Sebastian Hellmeier/Johannes Vüllers

Neue Entwicklungen und Herausforderungen bei der Kodierung von Ereignisdatensätzen

Kodierungsentscheidungen spielen eine zentrale Rolle bei der Erstellung quantitativer Datensätze. Der technische Fortschritt und die zunehmende Disaggregation der Daten bringen neue Herausforderungen für den Kodierungsprozess und die Konzeptformierung mit sich. In diesem Beitrag diskutieren wir, welche Chancen und Risiken sich aus der gestiegenen Verfügbarkeit von Daten ergeben und wie wir mit zu wenig oder mit zu vielen Informationen im Kodierungsprozess umgehen können. Aus unserer Sicht können technische Hilfsmittel derzeit vor allem unterstützend zur menschlichen Kodierung eingesetzt werden. Darüber hinaus betonen wir die Notwendigkeit eines transparenten Umgangs mit konzeptionell und empirisch unklaren Fällen.

1. Einleitung

Eine Reihe höchst relevanter Fragestellungen in der Forschung betreffen Phänomene, die in regelmäßigem Abstand an verschiedenen Orten auftreten: Führen Demokratien weniger häufig Krieg als Autokratien? Wann und warum entstehen Protestbewegungen? Die jüngsten Entwicklungen im Bereich des maschinellen Kodierens und die zunehmende Disaggregation der Datensätze bringen neue Möglichkeiten und Herausforderungen mit sich. Es stellt sich daher die Frage, wie mit diesen Neuerungen bestmöglich umgegangen werden kann und welche Folgen für den Kodierungsprozess und die Konzeptformierung absehbar sind.

Jede Art von Datensätzen ist von den jüngsten Entwicklungen im technischen Bereich betroffen, jedoch in unterschiedlichem Ausmaß. Jene Datensätze, die für ihre Kodierungen auf eine möglichst große Anzahl unterschiedlicher Informationsquellen zurückgreifen, sehen sich den größten Herausforderungen gegenüber. Dies gilt insbesondere für medienbasierte Ereignisdatensätze wie zum Beispiel die *Social Conflict Analysis Database* (Salehyan et al. 2012) oder das *Georeferenced Event Dataset* des *Uppsala Conflict Data Program* (Sundberg/Melander 2013).[1] Die Herausforderungen für Datensätze, die auf anderen Informationsquellen wie beispielsweise Regierungsdokumenten oder Berichten von Nichtregierungsorganisationen (*non-governmental organizations*, NGOs) basieren, werden wir aus Platzgründen daher nur vereinzelt diskutieren.

1 Siehe Otto (2018 in diesem Forum) zu einer umfassenderen Übersicht von medienbasierten Ereignisdatensätzen.

Nachdem wir die jüngsten Entwicklungen im Bereich der Sammlung von Ereignisdaten skizziert haben, konzentrieren wir uns in diesem Beitrag auf zwei zentrale Herausforderungen im Kodierungsprozess von Ereignisdaten.[2] Zunächst gehen wir auf die Folgen einer übermäßigen Informationsflut, aber auch auf das Fehlen ausreichender Informationen für Kodierungsentscheidungen ein. In einem weiteren Schritt diskutieren wir den Umgang mit konzeptionell und empirisch unklaren Fällen im Kodierungsprozess. Hier zeigen wir Zielkonflikte zwischen verschiedenen Gütekriterien der Konzeptformierung (Gerring 1999) und pragmatischen Kodierungsentscheidungen auf. Wir plädieren in diesem Beitrag dafür, neue technische Hilfsmittel gewinnbringend einzusetzen, ohne dabei zentrale konzeptionelle Entscheidungen an die Technik auszulagern. Vielmehr soll eine transparente Rückbindung der Kodierung an die Konzeptformierung erreicht werden.

2. *Jüngste Entwicklungen*

Die Sammlung von Ereignisdaten in der Politikwissenschaft weist eine lange Tradition auf, in der die Methodik der Datenerhebung regelmäßig hinterfragt wurde. Zwei wichtige Entwicklungen geben allerdings Anlass, vertieft über die Kodierung von Ereignisdaten nachzudenken: der technische Fortschritt und der Trend zur Disaggregation der Daten.

Die technischen Entwicklungen der letzten Jahre haben die Sammlung von Ereignisdaten effektiver und effizienter gemacht. Dabei sind Medienberichte nach wie vor die bevorzugte Datenquelle, obgleich dies vor allem auf pragmatische Überlegungen zurückzuführen ist (Koopmanns 1998). Die fortschreitende Digitalisierung von Zeitungsinhalten und die Verfügbarkeit von Online-Datenbanken wie *Factiva* oder *LexisNexis*, die systematisch nach Stichwörtern durchsucht werden können, ermöglichen mittlerweile nahezu eine Vollerhebung relevanter Nachrichtenberichte bei vergleichsweise geringem Aufwand.

Darüber hinaus sind die Verfahren maschineller Kodierungen, die in kürzerer Zeit mehr Informationen als menschliche Kodierer_innen verarbeiten können, weit fortgeschritten. So schätzen Beieler et al. (2016: 99), dass eine aktuelle Software zwischen 1.000 und 2.000 Sätze pro Sekunde auswerten und zwischen 4.000 und 8.000 Ereignisse pro Tag identifizieren kann. Laut der Studie haben diese maschinellen Verfahren eine Treffsicherheit von 70 bis 80 Prozent (Beieler et al. 2016: 98). Auch wenn vollautomatische Verfahren vermehrt zum Einsatz kommen – insbesondere bei der *Vorhersage* von Ereignissen – werden sie die Forschenden bei der Kodierung nicht ersetzen können. Ein Abgesang auf menschliche Kodierer_innen, wie er von einigen angestimmt wird (z. B. Schrodt 2012), ist verfrüht, da zahlreiche konzeptionelle Entscheidungen im Kodierungsprozess immer wieder neu getroffen werden müssen (Adcock/Collier 2001). Hinzu kommt, dass automatisiert

2 Wir gehen in unserem Beitrag nicht näher auf die ebenfalls wichtige Auswahl der Quellen ein und verweisen hier ebenfalls auf Otto (2018 in diesem Forum).

kodierte Daten nur schwer zu validieren sind. Häufig bleibt unklar, ob die Algorithmen wirklich im Einklang mit den verwendeten Konzepten stehen oder doch etwas völlig anderes messen. Der Vergleich mit händisch kodierten Ereignisdatensätzen fördert nur eine begrenzte Übereinstimmung zutage (Hammond/Weidmann 2014: 3-5).

Eine zweite Entwicklung der letzten Jahre ist die zunehmende Disaggregation von Ereignisdaten in Hinblick auf den zeitlichen und räumlichen Kontext der erfassten Ereignisse. Wurden in klassischen Datensätzen in erster Linie harte Fakten (*Wer? Was? Wie?*) erfasst, so werden heute genauere Informationen erhoben (z. B. Organisationsnamen, Aufspaltung der Ziele, angewandte Proteststrategien) (z. B. Cunningham et al. 2017). Mithilfe dieser Informationen können viele theoretische Annahmen wie zum Beispiel der Einfluss des lokalen Zugangs zu Informationstechnologien auf die Konfliktwahrscheinlichkeit erstmals empirisch untersucht werden. Dieser gestiegene Anspruch an Ereignisdatensätze führt gleichzeitig zu einem hohen Kodieraufwand.

Aus diesen Entwicklungen ergeben sich sowohl Chancen als auch Herausforderungen für den Kodierungsprozess. Auf der einen Seite kann die genauere Erfassung der Wirklichkeit zu mehr Kongruenz zwischen Theorie und Empirie führen, das heißt spezifische Aspekte eines theoretischen Ansatzes können erstmals quantitativ überprüft werden. Auf der anderen Seite sind zusätzliche Entscheidungen bei der Konzeptformierung und im Kodierungsprozess zu fällen, mit der Folge, dass die Reliabilität der Untersuchungsergebnisse unter Umständen nicht gewährleistet werden kann. Im Folgenden gehen wir auf beide Herausforderungen ein.

3. Informationsflut und Informationsknappheit

Während die Ursachen für Verzerrungen in der Berichterstattung und aufgrund der Quellenauswahl ein bekanntes, wenn auch nicht gelöstes Problem darstellen (vgl. Otto 2018 in diesem Forum), stellt sich die Frage nach dem Umgang mit der Informationsflut und Informationsknappheit im Kodierungsprozess.

3.1. Informationsflut

Die Digitalisierung hat dazu geführt, dass eine schier unendliche Menge an Informationen jederzeit abrufbar ist. Es muss nicht mehr allein auf Medienquellen zurückgegriffen werden, sondern Dokumente Internationaler Organisationen, die jene auf ihren Internetseiten der Öffentlichkeit zur Verfügung stellen, sind nützliche weitere Informationsquellen für die Kodierung von Ereignisdatensätzen (s. Knecht/Debre 2018 in diesem Forum). Dies hat den Aufwand der Datensammlung verringert, da der Zugang zu Informationen oft problemlos möglich ist.

Diese Informationsflut birgt große Herausforderungen für jene Datensätze, die Ereignisse (z. B. Protest, Gewalt) basierend auf Nachrichtenquellen erfassen. Die

Anzahl der zugänglichen Nachrichtenquellen ist in den letzten Jahren exponentiell gestiegen. Dies hat eine transparente Selektion des zu berücksichtigenden Quellenmaterials im Kodierungsprozess zunehmend an Bedeutung gewinnen lassen, da eine vollständige Auswertung aufgrund der begrenzten Ressourcen von Forschungsprojekten ausgeschlossen ist. Dabei ist die Berücksichtigung möglichst vieler Informationsquellen nicht unbedingt der beste Weg, um die Verzerrung im Kodierungsprozess gering zu halten. Eine große Anzahl unterschiedlicher Datenquellen verringert die Verzerrung nämlich nur, wenn jede Quelle selbst einen eigenen und von den anderen Quellen unterschiedlichen Verzerrungseffekt hat. Nur zusammengenommen gleichen sich die unterschiedlichen Effekte aus und kommen so einem weniger verzerrten Ideal näher (Marks 2007: 5-7). Die Auswahl der Informationsquellen sollte daher immer im Hinblick auf potenzielle Verzerrungen begründet werden.

Nach der Quellenauswahl müssen die Kodierer_innen eine Vielzahl an Informationen sichten, um jene für die Kodierung relevanten Ereignisse zu identifizieren. Eine Stichwortsuche zu weit verbreiteten Phänomenen fördert tausende Einträge zutage, von denen nur ein Bruchteil für die Kodierung relevante Informationen enthält. Die vollständige Kodierung jedes Ereignisberichts in den Quellen steht somit in keinem Verhältnis zum Ertrag. Bei der Vorbereitung und Sortierung des zu kodierenden Materials kann daher die maschinelle Kodierung von großem Nutzen sein. Mithilfe maschineller Lernverfahren haben Croicu und Weidmann (2015) Algorithmen entwickelt, die einen Großteil der irrelevanten Einträge bereits vor der händischen Kodierung aussortieren können. Ihr Programm ist in der Lage, die Anzahl der zu kodierenden Artikel zu halbieren und gleichzeitig über 90 Prozent der relevanten Berichte beizubehalten. Die Vorauswahl mithilfe des Computers verringert den Arbeitsumfang durch den Ausschluss irrelevanter Berichte deutlich. Diese Technik kann prinzipiell auch bei anderen textbasierten Datenerhebungen gewinnbringend eingesetzt werden, beispielsweise bei der Identifikation relevanter Passagen in langen Gesetzestexten oder Protokollen.

Die maschinelle Textkodierung ist ein vielversprechendes Hilfsmittel für die Kodierung einer überschaubaren Anzahl von Primärquellen wie etwa Resolutionen der *United Nations* (UN) (z. B. Beardsley 2013). Der größte Mehrwert der zunehmenden Digitalisierung besteht hier darin, dass die Dokumente mit geringem oder keinem Mehraufwand aus dem Internet heruntergeladen werden können. So veröffentlichen die UN oder die Europäische Union (EU) eine Vielzahl ihrer Dokumente im Netz, während etwa das *Uppsala Conflict Data Program* (UCDP) als Teil seiner Konfliktenzyklopädie eingescannte Originaldokumente der Friedensverträge bereitstellt. Hierdurch wird auch der Aufwand für qualitative Forschungsarbeiten reduziert.

3.2. Informationsknappheit

Trotz der schier unbegrenzten Informationsmenge im digitalen Zeitalter gibt es immer noch eine Vielzahl an Ereignissen, die wenig mediale Aufmerksamkeit erfahren. Dabei handelt es sich um Ereignisse und Länder, die aus Sicht der Medienschaffenden einen niedrigen Nachrichtenwert haben, also kleiner, weniger gewaltsam oder weltpolitisch unbedeutend sind (Herkenrath/Knoll 2011: 171-175). Fehlende Informationen zu bestimmten Ereignissen sind nicht nur ein Problem der Quellenauswahl, das etwa durch die Berücksichtigung lokaler Medienquellen verringert werden kann (Nam 2006: 283-284); Informationsknappheit ist auch ein Problem im Kodierungsprozess.

Die Anforderungen an den Informationsumfang einzelner Quellen sind mit der zunehmenden Disaggregation der Datensätze gestiegen. Es reicht nicht mehr, nur eine grobe Beschreibung der Identitäten von Aktivist_innen im Datensatz anzugeben. Stattdessen erfassen neuere Datensätze die Namen der Organisationen, die an dem Ereignis beteiligt waren, und die angewandten Proteststrategien (Cunningham et al. 2017). Die angesprochene Informationsflut führt nicht notwendigerweise dazu, dass die Datenquellen die notwendige Informationsqualität für disaggregierte Datensätze aufweisen, nur weil über die einzelnen Ereignisse nun in mehreren Quellen berichtet wird. Der Umfang des Problems der Informationsknappheit erhöht sich relativ zu der gewünschten Disaggregation einzelner Variablen, für die es zu einer großen Anzahl fehlender Werte kommen kann. Hier gibt es kein Patentrezept, das angewandt werden kann. Zu Beginn eines Kodierungsprozesses kann allein darauf geachtet werden, dass die Diskrepanz zwischen Datengenauigkeit und -verfügbarkeit nicht zu groß ist.

4. Umgang mit unklaren Fällen

Mit der Informationsflut und -knappheit steigt die Herausforderung des Umgangs mit unklaren Fällen während des Kodierungsprozesses. Im Folgenden unterscheiden wir zwischen Unklarheiten in Bezug auf die gewählten theoretischen Konzepte und in Bezug auf unklare Informationen in den Datenquellen.

4.1. Konzeptionelle Unklarheiten

Der Umgang mit Grenzfällen und konzeptionellen Graubereichen ist wichtig, um das untersuchte Phänomen klar ein- und abzugrenzen und die Reliabilität der Kodierungsentscheidungen zu gewährleisten. Die Konzeptbildung ist hierfür entscheidend, da sie die der Forschung zugrunde liegenden Konzepte bestimmter Begrifflichkeiten (z. B. Bürgerkrieg) operationalisierbar machen soll (Collier et al. 2012; Gerring 1999). Ohne Konzeptbildung besteht die Gefahr, dass sich mehrere Arbei-

ten zwar vermeintlich mit demselben Thema beschäftigen, aber allein aufgrund unterschiedlicher Konzeptspezifikationen zu unterschiedlichen Ergebnissen kommen. Abhängig von der Definition und dem Schwellenwert der Opferzahlen werden zum Beispiel die Auseinandersetzungen in Kolumbien wahlweise als Bürgerkrieg, bewaffneter Konflikt, Revolutionskrieg oder Krieg in den Datensätzen erfasst (Restrepo et al. 2006). Eine klare Abgrenzung des untersuchten Phänomens unter Rückbindung an bestehende Untersuchungen ist daher unerlässlich.

Die Herausforderung der Konzeptbildung schlägt sich auch auf den Kodierungsprozess nieder (Adcock/Collier 2001: 531). Der Zielkonflikt bei der Erstellung von Kodiervorgaben liegt oft darin, mit klaren Einschränkungen und Vorgaben ein allgemeines Phänomen zu erfassen (Konzeptbildung) und dennoch Raum für fallspezifische Besonderheiten zu lassen. Ein Pretest kann bei der Festlegung der Kodierregeln helfen, aber niemals alle Spezialfälle abbilden, die im Laufe des Kodierens auftreten. Ein Ausweg ist ein Kompromiss zwischen einer frühen Festlegung konkreter Definitionen sowie unveränderlichen Oberkategorien und einem flexiblen Umgang mit später notwendigen Anpassungen (Rucht/Neidhardt 1998: 83). Hierbei sollte der Grundsatz gelten, die Zahl der Ausnahmen möglichst gering zu halten. Dies erhöht die Nachvollziehbarkeit der Kodierregeln und hält den Kodieraufwand gering, da bei der Gewährung von Ausnahmen alle bereits erfassten Datenpunkte erneut überprüft werden müssten. Erfolgt keine erneute Überprüfung der Datenpunkte, so führt die neue Ausnahme zu einer Verzerrung des gesamten Datensatzes. Für den Fall, dass größere Anpassungen vorgenommen werden müssen, kann eine Änderungshistorie der Daten die notwendige Transparenz herstellen.

Es ist ebenso von Bedeutung, größtmögliche Transparenz bei konzeptionellen Unklarheiten herzustellen (Salehyan 2015: 107). Die genaue Beschreibung der Kodierungsentscheidungen ermöglicht es den Nutzer_innen der Daten, die betroffenen Fälle aus der Analyse auszuschließen oder eine alternative Kodierungsentscheidung zu treffen. Dies ist in erster Linie bei Datensätzen auf Grundlage von Primärquellen möglich, während der Umfang von medienbasierten Ereignisdatensätzen dies oft nicht erlaubt. Da unklare Fälle Teil eines jeden Kodierungsprozesses sind, ist es erstaunlich, wie wenige Datensätze hierzu Informationen in ihren öffentlich zugänglichen Versionen bereitstellen.

4.2. Empirische Unklarheiten

Ein weiteres Problem unklarer Fälle ist, dass die empirischen Informationen oft eine eindeutige Klassifizierung nicht zulassen. So kann ein Bericht über einen Straßenprotest den friedlichen Charakter der Aktivist_innen betonen und gleichzeitig auf vereinzelte gewaltbereite Steinewerfer_innen verweisen. Ist ein solcher Protest als gewaltfrei oder gewaltsam zu kodieren? Es gibt verschiedene Möglichkeiten, mit einem solch unklaren Fall umzugehen. Eine erste ist, dass man die Unterscheidung zwischen gewaltfrei vs. gewaltsam durch eine Einschränkung der eigentlichen Kodierungskategorien – zum Beispiel »primarily nonviolent« (Chenoweth/

Lewis 2013: 418-419) – ersetzt. Dies ist jedoch keine Veränderung des Konzepts von Gewaltfreiheit, sondern allein eine empirische Einschränkung, die die Kodierungsentscheidung bei unklaren Fällen nicht zwingend einfacher macht. So stellt sich zwangsläufig die Frage, wann ein Protest nicht mehr *vornehmlich* gewaltfrei ist.

Eine weitere Möglichkeit ist, die dichotome Kodierung komplexer Konzepte im Kodierungsprozess zu umgehen, indem die Untersuchungseinheit disaggregiert wird. Im Fall von gewaltfreien Kampagnen würden die Einzelereignisse einer Kampagne (z. B. Demonstrationen, Sit-ins) kodiert und nicht mehr die Kampagne als Ganzes. Dies erleichtert den Kodierungsprozess, da die Kodierregeln für abgrenzbarere Fälle transparenter sind und so über eine kleinere Anzahl von unklaren Fällen entschieden werden muss. Ein berechtigter Einwand ist, dass der Mehraufwand im Kodierungsprozess stark ansteigt.

Einzelne Fälle können empirisch unklar einzuordnen sein, da häufig mehrere Nachrichtenberichte für ein bestimmtes Ereignis vorliegen, die oft unterschiedliche Angaben machen, beispielsweise zur Anzahl der Protestteilnehmer_innen. Hinzu kommt, dass Berichte häufig die Positionen sich feindlich gegenüberstehender Akteure wiedergeben, die ein Interesse an einer nicht wahrheitsgetreuen Darstellung der Ereignisse haben (Day et al. 2015: 131-132). So werden staatliche Stellen die Anzahl verletzter Demonstrant_innen nach einer gewaltsamen Intervention niedriger angeben als die Protestorganisator_innen, und umgekehrt. Eine Gewichtung der widersprüchlichen Aussagen kann von den Forschenden meist nicht vorgenommen werden. Eine Möglichkeit damit umzugehen ist die explizite Unterscheidung zwischen dem *Ereignisbericht* und dem *Ereignis* selbst, was den transparenten Umgang mit Unsicherheiten in Bezug auf divergierende Informationen in den Quellen ermöglicht (Weidmann/Rød 2015). Auch hier ist die Empfehlung, im Kodierungsprozess ein möglichst großes Maß an Transparenz sicherzustellen, während die Nutzer_innen des Datensatzes letztlich die Entscheidung treffen müssen, welche Informationen in der Analyse berücksichtigt werden sollen.

Darüber hinaus stehen weitere Instrumente beim Umgang mit unklaren Fällen zur Verfügung. Denn selbst wenn einzelne Fälle nicht eindeutig kategorisiert werden können, so sollte dennoch mit ähnlichen unklaren Fällen gleich verfahren werden. Kommen Forscher_innen nicht zu vergleichbaren Ergebnissen bei der Kodierung eines Falles, kann dies zu systematischen Fehlern und Verzerrungen in den erhobenen Daten führen (Ruggeri et al. 2011). Die Ergebnisse sogenannter *intercoder-reliability*-Tests sollten daher standardmäßig berichtet werden, um mögliche Unklarheiten des Kodiervorgangs für Außenstehende transparent zu machen.

5. Fazit

Die Kodierung von Datenquellen ist ein zentrales Element bei der Erstellung von Datensätzen. Der technische Fortschritt und die gestiegenen Ansprüche an disaggregierte Ereignisdatensätze rücken die Bedeutung von zu treffenden Entschei-

dungen im Kodierungsprozess in den Fokus. In diesem Beitrag haben wir gezeigt, dass sowohl ein Übermaß an Quellenmaterial und Informationen als auch fehlende und ungenügende Informationen eine Herausforderung für den Kodierungsprozess sind. Software kann derzeit weniger bei der Kodierung an sich als bei der Erkennung irrelevanter Ereignisberichte gewinnbringend eingesetzt werden. Der Umgang mit unklaren Fällen ist eine zweite zentrale Herausforderung im Kodierungsprozess. Die Konzepte sollten eine hohe Reliabilität der Kodierungen sicherstellen, jedoch gleichzeitig Ausnahmen erfassen können, um der komplexen Realität gerecht zu werden. Der Umgang mit empirisch unklaren Fällen ist vergleichbar einfacher. Es gilt, möglichst im Kodierungsprozess große Transparenz über divergierende Informationen zu einem Ereignis herzustellen, während die Entscheidung über den Umgang mit diesen divergierenden Informationen im Analyseprozess getroffen werden muss.

Literatur

Adcock, Robert/Collier, David 2001: Measurement Validity: A Shared Standard for Qualitative and Quantitative Research, in: American Political Science Review 95: 3, 529-546.

Beardsley, Kyle 2013: The UN at the Peacemaking-Peacebuilding Nexus, in: Conflict Management and Peace Science 30: 4, 369-386.

Beieler, John/Brandt, Patrick T./Halterman, Andrew/Schrodt, Philip A./Simpson Erin M. 2016: Generating Political Event Data in Near Real Time: Opportunities and Challenges, in: Alvarez, Michael R. (Hrsg.): Computational Social Science Discovery and Prediction, Cambridge, 98-120.

Chenoweth, Erica/Lewis, Orion 2013: Unpacking Nonviolent Campaigns: Introducing the NAVCO 2.0 Dataset, in: Journal of Peace Research 50: 3, 415-423.

Collier, David/LaPorte, Jody/Seawright, Jason 2012: Putting Typologies to Work: Concept Formation, Measurement, and Analytic Rigor, in: Political Research Quarterly 65: 1, 217-232.

Croicu, Mihai/Weidmann, Nils B. 2015: Improving the Selection of News Reports for Event Coding Using Ensemble Classification, in: Research & Politics 2: 4, 1-8.

Cunningham, Kathleen Gallagher/Dahl, Marianne/Frugé, Anne 2017: Strategies of Resistance: Diversification and Diffusion, in: American Journal of Political Science 61: 3, 591-605.

Day, Joel/Pinckney, Jonathan/Chenoweth, Erica 2015: Collecting Data on Nonviolent Action: Lessons Learned and Ways Forward, in: Journal of Peace Research 52: 1, 129-133.

Gerring, John 1999: What Makes a Concept Good? A Critical Framework for Understanding Concept Formation in the Social Sciences, in: Polity 31: 3, 357-393.

Hammond, Jesse/Weidmann, Nils B. 2014: Using Machine-coded Event Data for the Micro-level Study of Political Violence, in: Research & Politics 1: 2, 1-8.

Herkenrath, Mark/Knoll, Alex 2011: Protest Events in International Press Coverage: An Empirical Critique of Cross-National Conflict Databases, in: International Journal of Comparative Sociology 52: 3, 163-180.

Knecht, Sebastian/Debre, Maria J. 2018: Die »digitale IO«: Chancen und Risiken von Online-Daten für die Forschung zu Internationalen Organisationen, in: Zeitschrift für Internationale Beziehungen 25: 1, 175-188.

Koopmans, Ruud 1998: The Use of Protest Event Data in Comparative Research: Cross-national Comparability, Sampling Methods and Robustness, in: Rucht, Dieter/Koopmans,

Ruud/Neidhardt, Friedhelm (Hrsg.): Acts of Dissent. New Developments in the Study of Protest, Berlin, 90-112.

Marks, Gary 2007: Introduction: Triangulation and the Square-Root Law, in: Electoral Studies 26: 1, 1-10.

Nam, Taehyun 2006: What You Use Matters: Coding Protest Data, in: PS: Political Science & Politics 39: 2, 281-287.

Otto, Sabine 2018: Herausforderungen und Möglichkeiten medienbasierter Bürgerkriegsdatensätze, in: Zeitschrift für Internationale Beziehungen 25: 1, 131-144.

Restrepo, Jorge A./Spagat, Michael/Vargas, Juan F. 2006: The Severity of the Colombian Conflict: Cross-Country Datasets Versus New Micro-Data, in: Journal of Peace Research 43: 1, 99-115.

Rucht, Dieter/Neidhardt, Friedhelm 1998: Methodological Issues in Collecting Protest Event Data: Units of Analysis, Sources and Sampling, Coding Problems, in: Rucht, Dieter/Koopmans, Ruud/Neidhardt, Friedhelm (Hrsg.): Acts of Dissent. New Developments in the Study of Protest, Berlin, 65-89.

Ruggeri, Andrea/Gizelis, Theodora-Ismene/Dorussen, Han 2011: Events Data as Bismarck's Sausages? Intercoder Reliability, Coders' Selection, and Data Quality, in: International Interactions 37: 3, 340-361.

Salehyan, Idean 2015: Best Practices in the Collection of Conflict Data, in: Journal of Peace Research 52: 1, 105-109.

Salehyan, Idean/Hendrix, Cullen S./Hamner, Jesse/Case, Christina/Linebarger, Christopher/Stull, Emily/Williams, Jennifer 2012: Social Conflict in Africa: A New Database, in: International Interactions 38: 4, 503-511.

Schrodt, Philip A. 2012: Precedents, Progress, and Prospects in Political Event Data, in: International Interactions 38: 4, 546-569.

Sundberg, Ralph/Melander, Erik 2013: Introducing the UCDP Georeferenced Event Dataset, in: Journal of Peace Research 50: 4, 523-532.

Weidmann, Nils B./Rød, Espen Geelmuyden 2015: Making Uncertainty Explicit: Separating Reports and Events in the Coding of Violence and Contention, in: Journal of Peace Research 52: 1, 125-128.