IJDC | *Brief Report*

Data Curation Strategies to Support Responsible Big Social Research and Big Social Data Reuse

Sara Mannheimer
Montana State University

Abstract

Big social research repurposes existing data from online sources such as social media, blogs, or online forums, with a goal of advancing knowledge of human behavior and social phenomena. Big social research also presents an array of challenges that can prevent data sharing and reuse.

This brief report presents an overview of a larger study that aims to understand the data curation implications of big social research to support use and reuse of big social data. The study, which is based in the United States, identifies six key issues relating to big social research and big social data curation through a review of the literature. It then further investigates perceptions and practices relating to these six key issues through semi-structured interviews with big social researchers and data curators.

This report concludes with implications for data curation practice: metadata and documentation, connecting with researchers throughout the research process, data repository services, and advocating for community standards. Supporting responsible practices for using big social data can help scale up social science research, thus enhancing our understanding of human behavior and social phenomena.

Submitted date 2022 ~ *Accepted* date 2022

Correspondence should be addressed to Sara Mannheimer, Montana State University Library, P.O. Box 173320, Bozeman, MT 59717. Email: sara.mannheimer@montana.edu

This paper was presented at International Digital Curation Conference IDCC22, online, 13-16 June, 2022

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

http://dx.doi.org/10.2218/ijdc.v17i1.823
DOI: 10.2218/ijdc.v17i1.823

Introduction

Big social research repurposes existing data from online sources such as social media, blogs, or forums, with a goal of advancing knowledge of human behavior and social phenomena. Big social research also presents an array of challenges. This brief report provides an overview of a larger study which identifies six key issues in big social research: context, data quality and trustworthiness, data comparability, informed consent, privacy and confidentiality, and intellectual property and data ownership. Partly due to these six challenges, big social data are less commonly shared in data repositories than other types of data, and they tend to lie on the periphery of data sharing and data reuse conversations. The larger study investigates perceptions and practices relating to these six key issues through semi-structured interviews with big social researchers and data curators. This brief report summarizes insights from these interviews and describes implications for data curation practice.

Defining Big Social Research

In this paper, I define big social research as research that uses large-scale data collected from online sources (such as social media, blogs, and online forums) to conduct social and behavioral research. Types of big social data include:

- Digital self-representation data: usernames, profile pictures, biographical information
- Social interaction data: timeline posts, online forum posts, content sharing, commenting, direct messaging
- Digital relationships data: Follower/following data, “likes”
- Metadata: Timestamps, geospatial data, type of operating system, type of device, application used to post (adapted from Olshannikova et al., 2017).

Big social research often uses unobtrusive data collection methods – that is, collecting data without directly contacting research participants (Bright, 2017), instead using application programming interfaces (APIs) or web scrapers to collect data from online sources such as social media platforms, online forums, or blogs. While it is possible to use smaller subsets of big social data to conduct traditional qualitative research, big social research is by definition large-scale. Big social researchers generally use computational social science methods, including natural language processing, sentiment analysis, network analysis, artificial intelligence, and deep learning techniques (Bankes et al., 2002; Berkout et al., 2019 ; Mason et al., 2014).

Key Issues for Big Social Research and Big Social Data Reuse

In this paper, I focus on six key issues in big social research: context, data quality and trustworthiness, data comparability, informed consent, privacy and confidentiality, and intellectual property and data ownership. The issues were identified through a review of the literature (Mannheimer, 2021), and further explored through interviews with big social researchers and data curators.

Context

Social media posts are short pieces of text, images, or video, taken from a larger context of personal and public life. This out-of-context effect is compounded when data are amassed at a large scale. In fact, context and meaning may never be accurately understood by big social researchers (Boyd & Crawford, 2012). Big social researchers I interviewed considered the out-of-context effect that could result from aggregating small snippets of text or images, which are part of the broader context social media users' lives and personas (Törnberg & Törnberg, 2018). Some big social researchers also talked about how APIs remove the context of the social media user interface, leaving only the content and metadata (Bruns & Weller, 2016). Some big social researchers also talked about structuring their research design and methods to support clearer context; for example, selecting data that had more inherent context, such as Tweets that included a geographical location tag. A final consideration regarding context is the importance of balancing providing enough contextual information with protecting user privacy.

Data Quality and Trustworthiness

Data quality can be affected by bots and fake accounts; as many as 15% of active Twitter accounts are bots (Varol et al., 2017). Data quality and trustworthiness can also be affected by missing data and sampling issues (Burgess & Bruns, 2012). Representativeness of data is also a key issue and was discussed by big social researchers I interviewed. Researchers collected data from social media platforms where people were discussing topics of interest, but they were aware that the users of any single social media platform are not representative of the population as a whole (Hargittai, 2020). Some researchers have attempted to create more representative datasets by blending big social data with smaller social datasets so as to "include perspectives that are both important to the data yet not necessarily present within it" (Croeser & Highfield, 2020, p. 673). This idea is discussed further in the next section, Data Comparability.

Data Comparability

Comparing and combining data can enhance the context and quality of data. However, this practice was rare for the big social researchers I spoke with. Researchers encounter challenges when trying to match participants across datasets (Stier et al., 2020), and when combining datasets that have used different data collection methods and different sampling methods (Bossetta, 2018; Martí et al., 2019). The wide variety of different filetypes, metadata fields, and metadata standards also make comparing and combining data difficult (Acker & Kriesberg, 2017).

Informed Consent

The large scale of big social research makes it difficult to obtain informed consent from each user. While social media terms of service may include consent to big social research, most users do not read the terms of service so closely as to constitute informed consent (Obar & Oeldorf-Hirsch, 2020). The European Union's General Data Protection Regulation (GDPR) discusses consent in Article 7; however, "it remains questionable whether the GDPR would in practice prevent the common 'click and forget' consent systems common to Internet interfaces" (Schneble et al., 2018). In 2015, the United States Health and Human Services provide guidelines for community focus groups and advisory boards could be a way to reduce harm for user communities (Secretary's Advisory Committee on Human Research Protections, 2015); however, these strategies have not been widely used in practice. None of the big social researchers I interviewed saw informed consent as a major issue for their research; most considered obtaining informed consent impractical and unnecessary.

Privacy and Confidentiality

“Public” and “private” spaces and activities can blur on social media. The theory of contextual integrity (Nissenbaum, 2009) is widely used to consider ideas about privacy online. This theory suggests that users have different expectations of privacy for their personal information, depending on the context. Users’ expectations of privacy and their strategies for protecting their privacy online continually change, depending on a variety of factors, including “physical environment, audience, social status, task or objective, motivation and intention, and … [the] information technologies in use” (Palen & Dourish, 2003, p. 131). Big social researchers interviewed for this study were concerned with protecting the privacy and maintaining confidentiality of the users represented in big social datasets. Researchers considered the privacy expectations of users, deidentified social media posts before publishing, and designed research with privacy in mind. While these actions support user privacy, they may also discourage data sharing; few big social researchers I spoke to had shared their data publicly.

Intellectual Property and Data Ownership

In the United States, intellectual property (IP) on social media is a relatively grey area of the law. Social media companies view big social data as corporate assets, and social media companies have even invoked the Computer Fraud and Abuse Act to try to prevent web scraping (Neuburger, 2021). Social media companies may also limit the data provided with APIs, and may implement policies to prevent sharing data that have previously been collected from the platform. For example, Twitter only allows publishing lists of Tweet IDs, which may be “rehydrated” later (Summers, 2017). Most researchers reported following social media terms of service, but some also made calculated decisions to bend them. Researchers also reported purchasing commercially available big social datasets to avoid IP concerns.

Implications for Data Curation Practice

This research suggests several tools, services, and strategies that data curators can use to help support responsible big social research and big social data reuse.

Metadata and Documentation

Metadata and description can be used to communicate context and data quality to future users, and standardized metadata and file formats can support data aggregation and comparability. When sharing big social datasets, data curators can ask researchers to include: information about communities and research participants; information about research questions and research methods; explanation of data collection, cleaning, analysis; documentation of potential errors, bias, and missing data; and related materials such as software, code, and related article DOIs. Data curators can also support use of metadata standards. The Data Documentation Initiative metadata schema, developed for social science data (Vardigan et al., 2008), has been adapted for big social data. Data curators who work with big social data can advocate for and develop interoperable metadata standards that are designed specifically for big social data.

Consultation throughout the research process

The big social researchers interviewed for this study rarely contacted data curators unless they were considering sharing data. This meant that the researchers were not able to benefit from data curators’ broad knowledge during the research process. Data curators should focus on connecting with researchers early in the research process – through partnerships with IRBs, university research support offices, and big data providers, and then aim to maintain this connection throughout the research process. Once these connections are established, data curators can support responsible big social research in several ways: they can encourage strategies such as focus groups or automated strategies for obtaining individual informed consent, they can help researchers with rights management and navigating social media terms of service to support data sharing and reuse, and they can help researchers conduct risk-benefit analysis for big social research and big social data sharing.

Data repository services

Weller and Kinder-Kurlanda suggest that data repositories can "fuel the discussions on: suitable documentation practices and metadata standards, different models for data access (e.g., embargoes, access to sensitive data), [and] practices for anonymization of social media datasets" (2016, p. 170). The data curators I interviewed reported engaging with these functions. Data curators can provide support with de-identification procedures that work for big social data – such as programmatically rewording quotes or adjusting images. Repositories can control access to big social data. Some curators also reported using data enclaves to protect privacy and intellectual property for big social data; these enclaves allow researchers to conduct data analysis on repository servers so that repository staff can conduct disclosure reviews on the analytical output, rather than an entire big social dataset. Data repositories can also implement data use agreements and facilitate data licensing to support responsible use and reuse.

Advocacy for Community Standards

A key takeaway from this study is that we need more concrete guidance for big social research. Institutional review boards, which act as the main compliance bodies in the United States, grant exempt status to most research involving secondary analysis, data reuse, or data scraping. The big social researchers interviewed reported cobbling together strategies for responsible practice from many sources, including conducting on-the-fly risk-benefit analyses, talking to colleagues and collaborators, and reading other studies.

Legislation and regulation may help, but the scholarly community needs to find ways to ensure epistemologically sound, ethical, and legal big social research and qualitative data reuse in the meantime. Some professional organizations produce ethical guidelines for data use and data sharing (APA Data Sharing Working Group, 2015; ASA, 2018), and the Association of Internet Researchers maintains an in-depth set of ethical guidelines (Franzke et al., 2020). The data curation community also produces resources such as the Data Curation Network data curation primers (Data Curation Network, 2022). However, it was rare for big social researchers or data curators to discuss standardized ethical guidelines or clear community best practices, which shows that such guidelines are not widely disseminated or adopted. Future work for curators could include advocacy for standardized data curation practices to support big social research and qualitative data reuse. Engaging with professional organizations such as Research Data Access and Preservation and the Digital Library Federation could support standardization in data curation practice. These practices could also be taught to the next generation of data curators through standardized curriculum in Library and Information Science graduate programs. As with any standard, the community will need to commit to regularly revising and updating these standard practices.

Conclusion

This brief report provides an overview of the key issues in big social research. Data curators can address these issues to some extent through data curation and advocacy. Supporting responsible practices for using big social data can help scale up social science research, thus enhancing our understanding of human behavior and social phenomena.

Acknowledgements

Many thanks to Vivien Petras at Humboldt University of Berlin and Michael Zimmer at Marquette University, whose support and guidance were vital to this research.

References

Acker, A., & Kriesberg, A. (2017). Tweets may be archived: Civic engagement, digital preservation and Obama White House social media data. *Proceedings of the Association for Information Science and Technology*, *54*(1), 1–9. https://doi.org/10.1002/pra2.2017.14505401001

APA Data Sharing Working Group. (2015). *Data sharing: Principles and considerations for policy development*. American Psychological Association. https://web.archive.org/web/20220120232933/https://www.apa.org/science/leadership/bsa/data-sharing-report

ASA. (2018). *American Sociological Association code of ethics*. American Sociological Association. https://web.archive.org/web/20220121025817/https://www.asanet.org/sites/default/files/asa_code_of_ethics-june2018.pdf

Bankes, S., Lempert, R., & Popper, S. (2002). Making Computational Social Science Effective: Epistemology, Methodology, and Technology. *Social Science Computer Review*, *20*(4), 377–388. https://doi.org/10.1177/089443902237317

Berkout, O. V., Cathey, A. J., & Kellum, K. K. (2019). Scaling-up assessment from a contextual behavioral science perspective: Potential uses of technology for analysis of unstructured text data. *Journal of Contextual Behavioral Science*, *12*, 216–224. https://doi.org/10.1016/j.jcbs.2018.06.007

Bossetta, M. (2018). The digital architectures of social media: Comparing political campaigning on Facebook, Twitter, Instagram, and Snapchat in the 2016 U.S. election. *Journalism & Mass Communication Quarterly*, *95*(2), 471–496. https://doi.org/10.1177/1077699018763307

boyd, d., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, *15*(5), 662–679. https://doi.org/10.1080/1369118X.2012.678878

Bright, J. (2017). 'Big social science': Doing big data in the social sciences. In N. G. Fielding, R. M. Lee, & G. Blank (Eds.), *The SAGE handbook of online research methods* (pp. 125–139). SAGE Publications. https://doi.org/10.4135/9781473957992.n8

Bruns, A., & Weller, K. (2016). Twitter as a first draft of the present: And the challenges of preserving it for the future. *Proceedings of the 8th ACM Conference on Web Science - WebSci '16*, 183–189. https://doi.org/10.1145/2908131.2908174

Burgess, J., & Bruns, A. (2012). Twitter archives and the challenges of "big social data" for media and communication research. *M/C Journal*, *15*(5), Article 5. https://doi.org/10.5204/mcj.561

Croeser, S., & Highfield, T. (2020). Blended data: Critiquing and complementing social media datasets, big and small. In J. Hunsinger, M. M. Allen, & L. Klastrup (Eds.), *Second international handbook of internet research* (pp. 669–690). Springer Netherlands. https://doi.org/10.1007/978-94-024-1555-1_15

Data Curation Network. (2022). *Data Curation Network Primers*. University of Minnesota Digital Conservancy. https://hdl.handle.net/11299/202810

Franzke, A. S., Bechmann, A., Ess, C. M., & Zimmer, M. (2020). Internet research: Ethical guidelines 3.0. *AoIR (The International Association of Internet Researchers)*. https://web.archive.org/web/20220402125705/https://aoir.org/reports/ethics3.pdf

Hargittai, E. (2020). Potential biases in big data: Omitted voices on social media. *Social Science Computer Review*, *38*(1), 10–24. https://doi.org/10.1177/0894439318788322

Mannheimer, S. (2021). Data Curation Implications of Qualitative Data Reuse and Big Social Research. *Journal of EScience Librarianship*, *10*(4). https://doi.org/10.7191/jeslib.2021.1218

Martí, P., Serrano-Estrada, L., & Nolasco-Cirugeda, A. (2019). Social media data: Challenges, opportunities and limitations in urban studies. *Computers, Environment and Urban Systems*, *74*, 161–174. https://doi.org/10.1016/j.compenvurbsys.2018.11.001

Mason, W., Vaughan, J. W., & Wallach, H. (2014). Computational social science and social computing. *Machine Learning*, *95*(3), 257–260. https://doi.org/10.1007/s10994-013-5426-8

Neuburger, J. D. (2021). Supreme Court vacates Linkedin-hiQ scraping decision, remands to Ninth Circuit for another look. *The National Law Review*, *XII*(105). https://web.archive.org/web/20220331113645/https://www.natlawreview.com/article/supreme-court-vacates-linkedin-hiq-scraping-decision-remands-to-ninth-circuit

Nissenbaum, H. (2009). *Privacy in context: Technology, policy, and the integrity of social life*. Stanford University Press.

Obar, J. A., & Oeldorf-Hirsch, A. (2020). The biggest lie on the internet: Ignoring the privacy policies and terms of service policies of social networking services. *Information, Communication & Society*, *23*(1), 128–147. https://doi.org/10.1080/1369118X.2018.1486870

Olshannikova, E., Olsson, T., Huhtamäki, J., & Kärkkäinen, H. (2017). Conceptualizing big social data. *Journal of Big Data*, *4*(1), Article 1. https://doi.org/10.1186/s40537-017-0063-x

Palen, L., & Dourish, P. (2003). Unpacking "privacy" for a networked world. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 129–136. https://doi.org/10.1145/642611.642635

Schneble, C. O., Elger, B. S., & Shaw, D. (2018). The Cambridge Analytica affair and internet-mediated research. *EMBO Reports*, *19*(8), e46579. https://doi.org/10.15252/embr.201846579

Secretary's Advisory Committee on Human Research Protections. (2015). *Attachment A: Human subjects research implications of "big data."* https://web.archive.org/web/20220409135614/https://www.hhs.gov/ohrp/sachrp-committee/recommendations/2015-april-24-attachment-a/index.html

Stier, S., Breuer, J., Siegers, P., & Thorson, K. (2020). Integrating survey data and digital trace data: Key issues in developing an emerging field. *Social Science Computer Review*, *38*(5), 503–516. https://doi.org/10.1177/0894439319843669

Summers, E. (2017, August 21). *The catalog and the hydrator*. Documenting the Now. https://news.docnow.io/the-catalog-and-the-hydrator-3299eddfe21e

Törnberg, P., & Törnberg, A. (2018). The limits of computation: A philosophical critique of contemporary Big Data research. *Big Data & Society*, *5*(2). https://doi.org/10.1177/2053951718811843

Vardigan, M., Heus, P., & Thomas, W. (2008). Data Documentation Initiative: Toward a Standard for the Social Sciences. International Journal of Digital Curation, 3(1), 107–113. https://doi.org/10.2218/ijdc.v3i1.45

Varol, O., Ferrara, E., Davis, C. A., Menczer, F., & Flammini, A. (2017). Online human-bot interactions: Detection, estimation, and characterization. *Proceedings of the Eleventh International AAAI Conference on Web and Social Media (ICWSM 2017)*, 10. https://web.archive.org/web/20220307114705/https://www.aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15587/14817

Weller, K., & Kinder-Kurlanda, K. E. (2016). A manifesto for data sharing in social media research. *Proceedings of the 8th ACM Conference on Web Science - WebSci '16*, 166–172. https://doi.org/10.1145/2908131.2908172