gesis Leibniz-Institut für Sozialwissenschaften

Privatheitsschutz durch Open Data und Trusted Third Parties: Plädoyer für die öffentliche Kontrolle sozialer Daten

Katharina Kinder-Kurlanda
GESIS – Leibniz-Institut für Sozialwissenschaften, Köln
Institut für Web Science and Technologies, Universität Koblenz
katharina.kinder-kurlanda@gesis.org @ka_kinder

Mitglied der Leibniz Gemeinschaft

Überblick

- Social-Media-Daten als Gegenstand datenökonomischer Überlegungen
- Schwierigkeiten in Zugang und Teilen der Daten
- Rolle von Plattformbetreibern
- Möglicheiten der Archivierung

Hintergrund

- Ethnografische / Praxeografische Studie der Datenpraktiken von Social-Media-Forschenden (mit Dr. Katrin Weller)
- Archivierungspraxis bei Social-Media-Daten im Datenarchiv der GESIS

- Diskussionen über Privacy und historische Entwicklung von Konzepten (mit Dr. Carsten Ochs)

Social-Media-(Big)-Daten

- Tweets, Facebook-Nachrichten oder andere nutzergenerierte Inhalte, z.B. Texte, Fotos, Links
- Meta-/Para-/Prozess-Daten, z.B. Lokation, Time-Stamps

Social-Media-(Big)-Daten

- Gegenstand akademischer Forschung
- Gegenstand eigener Forschung der Plattformen (opak)
- Bedeutung als zeitgenössische und historische Quelle (z.B. politische Aushandlungsprozesse)

International Journal of Communication 8 (2014), 646–672 1932–8036/20140005

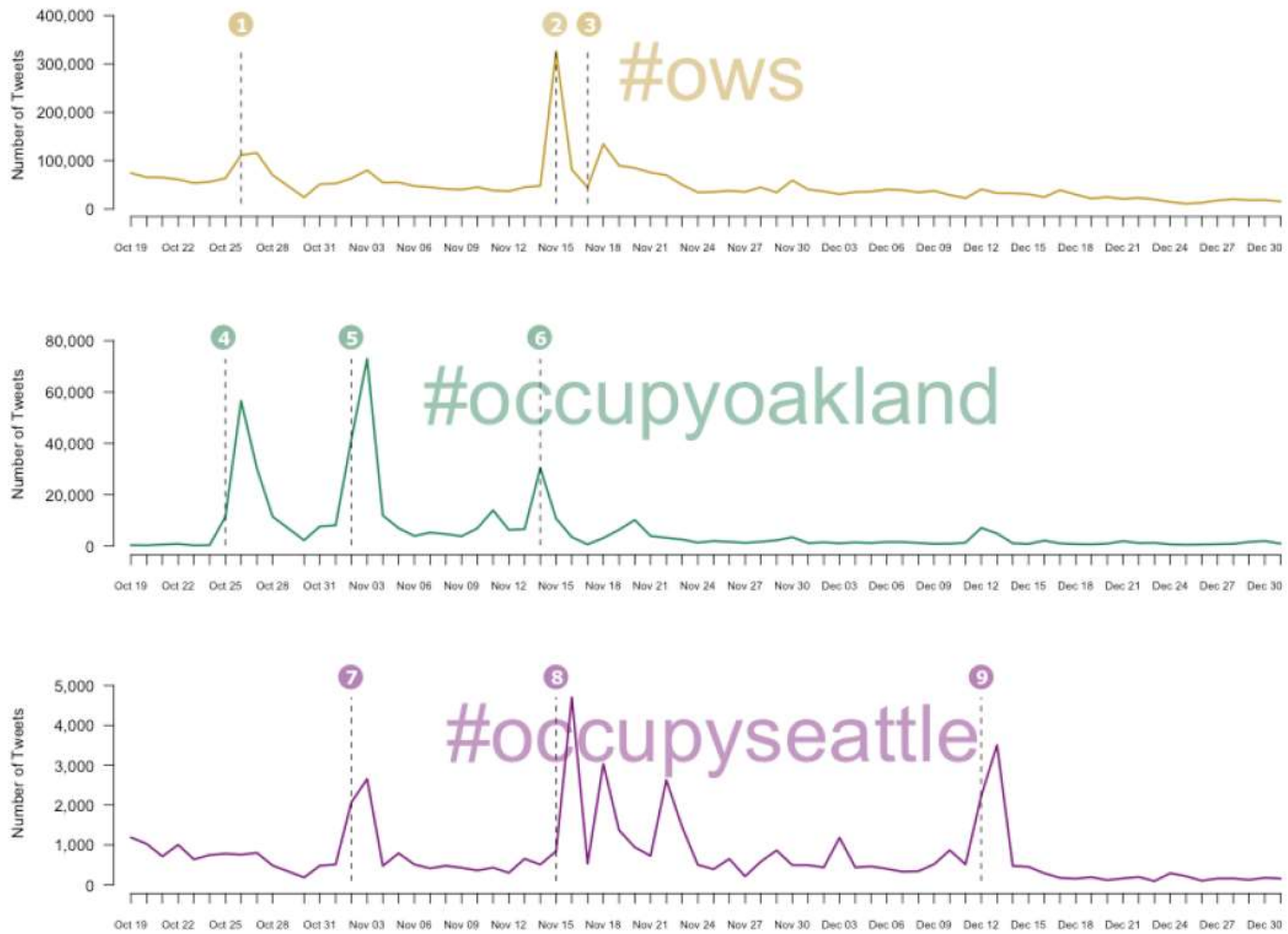
A Model of Crowd-Enabled Organization: Theory and Methods for Understanding the Role of Twitter in the Occupy Protests

SHEETAL D. AGARWAL[1]
W. LANCE BENNETT
COURTNEY N. JOHNSON
SHAWN WALKER
University of Washington, USA

This analysis establishes a conceptual framework, empirical criteria, and measures for deciding when technology-equipped crowd networks such as Occupy Wall Street behave as organizations. The framework is based on three principles that underlie most organizations: (1) resource mobilization; (2) responsiveness to short-term external conditions; and (3) coordinated long-term change, adaptation, or decline. We argue that Twitter played a coordinating role in Occupy as a connector and dynamic switching mechanism linking various networks. We develop methods for tracking how users embedded and shared links to resource locations. Using a database of some 60 million tweets, we examine different types of links distributed through different hashtags across time, showing how Occupy operated along each theoretical dimension as a networked organization.

Keywords: Occupy, Twitter, networks, networked organization, collective action, big data

Introduction



Agarwal, S. D., Bennett, W. L., Johnson, C. N., Walker, S. (2014). A Model of Crowd Enabled Organization: Theory and Methods for Understanding the Role of Twitter in the Occupy Protests. International Journal of Communication, 8(0), 27.

Social-Media-Forschungsdaten

Epistemologische, methodologische, ethische und praktische Probleme, z.B.

- Zusammensetzung aus menschlichen und nicht-menschlichen Nutzenden (z.B. Spam-Bots)
- Kein ‚Informed Consent', nur AGBs
- Sampling durch Plattformbetreiber (z.B. Driscoll/Walker 2014)

Social-Media-Forschungsdaten

- Raw data is an oxymoron (Gitelman 2013): Infrastrukturelle Gegebenheiten, Messinstrumente und Entscheidungen in Datenbereinigung und Analyse prägen die Daten
- Prozesse der Datengenerierung schwer dokumentierbar und daher schwer nachvollziehbar

Social-Media-Forschungsdaten

- Social-Media-Daten bieten keine unmittelbaren Einblicke in die Motivationen und Gedankenwelten der Nutzenden, z.B. komplexe Rolle der Selbstdarstellung im Internet: Nutzer und Avatar sind oft sehr verschieden und unterschiedliche Spielarten der Performanz kommen zur Anwendung (Turkle 1997)

Daten als Prozess, Boundary Object, Projektionsfläche…

- Versprechen von Big Data: übergreifende, allgemeine Erklärungsmuster, ‚paradigmatic change'
- Verschiedene Akteure (Forschende, Software Developer...) und (sozio-technische) Netzwerke (Ranking-Algorithmen, Softwaresysteme…) belegen die Daten mit unterschiedlichen Bedeutungen
- Praktische Aspekte der alltäglichen Datenarbeit (scrape, clean, analyse, share…) zeigen Brüche und subtile Widerständigkeiten

-> Mythologie ‚ubiquitous computing' (Dourish/Bell 2011)

Widerspricht der Idee, dass Big Data – wenn nur die Probleme überwunden werden können – allumfassende Erklärungen liefern wird

Problemstellungen

- Wie können die Aussagen von Social-Media-Forschung überprüft werden?
- Welche (ethischen) Verpflichtungen ergeben sich für die Wissenschaft in Bezug auf Aussagekraft und Nachvollziehbarkeit?

Teilen von Forschungsdaten als Lösung?

- Teilen ermöglicht Qualitätskontrolle, Peer Review, Relativierung der Zugangsungleichheiten…
- Aber: Keine informierte Einwilligung, Schwierigkeiten der Anonymisierung
- Daten dürfen aufgrund der AGBs der Plattformbetreiber oft nicht geteilt oder zu Reproduktionszwecken archiviert werden
  - „fog of confusion“ (Thomas und Walport 2008)
  - „grey market“ (Weller/Kinder-Kurlanda 2015)

Datenzugangsmöglichkeiten

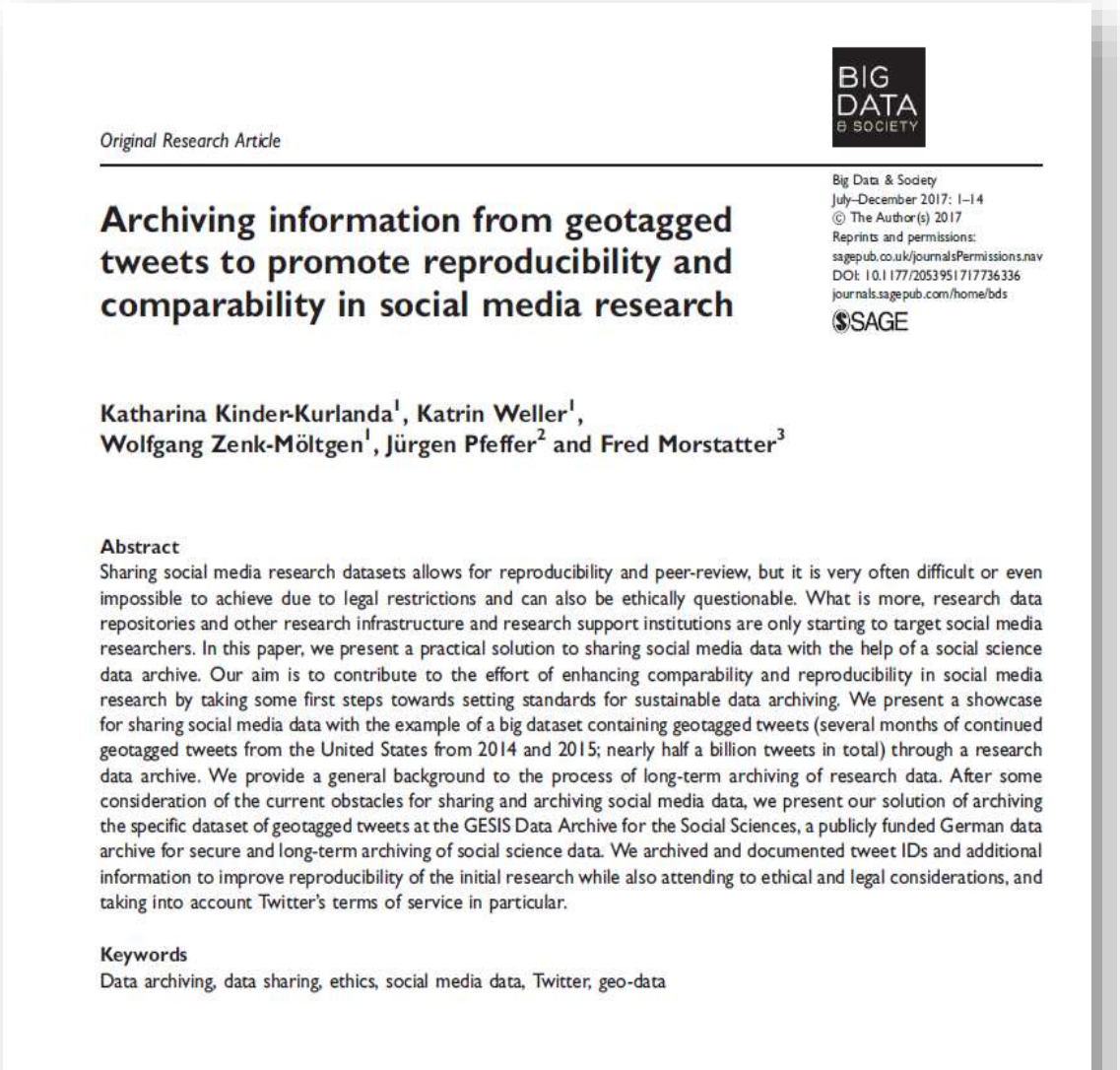
- Eigenes Sammeln öffentlich zugänglicher Inhalte
- Nutzung von für die Forschung bereitgestellten Plattform-APIs
- Kooperationen mit Unternehmen
- Kaufen der Daten von Data Resellers

- -> bestimmen Umfang / Möglichkeiten zum Teilen

gesis Leibniz-Institut für Sozialwissenschaften

Archivierung von Twitter-Daten

- Zugang über API
- Nur Tweet-IDs können archiviert werden
- ‚Rehydration‘
- Umfangreiche Dokumentation und zusätzliche Informationen



Original Research Article

BIG DATA & SOCIETY

Big Data & Society
July–December 2017: 1–14

DOI: 10.1177/2053951717736336
journals.sagepub.com/home/bds
SAGE

Archiving information from geotagged tweets to promote reproducibility and comparability in social media research

Katharina Kinder-Kurlanda[1], Katrin Weller[1], Wolfgang Zenk-Möltgen[1], Jürgen Pfeffer[2] and Fred Morstatter[3]

Abstract

Sharing social media research datasets allows for reproducibility and peer-review, but it is very often difficult or even impossible to achieve due to legal restrictions and can also be ethically questionable. What is more, research data repositories and other research infrastructure and research support institutions are only starting to target social media researchers. In this paper, we present a practical solution to sharing social media data with the help of a social science data archive. Our aim is to contribute to the effort of enhancing comparability and reproducibility in social media research by taking some first steps towards setting standards for sustainable data archiving. We present a showcase for sharing social media data with the example of a big dataset containing geotagged tweets (several months of continued geotagged tweets from the United States from 2014 and 2015; nearly half a billion tweets in total) through a research data archive. We provide a general background to the process of long-term archiving of research data. After some consideration of the current obstacles for sharing and archiving social media data, we present our solution of archiving the specific dataset of geotagged tweets at the GESIS Data Archive for the Social Sciences, a publicly funded German data archive for secure and long-term archiving of social science data. We archived and documented tweet IDs and additional information to improve reproducibility of the initial research while also attending to ethical and legal considerations, and taking into account Twitter's terms of service in particular.

Keywords

Data archiving, data sharing, ethics, social media data, Twitter, geo-data

Ungelöste Probleme

- Keine informierte Einwilligung
- Unzureichende Werkzeuge der Dokumentation
- Abhängigkeit von der API für die Rehydration
- ‚Decay' der Daten

Neuformation datenökonomischer Konstellationen

- Proprietäre Daten als Ware (Leonelli 2014) haltende Social-Media-Plattformbetreiber oder andere Internetfirmen
- Forschung in Unternehmen lässt in bislang nicht gekanntem Umfang weitreichende technische Entwicklungen großenteils in den Händen der Privatwirtschaft (Ekbia 2016)

Neuformation datenökonomischer Konstellationen

- Wissenschaftliche Institutionen als klassische Produzenten des „Wissens über Gesellschaft“ die ethisch und institutionell auf politische Neutralität und scientific best practice verpflichtet sind
- Nutzende, aber auch Nicht-Nutzende, die Gegenstand von Analysen und von den Verschiebungen der epistemischen Gewichte betroffen sind

Alltägliche Praktiken in der Datenökonomie

- Ungleichheiten in Bezug auf die Menge und Qualität an Social-Media-Daten zu denen Forschende Zugang haben
- Finanzielle Mittel und institutionelle Kollaborationen mit der Industrie ermöglichen besseren Datenzugang als APIs und Crawlen
- Chancenungleichheiten, aber: Unterschiede graduell: persönliche Netzwerke, geografische Verortung (z.B. Datenschutzgesetzgebungen)
- Datenökonomie als Rahmung individuell stark unterschiedlicher Möglichkeiten und Praktiken

Fazit

- Überprüfung der Validität der Aussagen in Social-Media-Daten-basierter Forschung ist durch den eingeschränkten Datenzugang begrenzt

Öffentliche Datenarchive als *Trusted Third Parties*

„…what is necessary is an acknowledgment by the platforms that researchers have a legitimate need to share their data with each other, in a controlled and ethical way, and that the best way to do so is openly and transparently through the carefully controlled data repositories that already serve other academic fields” *(Bruns & AoIR 2018)*

Öffentliche Datenarchive als *Trusted Third Parties*

- Verfügen über die entsprechenden Erfahrungen und Strukturen, um diese Rolle einzunehmen:
  - öffentlich finanziert
  - Datenmanagement-Kompetenzen
  - Vielfältige Lösungen für kontrollierte und gesicherte Zugangswege

Öffentliche Datenarchive als *Trusted Third Parties*

- Würden Anbieter die durch Nutzung ihrer Angebote generierten Daten der Gesellschaft in den Infrastrukturen solcher *Trusted Third Parties* „einlagern“, die diese Daten treuhänderisch verwalten, so könnten daraufhin im Rahmen öffentlicher Aushandlungsprozesse Spielregeln sowohl für Zugänge als auch für Nutzungsweisen formuliert werden