Der Event Crawl als Ansatz für den Aufbau von Webarchiven am Beispiel von politischen Wahlkämpfen

Eckl, Markus
markus.eckl@uni-passau.de
Uni Passau, Deutschland

Gassner, Sebastian
Sebastian.Gassner@uni-passau.de
Uni Passau, Deutschland

Die vielfältigen internationalen Aktivitäten im Handlungsfeld Webarchivierung zeigen, dass der Aufgabe, die Inhalte im Web als eine neue Form von Quellenmaterial für die Wissenschaft dauerhaft zu sichern und zugänglich zu halten, in den Gedächtnis- und Forschungsinstitutionen inzwischen eine große Bedeutung beigemessen wird (Aubry 2010). So nimmt der Aufbau und die Erforschung von Webarchiven auch in den Sozial- und Geisteswissenschaften eine immer wichtigere Rolle ein (Brügger 2012; Milligan et al. 2019). Zu konstatieren ist aber, dass sowohl die Methoden ihrer Erzeugung, sowie die Analyse fachwissenschaftlicher Fragestellungen noch nicht hinreichend erforscht sind.

Politische Wahlkämpfe sind beispielsweise nicht nur für die Politikwissenschaft, sondern auch für andere sozial- und geisteswissenschaftliche Disziplinen ein wichtiges Forschungsfeld. Angesichts der zunehmenden Verlagerung von Wahlkämpfen auf virtuelle Arenen ergeben sich methodische Herausforderungen, wie diese politischen Diskurse beobachtet und analysiert werden können. So steigt der Umfang von zu untersuchenden Inhalten und Diskursen enorm an und herkömmliche, qualitative Analysemethoden stoßen an ihre Grenzen. Zum anderen verschwinden nach einiger Zeit immer mehr relevante Inhalte im Web und können nicht mehr abgerufen werden. In diesem Zusammenhang wird der Aufbau und die Nutzung von Webarchiven immer bedeutender, wobei angesichts des diachronen Verlaufs von Wahlkämpfen auch spezifische Strategien des Web-Crawlings notwendig sind (Eckl & Rehbein 2018). Der Event Crawl, als ein möglicher Ansatz der Webarchivierung, kann diese besonderen Anforderungen nicht nur berücksichtigen, sondern er ermöglicht auch die Archivierung digitaler Diskurslandschaften von Ereignissen (Brügger 2012, Rogers 2019).

Das Poster möchte auf Grundlage von zwei durchgeführten Event Crawls, dem bayerischen Landtagswahlkampf 2018 und dem Europawahlkampf 2019, die methodischen Herausforderungen und deren Lösungsansätze darlegen und die Möglichkeiten hinsichtlich der Analyse dieses Webarchives mit Methoden der Digital Humanities diskutieren. Bei beiden Wahlkämpfen wurden Webseiten von Medienhäusern, Parteien und Politikern (+social media accounts) mehrfach gecrawlt und daraus ein Archiv mit einem Umfang von mehr als 4 TB aufgebaut.

In Abgrenzung zum Web Scraping einerseits, bei dem Inhalte von möglichst vielen Webseiten automatisiert gecrawlt werden, und dem Selektiven Crawling andererseits, bei dem eine sehr begrenzte Anzahl von Webseiten gecrawlt werden, besteht hinsichtlich des Event Crawls die Herausforderung, die Grenzen des Events und somit die relevanten Webinhalte zu bestimmen. Durch eine sachliche, zeitliche und akteurszentrierte Eingrenzung des Gegenstandes ist es möglich, relevante Webseiten und -inhalte zu bestimmen und dieses Vorgehen kann einen wichtigen Beitrag hinsichtlich der Diskussion über die Vollständigkeit von Webarchiven liefern (Weber & Napoli 2018). Wir teilen hier die Auffassung von Brügger (2018), dass Webarchive in der Regel unvollständig sind und eine hohe Selektivität aufweisen. Auch wenn durch unsere Eingrenzungen sich kein vollständiges Webarchiv aufbauen lässt, ist es dadurch dennoch möglich, approximativ diesem Ziel näher zu kommen. Vielmehr soll durch die gewählte Abgrenzung des Untersuchungsgegenstandes, die zentralsten Diskurse der beiden Wahlkämpfe erfasst werden. Im Gegensatz zu manch anderen Webarchiven, die eine hohe Selektivität aufweisen, nicht zuletzt aufgrund anderer Crawlmethoden, sind hier vielversprechende Ergebnisse zu erwarten.

Die zweite Herausforderung ergibt sich hinsichtlich der zeitlichen Taktung der durchgeführten Crawls. Denn durch die diachrone zeitliche Entwicklung von Wahlkämpfen, wie zum Beispiel dem diachronen Posten von Inhalten auf Blogs oder Medienwebsteinen, sowie dem Verschwinden und dem Löschen von Webinhalten noch während des Beobachtungszeitraums, muss ein und dieselbe Webseite mehrfach gecrawlt werden. Zu diskutieren ist, welche Taktungen für welche Webseiten notwendig sind und inwieweit sich durch unterschiedliche Vorgehensweisen die erstellten Korpora unterscheiden. Es muss auch die Frage geklärt werden, ob gesamte Webseiten oder nur relevante Inhalte der Webseiten häufiger zu crawlen sind. Neben ökonomisch und technisch beschränkten Mitteln ist hier ein Vorgehen zu wählen, welches sich auch an mögliche fachdisziplinäre Forschungsfragestellungen orientiert.

Ebenfalls soll die Schnittstellen von WARC Dateien diskutiert werden, die genutzt werden können, um analysefähige Korpora zu erstellen. Eine WARC ist ein genormtes Archivformat, das die Inhalte der gecrawlten Webseiten, sowie Metadaten zu den spezifischen Crawls enthält. Dieser Vorgang ist von großer Bedeutung, da dabei nicht nur die WARC Datei entpackt wird, sondern es findet auch eine Filterung, Gruppierung und

Extraktion der Daten statt. Auch wenn dafür bereits einige Programme entwickelt wurden, wie zum Beispiel “ArchivSpark” (Holzmann, Goel & Anand 2017) oder das “Archive Unleashed Toolkit” (Lin et al. 2017), braucht es zum Beispiel für den automatisierten Aufbau eines hochwertigen Textkorpus mit Metadaten aus den Webseiten, eine an den spezifischen Webseiten orientierte Extraktion der Textinhalte. Diese Anforderung ergibt sich, weil Webseiten von verschiedenen Quellen in ihrem Aufbau unterschiedlich sind. Zusätzlich können Webseiten im Laufe der Zeit ihr Aussehen verändern, wodurch in einem Webarchiv für eine Quelle mehrere Layouts enthalten sein können. Wurde nun bei der Verarbeitung einer relevanten Webseite ein Layout identifiziert, wird die Position der gewünschten Daten mit Hilfe von sogenannten *CSS Paths* spezifiziert und die Extraktion kann erfolgen. Nach der Extraktion werden die Daten in einer Mongodb Datenbank zur weiteren Verarbeitung abgelegt.

Nach den methodologischen Überlegungen hinsichtlich des Aufbaus eines Webarchievs, der Beschreibung des Event Crawls und der Erstellung einer MongoDB Datenbank mit Metadaten aus den WARC Dateien (wie z.B. Datum, Überschrift, Verfasser),

ist es auch unter Zuhilfenahme von Methoden der Digital Humanities nun möglich, Archive auf Basis fachwissenschaftlicher Fragestellungen zu untersuchen. Auf Grundlage des beschriebenen Archivs zum Europawahlkampf 2018 können unterschiedliche politikwissenschaftliche Forschungsfragen gestellt werden. Exemplarisch kann untersucht werden, welche Themen auf den Parteienwebsetien im Rahmen des Europawahlkampf 2018 diskutiert wurden. Weiter kann danach gefragt werden, wie die Themenkonjunktur im Laufe des Wahlkampfes war? Eine solche fachwissenschaftliche Fragestellung kann untersuchen, inwieweit die Europawahl 2018 als eine “second order election” zu verstehen ist (Hix, S. & Marsh 2011, Weber 2009). Darunter versteht man den Sachverhalt, dass in europäische Wahnen häufig nationale und nicht europäische Themen im Wahlkampf diskutiert werden. Um diese Fragestellung zu bearbeiten wurde aus der MongoDB Datenbank ein spezifischer Textkorpus mit zusätzlichen Metadaten aus den jeweiligen Webseiten erstellt. Für die Untersuchung und Ermittlung von Wahlkampfthemen fanden Methoden des Topic Modelings Anwendung, wobei hierfür das R Package “Structural Topic Modeling” von Roberts et al. (2019) genutzt wurde. Für weitere Analysen wurde zudem unter anderem auf Methoden der Netzwerkanalyse zurückgegriffen. Erste Ergebnisse können auf GitHub eingeshen werden (https://github.com/MarkusEckl/web_archive_and_stm).

Bibliographie

Aubry, Sara (2010): *Introducing Web Archives as a New Library Service: the Experience of the National Library of France*, in: *LIBER Quarterly*, 20(2), S. 179–199. DOI: http://doi.org/10.18352/lq.7987

Brügger, Niels (2012): *Historical Network Analysis of the Web*in: Social Science Computer Review 31 (3), 306–321.DOI: https://doi.org/10.1177/0894439312454267

Eckl, Markus / Rehbein, Malte (2018): Methoden der Digital Humanities in Anwendung für den Aufbau und die Nutzung von Webarchiven. Konferenz zur Bewahrung digitalen kulturellen Erbes. Deutsche Nationalbibliothek. Frankfurt am Main. https://www.dnb.de/SharedDocs/Downloads/DE/Kulturell/konferenzeDigKultErbe2018_MarkusEckl.html?nn=56454 (letzter Zugriff 12. September 2019).

Holzmann, Helge / Goel, Vinay / Anand, Avishek (2016): *ArchiveSpark*in: Nabil R. Adam / Boots Cassel / Yelena Yesha / Richard Furuta / Michele C. Weigle (Hg.): JCDL'16. Proceedings of the 16th ACM/IEEE-CS Joint Conference on Digital Libraries : June 19-23, 2016, Newark, NJ, USA. DOI: https://doi.org/10.1145/2910896.2910902

Lin, Jimmy / Milligan, Ian / Wiebe, Jeremy / Zhou, Alice (2017): *Warcbase* in: Journal of Cultural Heritage 10 (4), 1–30.

Roberts, Molly E., / Stewart, Brandon M. / Tingley, Dustin (2019): stm: An R Package for Structural Topic Models. *Journal of Statistical Software*,**91** (2), 1–40. DOI: 10.18637/jss.v091.i02 .

Rogers, Richard (2019): „Periodizing Web Archiving: Biographical, Event-Based, National and Autobiographical Traditions“ in: Brügger, Niels / Milligan Ian (Hg.): *The SAGA Handbook of Web History*. London: SAGE.

Weber, Matthew S. / Napoli, Philip M. (2018): *Journalism History, Web Archives, and New Methods for Understanding the Evolution of Digital Journalism* in: Digital Journalism, 6:9, 1186-1205. DOI: https://doi.org/10.1080/21670811.2018.1510293