FORUM

„Let's archive!" Die Dokumentation internetbasierter Daten als neue Herausforderung für die europäische Integrationsforschung

Daniel Göler und Florence Reiter[*]

Web-based data collections are having a growing impact on European integration research. However, analysing this type of data is becoming increasingly challenging for researchers. A so-called third methodological level can be identified, namely the field of data archiving, which is currently hardly mentioned in the methodological debate. Thus, researchers must deal with web archiving and its technological possibilities and limitations if they want to base their work on web-based data. This is important in order to ensure replicability and reliability while collecting, storing and archiving web-based data, which cannot be covered by traditional methods.

Die Digitalisierung der Kommunikationswege und -strukturen hat die europäische Integrationsforschung in den letzten Jahren nachhaltig beeinflusst. Die Anzahl der wissenschaftlichen Abhandlungen über die Auswirkungen dieser Veränderungen auf Wirtschaft, Gesellschaft und Politik in der Europäischen Union (EU) ist gestiegen, wie etwa Untersuchungen über neue Formen der politischen Kommunikation im Europawahlkampf 2019 zeigen.[1] Im Vergleich zu anderen Teildisziplinen der Politikwissenschaft gibt es in der europäischen Integrationsforschung zwei weitere Besonderheiten. Zum einen hat auf theoretischer Ebene die Debatte über den Postfunktionalismus zu einer stärkeren Hinwendung der Forschung zu Fragen von kultureller Identität und Einstellungen zum europäischen Integrationsprozess geführt,[2] die eine Auseinandersetzung mit neuen Medien als Untersuchungsgegenstand zwingend machen. Zum anderen ist auf empirischer Ebene im letzten Jahrzehnt ein Anstieg der Anzahl populistischer EU-skeptischer Parteien zu verzeichnen, die sich stärker als etablierte Parteien in ihren Kommunikationsstrategien sozialer Medien

* Univ.-Prof. Dr. Daniel Göler, Inhaber des Jean-Monnet-Lehrstuhls für Europäische Politik, Universität Passau. Florence Reiter, Wissenschaftliche Mitarbeiterin, Jean-Monnet-Lehrstuhl für Europäische Politik sowie Lehrstuhl für Digital Humanities, Universität Passau.

Der Beitrag entstand im Rahmen des DFG-Projekts „Methoden der Digital Humanities in Anwendung für den Aufbau und die Nutzung von Webarchiven", das in Kooperation zwischen der Bayerischen Staatsbibliothek sowie dem Lehrstuhl für Digital Humanities und dem Jean-Monnet-Lehrstuhl für Europäische Politik der Universität Passau durchgeführt wird.

1 Vgl. Martin Fuchs/Josef Holnburger: #ep2019 – Die digitalen Parteistrategien zur Europawahl 2019 (Kurzstudie der Friedrich-Ebert-Stiftung), Hamburg/Berlin 2019; Chiara Valentini: Social media use by main EU political parties during EP elections 2019, in: Niklas Bolin/Kajsa Falasca/Marie Grusell/Lars Nord (Hrsg.): Euroflections, Sundsvall 2019, S. 80f.

2 Vgl. Liesbet Hooghe/Gary Marks: Grand theories of European integration in the twenty-first century, in: Journal of European Public Policy 8/2019, S. 1113ff.; Liesbet Hooghe/Gary Marks: A Postfunctionalist Theory of European Integration. From Permissive Consensus to Constraining Dissensus, in: British Journal of Political Science 1/2009, S. 1ff.

DOI: 10.5771/0720-5120-2019-4-321

bedienen[3] und deren Wählerschaft häufig das Internet als primäre Informationsquelle nutzt.[4]

Bei der Analyse dieser Entwicklungen haben es WissenschaftlerInnen in der EU-Forschung nicht nur mit neuen Phänomenen zu tun, sondern auch mit einer neuen Art von Daten. Dabei lässt die Einbeziehung der „digitalen Empirie"[5] die Datenmenge, die verarbeitet werden muss, stark ansteigen. Die hieraus entstehenden methodischen Probleme werden schon seit einigen Jahren in den Sozialwissenschaften diskutiert[6] und sind seit Kurzem Gegenstand intensiver Debatten der Teilgebiete der Internationalen Beziehungen und der Vergleichenden Regionalismusforschung[7] und damit auch der EU-Forschung. Vor allem die Herausforderungen der teilweise so bezeichneten „digitale[n] Revolution in den Sozialwissenschaften"[8] für die Datenerhebung und Datenauswertung sind hierbei intensiv diskutiert worden.[9] Der vorliegende Beitrag ergänzt diese beiden Punkte um einen weiteren Aspekt: die Datenarchivierung bzw. Datenkonservierung als „dritte methodische Ebene", der sich ForscherInnen auf dem Gebiet der europäischen Integration verstärkt annehmen müssen. Denn neben den bereits genannten Problemen haben internetbasierte Daten häufig eine Eigenschaft, die eine besondere Herausforderung für die wissenschaftliche Auseinandersetzung mit ihnen darstellt: die – im Vergleich zu klassischen Daten – hohe Fluidität. Bezieht man z.B. in der Analyse eines Europawahlkampfs neben klassischen Printmedien ebenfalls die Online-Angebote von Zeitungen und Rundfunkanstalten ein, so steht man vor dem Problem, dass einzelne Beiträge teils mehrmals täglich aktualisiert bzw. „weitergeschrieben" werden. Bei der Auswertung von Social-Media-Seiten einzelner EuropapolitikerInnen oder von Parteien kann man zudem in der Regel nicht sämtliche Tweets und Posts einsehen und damit in die Analyse einbeziehen, denn durch unterschiedliche Abfragezeitpunkte und Filteralgorithmen werden immer nur selektive (und sich verändernde) Ausschnitte der Gesamtdatenmenge angezeigt. Hinzu kommt die „Besonderheit der Sozialen Medien […], dass sie auf den Beiträgen der Benutzer *(user generated content)* basieren und damit auch Informationen […] enthalten, die kaum professionell oder institutionell gefiltert"[10] und nicht systematisch archiviert bzw. erhalten werden. Auch die Möglichkeit der Korrektur und Löschung von Posts oder gar ganzer Seiten stellt für ForscherInnen ein Problem dar. Ein Beispiel für Letzteres ist etwa die Entscheidung des Parteivorsit-

3 Vgl. Max Schaub/Davide Morisi: Voter mobilization in the echo chamber: Broadband internet and the rise of populism in Europe, Collegio Carlo Alberto: CAN Research Paper 584/2019.
4 Vgl. Nicola Maggini: Understanding the Electoral Rise of the Five Star Movement in Italy, in: Czech Journal of Political Science 1/2014, S. 37ff., hier S. 57.
5 Sebastian Knecht/Maria J. Debre: Die „digitale IO": Chancen und Risiken von Online-Daten für die Forschung zu Internationalen Organisationen, in: Zeitschrift für Internationale Beziehungen 1/2018, S. 175ff., hier S. 175.
6 Vgl. u.a. David A. Karpf: Social Science Research Methods in Internet Time, in: Information, Communication & Society 5/2012, S. 639ff.; W. Lance Bennett: The Personalization of Politics: Political Identity, Social Media, and Changing Patterns of Participation, in: The ANNALS of the American Academy of Political and Social Science 1/2012, S. 20ff.
7 Vgl. Knecht/Debre: Die „digitale IO", 2018.
8 Carolin Kaiser: Soziale Medien als Mittel der Produktgestaltung (Co-Creation), in: Christian König/Matthias Stahl/Erich Wiegand (Hrsg.): Soziale Medien. Gegenstand und Instrument der Forschung, Wiesbaden 2014, S. 171ff., hier S. 185.
9 Vgl. Markus Strohmaier/Maria Zens: Analyse Sozialer Medien an der Schnittstelle zwischen Informatik und Sozialwissenschaften, in: Christian König/Matthias Stahl/Erich Wiegand (Hrsg.): Soziale Medien. Gegenstand und Instrument der Forschung, Wiesbaden 2014, S. 73ff.
10 Ebenda, S. 73.

zenden der Grünen, Robert Habeck, seine Accounts bei Facebook und Twitter zu löschen.[11] Diese Fluidität internetbasierter Daten stellt aber ein sehr grundsätzliches Problem hinsichtlich der intersubjektiven Nachvollziehbarkeit sowie der Reliabilität dar und betrifft damit zwei Kernkriterien guten wissenschaftlichen Arbeitens. Denn wenn die einer Analyse zugrunde liegenden Daten nicht für eine Überprüfung durch Dritte zur Verfügung stehen, ist ein Kernelement der Wissenschaftlichkeit aufgehoben. Für eine Analyse etwa des Europawahlkampfs 2019 unter Einbeziehung von Social-Media-Kanälen stellt sich damit nicht nur die Frage, wie WissenschaftlerInnen diese Daten erheben und für sich auswerten, sondern auch, wie sie diese so sichern und dokumentieren, dass ihre Analyse mit zeitlichem Abstand nochmals durchgeführt und nachvollzogen werden kann.

Entsprechend unterstreichen nahezu alle Positionspapiere führender Wissenschaftsorganisationen, dass es unverzichtbar für gute wissenschaftliche Praxis ist, die einer Analyse zugrunde liegenden Primärdaten (zumindest mittelfristig) zu sichern und bei Bedarf anderen ForscherInnen zur Überprüfung zur Verfügung zu stellen. So betont die Hochschulrektorenkonferenz in ihrer Empfehlung „Gute wissenschaftliche Praxis an deutschen Hochschulen": „Jede Wissenschaftlerin und jeder Wissenschaftler ist zur vollständigen Datendokumentation verpflichtet."[12] Ein gemeinsames Positionspapier des Allgemeinen Fakultätentags, der Fakultätentage und des Deutschen Hochschulverbands unterstreicht, dass „Forschungsergebnisse und die ihnen zugrunde liegenden Daten […] ebenso genau dokumentiert werden und überprüfbar sein [müssen …] wie die Interpretationsleistungen und ihre Quellen".[13] Konkretisiert wird diese Dokumentationspflicht in der Denkschrift „Sicherung guter wissenschaftlicher Praxis" der Deutschen Forschungsgemeinschaft (DFG), welche die Anforderung aufstellt, dass „Primärdaten als Grundlagen für Veröffentlichungen […] auf haltbaren und gesicherten Trägern in der Institution, wo sie entstanden sind, zehn Jahre lang aufbewahrt werden"[14] sollen. Diese Zehnjahresfrist nennen ebenfalls die „Regeln zur Sicherung guter wissenschaftlicher Praxis" der Max-Planck-Gesellschaft.[15] Auch der „European Code of Conduct for Research Integrity" der „All European Academies" fordert, dass „[r]esearchers, research institutions and organisations ensure appropriate stewardship and curation of all data […] with secure preservation for a reasonable period".[16]

Datenarchivierung als dritte methodische Ebene in der europäischen Integrationsforschung

Nimmt man die Forderungen nach umfassender Archivierung von Primärdaten ernst, bedeutet dies, dass in der europäischen Integrationsforschung im Rahmen der Auseinandersetzung mit internetbasierten Daten neben der Frage der Datenerhebung und der Da-

11 Vgl. Philipp Saul: Habeck will seine Accounts bei Facebook und Twitter löschen, in: SZ.de, 7. Januar 2019.
12 Hochschulrektorenkonferenz: Gute wissenschaftliche Praxis an deutschen Hochschulen. Empfehlung der 14. Mitgliederversammlung der HRK am 14. Mai 2013 in Nürnberg, S. 3.
13 Gemeinsames Positionspapier des Allgemeinen Fakultätentags (AFT), der Fakultätentage und des Deutschen Hochschulverbands (DHV): Gute wissenschaftliche Praxis für das Verfassen wissenschaftlicher Qualifikationsarbeiten, 9. Juli 2012, S. 2.
14 Deutsche Forschungsgemeinschaft: Sicherung guter wissenschaftlicher Praxis, ergänzte Auflage, Weinheim 2013, S. 21.
15 Vgl. Max-Planck-Gesellschaft: Regeln zur Sicherung guter wissenschaftlicher Praxis, März 2009, S. 4.
16 All European Academies (ALLEA): The European Code of Conduct for Research Integrity, Berlin 2017, S. 6.

tenverarbeitung bzw. -auswertung, die in der aktuellen Methodendebatte bereits intensiv diskutiert werden,[17] die Datenarchivierung bzw. -sicherung als dritte methodische Ebene unverzichtbar wird. Nur so können intersubjektive Nachvollziehbarkeit sowie Reliabilität von Forschungsarbeiten sichergestellt werden. Die Herausforderungen, die sich hieraus ergeben, werden in den verschiedenen Ansätzen zum Aufbau von Webarchiven und den jeweils eingesetzten Webarchivierungstools deutlich. Diese Tools dienen dazu, internetbasierte Daten so zu sichern und zu dokumentieren, dass diese auch mit größerem zeitlichen Abstand noch für wissenschaftliche Analysen – in der Regel mit Methoden aus dem Bereich der Digital Humanities – herangezogen werden können.

Chancen und Grenzen der Webarchivierung für die wissenschaftliche Arbeit im Allgemeinen und die europäische Integrationsforschung im Besonderen zeigen sich in der systematischen Erfassung, Aufbereitung und Archivierung von Websites, Online-Berichten, Social-Media-Debatten und Online-Kommentaren, die etwa für die Analyse gesellschaftlicher Debatten über die EU und ihre Politiken oder aber auch für die Untersuchung von Europawahlkämpfen mittlerweile unverzichtbar sind. Dabei muss die Archivierung zum einen so erfolgen, dass die Datenkorpora mit zeitlichem Abstand und durch Dritte nutzbar sind, sodass die intersubjektive Nachvollziehbarkeit sowie die Reliabilität der Analyse gegeben sind. Zum anderen sollten die Daten in einer solchen Form archiviert werden, dass sie auch mit computergestützten Methoden der Digital Humanities auswertbar sind. Zu nennen wären hier insbesondere Text-Mining-Verfahren wie Topic Modeling, Word2Vector oder Sentiment-Analyse. Text-Mining-Verfahren sind „computergestützte Verfahren für die semantische Analyse von Texten […], welche die automatische bzw. semi-automatische Strukturierung von Texten, insbesondere sehr großen Mengen von Texten, unterstützen".[18] Sie bieten somit ein vielfältiges Methodenspektrum, um große Textmengen nach bestimmten Fragestellungen zu untersuchen. Die Chancen, die sich aus solchen computergestützten Methoden für die europäische Integrationsforschung ergeben, sind vielfältig und eröffnen für die Analyse der gesellschaftspolitischen Grundlagen des Integrationsprozesses neue Forschungsperspektiven. Gerade deshalb ist es aber unverzichtbar, die hierfür im Internet zu findenden relevanten Daten zuverlässig und nachhaltig zu archivieren.

Probleme der Nutzung bestehender Webarchive für die Integrationsforschung

Während Webarchivierung in der europäischen Integrationsforschung noch eine untergeordnete Rolle spielt, ist die Entwicklung in anderen Bereichen – wie beispielsweise in den Digital Humanities oder bei Gedächtnisinstitutionen wie Archiven und Bibliotheken – deutlich fortgeschrittener. Durch die Archivierung von Websites können ForscherInnen einerseits die Entwicklung des Internets dokumentieren und analysieren. Andererseits ermöglicht Webarchivierung, „to document our findings when we study today's web, since in practice most web studies preserve the web in order to have a stable object to study and

17 Vgl. Jasmin Haunschild/Anja P. Jakobi: „Big Data" oder „Dunkelziffer"? – Wie Studierende aus schwieriger Datenlage lernen können, in: Zeitschrift für Internationale Beziehungen 1/2018, S. 221ff., hier S. 221.

18 Gerhard Heyer/Uwe Quasthoff/Thomas Wittig: Text Mining: Wissensrohstoff Text. Konzepte, Algorithmen, Ergebnisse, Bochum 2006, S. 3.

refer to when the analysis is to be documented (except for studies of the live web)".[19] Webarchivierung als Prozess und Webarchive als Datengrundlage für Analysen sind somit nicht nur für WebhistorikerInnen, sondern durch die zunehmende Verlagerung von gesellschaftlichen und politischen Kommunikationsprozessen in das Internet auch für andere Forschungsbereiche unverzichtbar. In der europäischen Integrationsforschung gilt dies besonders im Hinblick auf die Untersuchung von Identitätsfragen und die diesbezüglichen politischen und gesellschaftlichen Debatten, die sich zu einem Großteil auf Social-Media-Plattformen abspielen. Vor allem seit der sogenannten Migrationskrise wird diesen Fragen in der Integrationsforschung ein großes Gewicht beigemessen, insbesondere aus postfunktionalistischer Perspektive. So betonen Liesbet Hooghe und Gary Marks: „Postfunctionalism puts the spotlight on identity politics [… and] shows how the migration crisis has intensified a cultural divide across Europe that pits proponents of a multicultural, open, Europe against its opponents."[20] Eine Integrationsforschung, die diese Entwicklungen ohne Berücksichtigung von internetbasierten Daten untersucht, würde einen Teil der heutigen gesellschaftlichen Realität systematisch ausblenden.

Trotz der gewachsenen Bedeutung internetbasierter Daten und zahlreicher Aktivitäten der Webarchivierung gelten Webarchive als „unknown, and certainly underused, primary source".[21] Dies liegt unter anderem daran, dass die ständige Weiterentwicklung von Webtechnologien und das dadurch entstehende dynamische Umfeld der Webarchivierung archivierende Institutionen laufend vor neue Herausforderungen stellen.[22] Denn neben den verschiedensten Arten von Textdateien – dazu zählen HTML-Sites, Word-Dokumente oder PDF-Dokumente – beinhalten Websites unterschiedlichste Video-, Bild- und Sound-Dateien, animierte GIFs oder auch eingebettete Bild- und Videodateien von Social-Media-Plattformen wie YouTube, Instagram oder Facebook. Diese vielfältigen Datenarten sollten im idealtypischen Archivierungsprozess berücksichtigt werden, was aber in der Realität noch schwer umsetzbar ist.[23] Denn die Vielfalt dieser Daten übersteigt derzeit die Möglichkeiten der Webarchivierungstools.[24]

Zwar arbeiten verschiedene Institutionen schon seit einigen Jahren daran, Webinhalte als Datenmaterial für die Wissenschaft zu sichern und zugänglich zu machen; die Nutzbarkeit für die Wissenschaft ist jedoch teils problematisch. Neben zum Teil eingeschränkten Zugangs- und Nutzungsmöglichkeiten dieser Webarchive besteht das Hauptproblem in deren Unvollständigkeit. Selbst das „Internet Archive",[25] das als kostenlose digitale Bibliothek seit 1996 Websites archiviert – mittlerweile sind es über 330 Milliarden[26] – und über die

19 Niels Brügger: Web Archiving – Between Past, Present, and Future, in: Mia Consalvo/Charles Ess (Hrsg.): The Handbook of Internet Studies, Malden/Oxford/Chichester 2011, S. 24ff., hier S. 24.
20 Hooghe/Marks: Grand theories, 2019, S. 1122.
21 Jane Winters: Coda: Web archives for humanities research – some reflections, in: Niels Brügger/Ralph Schroeder (Hrsg.): The Web as History. Using Web Archives to Understand the Past and the Present, London 2017, S. 238ff., hier S. 238.
22 Vgl. Nicholas Taylor: Reflections on the 2016 IIPC General Assembly and Web Archiving Conference, in: Stanford Libraries, 12. Mai 2016.
23 Vgl. ebenda.
24 Vgl. Winters: Coda, 2017, S. 243.
25 Internet Archive, abrufbar unter: https://archive.org/ (letzter Zugriff: 14.10.2019).
26 Vgl. Internet Archive: About the Internet Archive, abrufbar unter: https://archive.org/about/ (letzter Zugriff: 14.10.2019).

Serveranwendung „Wayback Machine“[27] online zur Verfügung stellt, weist erhebliche Lücken auf. So hat sich in einem DFG-Projekt zum Aufbau von Webarchiven am Beispiel der Europawahl 2019[28] gezeigt, dass das „Internet Archive“ etwa vom Tag der Europawahl in Deutschland, dem 26. Mai 2019, und dem Vortag keine Archivierung der Website der ZDF-Nachrichtensendung „heute“ vorgenommen hat.[29] Versucht man die archivierte Website der Frankfurter Allgemeinen Zeitung vom deutschen Wahltag und dem Vortag zu öffnen, erscheint eine Fehlermeldung.[30] Darüber hinaus sind wichtige Foren der politischen Kommunikation wie Social-Media-Plattformen nicht vollständig archiviert. Für das wissenschaftliche Arbeiten bedeutet dies, dass Studien, die sich nur auf bestehende Webarchive stützen, immer der Gefahr einer unzulänglichen Datengrundlage ausgesetzt sind, zumal meist nicht ersichtlich ist, welche systematischen und technischen Grenzen die Erstellung bestehender Archive beeinflusst haben. Gerade bei aktuellen Fallstudien sind daher eine eigene Datenerhebung und Datenarchivierung unverzichtbar. In der genannten Projektstudie zu den Europawahlen 2019 wurde deshalb versucht, im Rahmen eines sogenannten Event-Crawls Websites und Social-Media-Seiten von Parteien und PolitikerInnen sowie die Online-Berichterstattung von Zeitungen und Nachrichtensendern zu archivieren. „Crawl“ ist dabei die gängige Bezeichnung für den Prozess der Webarchivierung und wird definiert als „the process of building a collection of webpages by starting with an initial set of URLs (or links) and recursively traversing the corresponding pages to find additional links“.[31] Neben dem Crawlen der einzelnen Websites gehört zum Archivierungsprozess idealerweise auch eine Qualitätskontrolle jedes einzelnen Vorgangs, die gewährleisten soll, dass der Archivierungsprozess erfolgreich war und Daten auch in der Zukunft noch verfügbar sind. Dies ist allerdings sehr aufwendig und setzt der Erstellung von allgemeinen Webarchiven Grenzen.

Herausforderungen bei der Erstellung individueller Webarchive

Eine ganze Reihe von Problemen, die sich bei der Erstellung umfassender Webarchive stellen, zeigt sich gleichermaßen bei individuellen bzw. event- und projektbezogenen Archiven. An erster Stelle sind hierbei die technischen Herausforderungen zu nennen. So bieten Webarchivierungstools wie das „Web Curator Tool“ (WCT)[32] oder der „Webrecorder“[33] zwar die Möglichkeit, Websites zu erfassen und zu archivieren; die Anwendung birgt jedoch auch zahlreiche Probleme. Hinzu kommt, dass die Tools ständig weiterentwickelt werden. Bei länger laufenden Event-Crawls stellt sich damit das Problem, dass neue Versionen zwar bessere Ergebnisse erbringen können. Allerdings wird damit auch die Vergleich-

27 Internet Archive: Wayback Machine, abrufbar unter: https://web.archive.org/ (letzter Zugriff: 14.10.2019).
28 Vgl. Universität Passau: Webarchive. DFG-Projekt Webarchive – Internet für die Nachwelt archivieren, abrufbar unter: https://www.uni-passau.de/forschung/forschungsprojekte/webarchive/ (letzter Zugriff: 18.10.2019).
29 Vgl. Internet Archive: Wayback Machine: Suche nach www.zdf.de/nachrichten, abrufbar unter: https://web.archive.org/web/*/zdf.de/nachrichten (letzter Zugriff: 15.10.2019).
30 Vgl. Internet Archive: Wayback Machine: Öffnen eines Crawls von www.faz.net mit Fehlermeldung, abrufbar unter: https://web.archive.org/web/20190525155240/https://www.faz.net/ sowie https://web.archive.org/web/20190526092534/https://www.faz.net/ (letzter Zugriff: 15.10.2019).
31 Gabe Ignatow/Rada F. Mihalcea: An Introduction to Text Mining. Research Design, Data Collection, and Analysis, Los Angeles u.a. 2018, S. 82.
32 Web Curator Tool, abrufbar unter: http://webcurator.sourceforge.net/ (letzter Zugriff: 14.10.2019).
33 Webrecorder, abrufbar unter: https://webrecorder.io/ (letzter Zugriff: 14.10.2019).

barkeit der Daten eingeschränkt, da die Datenerhebung während des Event-Crawls nach unterschiedlichen (technischen) Standards erfolgt. In dem genannten Forschungsprojekt zur Europawahl 2019 verwendete die Bayerische Staatsbibliothek zunächst die von ihr bisher für die Archivierung von Internetquellen genutzte WCT Version 1.6.1. Diese erwies sich während eines Pretests als ungeeignet für die Archivierung von Facebook- und Twitter-Seiten, deren Einbeziehung in die geplante Analyse des Europawahlkampfs 2019 aber unverzichtbar war. Deshalb wurde für Social-Media-Seiten auf das browserbasierte Tool „Webrecoder" zurückgegriffen. Bei diesem ist die Archivierung mittels händischen Scrollens der jeweiligen Website möglich, was jedoch einen hohen Personalaufwand bedeutet. Aber auch bei der Archivierung klassischer Websites kam die WCT Version 1.6.1 an ihre Grenzen, da sie verschlüsselte Websites nicht erfolgreich harvesten[34] konnte. Daher wurde parallel die zu diesem Zeitpunkt noch in der Entwicklung befindliche WCT Beta-Version 1.7 mit einer neueren Version des Heritrix-Crawlers, der die Basis für das WCT bildet, eingesetzt. Da bei dieser Beta-Version immer wieder technische Schwierigkeiten auftraten, konnte der ursprünglich wöchentlich geplante Archivierungs-Rhythmus nicht für alle Websites eingehalten werden.

Für die letztendliche Erfassung der Websites im Umfeld der Europawahlen konnte dann auf die neueste Version des WCT (Version 2.0) zurückgegriffen werden, die sich als deutlich stabiler als die im Pretest verwendete Version erwies. Dieses Beispiel zeigt sehr gut, wie stark Korpusbildung und Datenerfassung im digitalen Bereich durch technische Rahmenbedingungen beeinflusst werden. Denn auch diese neueste Version verfügte nicht über eine unbegrenzte Crawl-Kapazität, was ihre Anwendung erheblich einschränkte, insbesondere hinsichtlich der Frequenz der Datenerhebung. So war es in dem genannten Projekt selbst mit einem in der Datenerfassung und Datenarchivierung erfahrenen und ressourcenstarken Partner wie der Bayerischen Staatsbibliothek nicht möglich, mehr als acht Medien-Websites täglich zu crawlen, wobei hier nochmals eine weitere Beschränkung auf die jeweilige Startseite, Politikseite und Themenseite zur Europawahl erforderlich war.

Neben den Webarchivierungstools birgt auch das Dateiformat zur Archivierung Schwierigkeiten. Denn das in der Webarchivierung mittlerweile standardmäßig verwendete Webarchive-Dateiformat WARC[35] erfordert sowohl bei der Datenarchivierung als auch bei der späteren Nutzung der archivierten Daten vertiefte informationstechnische Kenntnisse, welche in der klassischen geistes- und sozialwissenschaftlichen Methodenausbildung bisher nicht vorkommen. Denn WARC-Dateien können nicht ohne umfangreiche Vorverarbeitungsschritte mit Text-Mining-Verfahren der Digital Humanities analysiert werden. Die einzelnen Dateien enthalten für die Textanalyse großteils irrelevante Daten wie beispielsweise Header und Footer, Werbung, Menüs, Bilder und Videos. Die relevanten HTML-Sites enthalten wiederum zu einem großen Teil Informationen, die nicht benötigt werden (HTML-Tags, Javascript etc.). Um von den WARC-Dateien an den gewünschten (Text-)Inhalt zu gelangen, ist ein mehrstufiger Extraktionsprozess notwendig, der nicht ohne nötige informationstechnische Kenntnisse durchführbar ist. Insgesamt lässt sich festhalten, dass die Komplexität der internetbasierten Daten und die Grenzen der bisher zur Verfügung

34 „Harvesten" kann als Synonym für „crawlen" verwendet werden.

35 WARC steht für „Web ARChive Archivformat". Es handelt sich dabei um ein Verfahren für die Kombination mehrerer digitaler Ressourcen in einer aggregierten Archivdatei mit Metadaten.

stehenden Tools zu ihrer Archivierung dazu führen, dass die Datenerfassung, -aufbereitung und -archivierung sehr ressourcenintensiv und voraussetzungsvoll sind.

Webarchivierung als Teil der Methoden der Sozialwissenschaften

Die vorangegangenen Beispiele unterstreichen, dass die „digitale Revolution" in den Geistes- und Sozialwissenschaften auch die europäische Integrationsforschung vor die Herausforderung stellt, sich mit den technischen Möglichkeiten und Grenzen der systematischen Archivierung von internetbasierten Daten auseinanderzusetzen. Denn mit dem klassischen Methodeninstrumentarium lassen sich internetbasierte Daten für die Forschung nicht in einer Weise erschließen, die den Ansprüchen an intersubjektive Nachvollziehbarkeit und Reliabilität gerecht würde. Hinzu kommt, dass die Beeinflussung des erfass- und archivierbaren Materials durch sich ständig wandelnde technische Möglichkeiten ein verändertes Problembewusstsein für entsprechende Entwicklungen bei VerfasserInnen und RezipientInnen wissenschaftlicher Studien voraussetzt. Von daher ist es dringend geboten, die Sensibilität für Fragen der Webarchivierung und das Verständnis für die komplexen technischen Rahmenbedingungen im Umgang mit internetbasierten Daten zu erhöhen. Dies ist auch eine Aufgabe für die Entwicklung künftiger Curricula an Universitäten, wo die Archivierung und Analyse internetbasierter Daten bisher in der Methodenausbildung europawissenschaftlicher Studiengänge praktisch nicht vorkommen. Der Europaforschung könnte bei der Bewältigung dieser Aufgabe allerdings zugutekommen, dass sie aufgrund ihrer Interdisziplinarität traditionell enge Bezüge zu den Geschichtswissenschaften hat, in denen zumindest das Problembewusstsein für Fragen der Datenarchivierung Teil der Fachidentität ist. Hieran anknüpfend sollten die Europawissenschaften in den nächsten Jahren einen intensiven Diskurs über den Umgang mit und die Archivierung von internetbasierten Daten führen. Denn die durch den Postfunktionalismus angestoßene Hinwendung der EU-Integrationsforschung zu Identitätsfragen sowie die durch den Aufstieg links- und rechtspopulistischer Parteien beschleunigte Verlagerung politischer Kommunikationsprozesse auf Social-Media-Plattformen machen eine verstärkte Einbeziehung internetbasierter Daten in die wissenschaftliche Auseinandersetzung mit dem europäischen Integrationsprozess unverzichtbar. Ebenso unverzichtbar sind daher Fähigkeiten und Techniken zum Umgang mit dieser – für Geistes- und SozialwissenschaftlerInnen – relativ neuen Datenkategorie.