters & Editors". https://www.ire.org/nicar/database-library/

Hamburgisches Transparenzportal. http://transparenz.hamburg.de/