STUDIES IN SOCIAL JUSTICE
Volume 15, Issue 3, 442-449, 2021

Dispatch

Ethical Approaches to Youth Data in Historical Web Archives

KATIE MACKINNON
University of Toronto, Canada

My doctoral research focuses on the experiences of young people learning about and exploring the World Wide Web from Canadian homes, schools, libraries and community centres between 1994-2004. While there are many intersecting facets of my research that include federal policy interventions, public discourse in Canadian media, and oral interviews, I engage significantly with web archives in order to provide perspectives from young and marginalized people who were creating websites and community on the early web. My research has focused on GeoCities, one of the most popular web hosting platforms between 1996-1999.

GeoCities users, called homesteaders, could build websites for free in different neighbourhoods that reflected interests and hobbies, like the WestHollywood (LGBTQ+) or EnchantedForest (Youth) neighbourhoods. When the platform was removed from the web in 2009, there were significant archival efforts to preserve the once-thriving online community in the Internet Archive. For researchers, this archive poses significant ethical, methodological and epistemological issues. Although it is a valuable resource for researching a history of the online communities on the early web, it also creates opportunities for harmful data practices while also calling into question individuals' "right to be forgotten" (EU, 2016b). This dispatch explores some ethical questions that have emerged through my research on digital experiences of young people throughout the 1990-2000s and the use of archived web materials created at that time by young people who were under the age of 18.

Correspondence Address: Katie Mackinnon, Faculty of Information, University of Toronto, Toronto, ON, M5S 3G6; Email: katherine.mackinnon@mail.utoronto.ca

ISSN: 1911-4788

Background

Debates over best practices relating to internet research ethics have been ongoing for the past two decades amongst scholars from a variety of fields, with numerous professional organizations developing guidelines (NESH, 2016; Shelley-Egan, 2015; Utrecht Data School, 2019). Ethics protocols, in general, are modelled after human rights and protection principles that emerged from medical science and were consolidated in the UN Declaration of Human Rights (United Nations, 1948), the Nuremberg Code (1949) and the Belmont Report (Office of the Secretary, 1979). The focus is typically on minimizing harm and maximizing benefits, with overall principles of respect and justice, resulting in general research practices around informed consent, confidentiality, privacy and anonymity.

Digital research has renewed debates over the implications of these concepts because "[the concept of] harm can be difficult to operationalize in a socio-technical context [in a way] that is persistent, replicable, scalable, and searchable" (boyd, 2010). It is generally agreed by internet researchers (Ess, et al., 2002; franzke, et al., 2019; Markham & Buchanan, 2012) that there should be no fixed rules for ethically sound research with digital media because it is impossible to establish protocols that can be applied across all contexts, cultures and research fields. Instead, a case-by-case approach that follows similar questioning and reflection – around procedure, tools, collection, storage, processing and documentation – has been encouraged (Crossen-White, 2015; Ess, et al., 2002; franzke, et al., 2019; Leurs, 2017; Lomborg, 2018; Markham & Buchanan, 2012).

As ordinary individuals and their public activities, as well as their personal lives, are less well-represented in historical records before the age of the internet (Lin et al., 2020; Lomborg, 2018), web archives hold the potential to shift the types of history that can be written, where terabytes of information are collected, stored and preserved from lay people and sources so that history can be written from below. This opens up radical potential for history on the margins of society, giving voice and power to communities and subcultures typically absent or erased from the records. However, this potential creates a pressing need for stronger, more robust ethical frameworks, guidelines and protocols to ensure that histories of the early web do not deploy systems of data exploitation and oppression. It is imperative for researchers to consider whose stories are being told, who is equipped to tell them, and what kinds of vulnerability and harm one might encounter and create when doing so.

If we consider how throughout the past 15 years young peoples' data have been subject to commodification, surveillance, and archiving without consent (Grimes & Chung, 2005; Steeves, 2015; Van Dijck et al., 2018), researchers who engage with archived web material have a responsibility to develop better practices of care for web materials created by young people in the past. The overview below demonstrates the trajectory of web ethics discussions and debates that I have explored in my approach to web archival research, and points to sensitizing concepts applicable to developing an ethics of care. While defining the contours of historical web research ethics and bringing together pieces in conversation with each other, I also demonstrate some of the gaps and oversights that are addressed later in critical feminist scholarship.

Ethics Protocols and Frameworks Overview

Holly Crossen-White (2015) describes this as a "new era" in historical research, where "the use of technology to explore the lives of individuals from the past in greater detail is really part of an ongoing wider global debate on the use of technology" (p. 110). For example, in 2014 The Court of Justice of the European Union ruled that in accordance with the EU's 1995 Data Protection Directive, individuals have the "right to be forgotten" (Art. 17) in online search engines like Google. This language was seen again in the European Union's General Data Protection Regulation (EU, 2016a), which applied the "right to erasure" to prevent "the indefinite storage and trade in electronic data, placing limits on the duration and purpose" of the data (Tsesis, 2014, p. 433). It also states that individuals may request that data be deleted should it become irrelevant, inaccurate, or cause harm that is not outweighed by a public benefit in retaining the data. This directive has become relevant to internet researchers for many reasons, but particularly applies to my use of web archives for historical research.

An early response to concerns over a lack of ethical protocols for internet research brought scholars together in 2002 in a working group to create the Association of Internet Researchers (AoIR) ethical guidelines, which have since been updated in 2012 and 2019 (Ess, et al., 2002; franzke, et al., 2019; Markham & Buchanan, 2012). The guiding principles declare that researchers must weigh harms to research communities, balance the rights of subjects with the social benefits of the research and consider the possibility of increasing the vulnerability of the researched communities. Concepts explored in the guidelines include informed consent, methods of accessing data, analyses, potential findings that can create harms, context and social vulnerability, responding to many of the gaps from previous ethical standards adopted by universities and research communities.

Taking up the AoIR guidelines in exploring ethical issues associated with providing access to web archives, Lin et al. (2020) survey different ethical frameworks that scholars and systems builders have been engaging with to discuss what types of analysis can and should be conducted. They grapple with the notion that while researchers have the computational tools to scrape web data for many types of analyses, a more pertinent question is whether it is ethical to pursue these types of academic avenues of inquiry. They focus on content-based retrieval, large-scale distant reading and user re-identification for research engaged with digital history. Lin et al. (2020) determine two significant guiding principles relevant for developing better practices of care: context (what the original creators expected of their materials) and scale (the conceptual and experiential distance between the researcher and the content creator).

Stine Lomborg (2018) also references these concepts, and asks, "how can we ensure that the voluntary sharing of personal data at one point in time does not come to negatively impact the research subject at a later point in time?" (p. 204). She also states that "researching children, social or politically marginalized groups and physically or mentally vulnerable individuals entails a great ethical responsibility to protect participants from being bullied or put on undesirable public display, regardless of whether the online activities we study revolve around their vulnerability" (p. 205). These questions highlight some of the most prominent concerns with the use of web archives for my research on youth cultures online, and reflect some of the issues with archives like the Internet Archive.

In proposing an answer to these questions, Lomborg highlights Helen Nissenbaum's (2010) "contextual integrity," an expectations-based framework for ethical reasoning about privacy and the protection of human subjects from harm in digital contexts. Instead of thinking of privacy as a matter of universal principles about who should have access to what information in which contexts, Nissenbaum suggests a lens of privacy as contextual integrity to consider whether something must be treated as private or not by looking at the specific context, data and actors involved. The principle of contextual integrity invites researchers to dwell on the possible ethical consequences of repurposing historical web data, from the original context, to the archive and on to the context of web history. In this framework, the relationship between actors, content and contexts are bound together, and also invoke "the distance principle" (Markham & Buchanan, 2015), which measures the conceptual and experiential distance between the object of study and the person who produced it (p. 611).

Katrin Tiidenberg (2018) explores how the applications of digital research ethics – such as seeking informed consent and manipulating data for confidentiality, privacy and anonymity – is an on-going issue in digitally saturated contexts. They explain that since sharing has become a default relationship between the self and technological infrastructures, which is both manipulated by the infrastructure to create and collect more valuable data and through the very nature of digital communication, a lack of privacy has been normalized. Privacy is instead constituted as an individual burden, where individuals are responsible for modulating their platform settings in relation to their awareness of privacy needs, which removes responsibility from corporations and platforms, as well as from researchers.

Critical Feminist Frameworks

A feminist ethics of care approach to web archives responds to many of the shortcomings in earlier ethical frameworks and practices. In the AoIR 3.0 guidelines, aline franzke (2019) presents an overview of feminist ethics of care to demonstrate how the intersections of power relationships produce inequalities that are relevant in the study of internet ethics. Some key principles highlighted call for the assessment of relationships between the researchers, the researched, and the wider research community. These guidelines are integral to the development of ethical commitments in research practices, however their application to specific research contexts is not yet widely accepted.

Koen Leurs (2017), in his research with the digital practices of young Londoners, created a roadmap to alternative data-analysis practices that explores what a social-justice oriented, feminist data study could look like. His framework allows for "attention to human meaning-making, context-specificity, inter/dependencies, temptations, as well as benefits and harm" with a moral focus on "relationality, responsibility, inter-subjectivity and the autonomy of the research participants" (Leurs, 2017, p. 140). Leurs' roadmap highlights five statements about data-related research practices, asserting that people are more than digital data (i.e., data are limited, ahistorical, decontextualized), data is context-specific and performative, data is dependent and relational to the platform it was created on, we are tempted to over-invest in digital representations of people rather than individuals themselves, and we benefit in various ways from our studies that do not always benefit research participants. Leurs' roadmap demonstrates one of the most robust approaches available and was paired with qualitative interviews with the young people whose data he was engaging. My approach to studying digital practices of Canadian youth 20 years ago borrows from this roadmap, but there are additional challenges to researching older web materials.

In the same vein as Leurs, *The Feminist Data Manifest-NO* (2019), a project led by Marika Cifor and Patricia Garcia and supported by numerous techno-feminist scholars, pulls at the intricacies of data-based research practices and suggests new pathways for considering our ethical responsibility to the human-centric web data. It declares that data is "always and variously attached to bodies" and vows to "interrogate the biopolitical implications of data with a keen eye to gender, race, sexuality, class, disability, nationality, and other forms of embodied difference." *Data Feminism* (D'Ignazio & Klein, 2020) echoes these sentiments, calling for changes to oppressive data collection, storage and manipulation tactics in academic research. They reaffirm that, despite the apparent ubiquity and novelty of data harvesting, the act of collecting and recording data about people is not new at all. In fact, it has long been employed as a technique of power and control over the lives of the people whose data are collected. The relationship between data and power is clear historically, whether it is through the logs of people captured and sold into slavery or the biometric technologies that are deployed to surveil Black people (Browne, 2015). With this in mind, the question of data ethics requires a framework of data justice rather than ethics alone, so that inequalities of the past and research and archive practices are not replicated into the future.

Gaps & Conclusions

One of my main challenges to researching historical web archives is grappling with conceptual and experiential distance between researchers and the researched; websites become defunct, users have grown up and changed names, and communities have splintered and dispersed. While deploying contextual integrity tactics and evaluating the scale of archived web materials are significant ethical developments, getting in contact with those whose data is stored in archives is still an important ethical dimension for many internet researchers. Recent scholarship on "growing up online" (Adair, 2019; Eichhorn, 2019; Robards & Lincoln, 2020), has rejected or interrogated the archive as a site of study because of its inherently exploitative relationship with online communities. Cass Adair has written about the ethical difficulties in researching and writing about queer and trans young people online in the 2000s in their own doctoral research. In "Delete Yr Account," they write that while archives of queer and trans life and survival are scarce and therefore valuable, "there is a significant anti-archival trans politics… [that] have emerged not out of desire for self-annihilation, but out of resistance to being siphoned up, pinned down, by state or corporate collection" (Adair, 2019). As I begin the collection phase of my doctoral research this means critically assessing my use of web archives and building from feminist frameworks for engaging with these materials, while developing an ethics of care that can be applied to studying online communities in the past.

References

Adair, C. (2019). "Delete Yr Account: Speculations on Trans Digital Lives and the Anti-Archival." *Digital Research Ethics Collaboratory.* Retrieved April 23, 2021 from http://www.drecollab.org/delete-yr-account-part-i/

boyd, d. (2010). Social network sites as networked publics: Affordances, dynamics, and implications. In Z. Papacharissi (ed.), *Networked self: Identity, community, and culture on social network sites* (pp. 39-58). Routledge.

Browne, S. (2015). *Dark matters: On the surveillance of Blackness.* Duke University Press.

Cifor, M., Garcia, P., Cowan, T. L., Rault, J., Sutherland, T., Chan, A., Rode, J., Hoffmann, A.L., Salehi, N., & Nakamura, L. (2019). *Feminist Data Manifest-No.* https://www.manifestno.com/

Crossen-White, H. L. (2015). Using digital archives in historical research: What are the ethical concerns for a 'forgotten' individual? *Research Ethics, 11*, 108-119.

D'Ignazio, C., & Klein, L. (2020). *Data feminism.* MIT Press.

Eichhorn, K. (2019). *The end of forgetting: growing up with social media.* Harvard University Press.

Ess, C., & Association of Internet Researchers Ethics Working Committee. (2002). *Ethical decision-making and Internet research: Recommendations from the AoIR ethics working committee.* http://aoir.org/reports/ethics.pdf

franzke, a. s., Bechmann, A., Zimmer, M., Ess, C. and the Association of Internet Researchers. (2019*). Internet research: Ethical Guidelines 3.0.* https://aoir.org/reports/ethics3.pdf

EU (European Union). (2016a). *Regulation (EU) 2016/679 of The European Parliament and of the Council on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC* (General Data Protection Regulation). Retrieved April 23, 2021 from https://gdpr-info.eu/

EU (European Union). (2016b). *General data protection regulation (GDPR) Article 17: Right to be forgotten.* Retrieved April 23, 2021 from: https://gdpr-info.eu/art-17-gdpr/

Grimes, S., & Chung, G. (2005). Data mining the kids: Surveillance and market research strategies in children's online games." *Canadian Journal of Communication, 30*(4), 527-548.

Leurs, K. (2017). Feminist data studies. Using digital methods for ethical, reflexive and situated socio-cultural research. *Feminist Review, 115*(1), 130-154.

Lin, J., Milligan, I., Oard, D. W., Ruest, N., & Shilton, K. (2020, March 14-18). *We could, but should we? Ethical considerations for providing access to GeoCities and other historical digital collections* [Conference presentation]. CHIIR '20, Vancouver, BC, Canada.

Lomborg, S. (2018). Ethical considerations for web archives and web history research. In N. Brügger & I. Milligan (Eds.), *SAGE handbook of web history* (pp. 199-219). Sage.

Markham, A., & Buchanan, E. (2012). *Ethical decision-making and internet research: Recommendations from the AoIR Ethics Research Committee (Version 2.0)*. Retrieved April 23, 2021 from: http://aoir.org/reports/ethics2.pdf

Markham, A. N., & Buchanan, E. (2015). Ethical considerations in digital research contexts. In J. Wright (Ed.), *Encyclopedia for social & behavioral sciences* (pp. 606-613). Elsevier.

NESH (National Committees for Research Ethics in Norway). (2016). *Guidelines for research ethics in social sciences, law and the humanities.* Retrieved April 23, 2021 from https://www.etikkom.no/Aktuelt/publikasjoner/Guidelines-for-research-ethics-in-the-social-sciences-law-and-the-humanities/

Nissenbaum, H. (2010). *Privacy in context: Technology, policy, and the integrity of social life*. Stanford University Press.

Nuremberg Code.(1949). *Trials of war criminals before the Nuremberg Military Tribunals under Control Council Law*. No. 10, vol. 2. (pp. 181-182). U.S. Government Printing Office. Retrieved April 23, 2021 from: https://www.loc.gov/rr/frd/Military_Law/pdf/NT_war-criminals_Vol-II.pdf

Office of the Secretary of the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1979). *The Belmont report: Ethical principles and guidelines for the protection of human subjects of research. Technical report. Department of Health, Education, and Welfare.* Retrieved April 23, 2021 from https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/index.html

Robards, B., & Lincoln, S. (2020). *Growing up on Facebook.* Peter Lang.

Shelley-Egan, C. (2015). *SATORI ethics assessment in internet research ethics*. Retrieved April 23, 2021 from https://satoriproject.eu/media/2.d.2-Internet-research-ethics.pdf

Steeves, V. (2015). Now you see me: Privacy, technology, and autonomy in the digital age. In G. DiGiacomo (Ed.), *Current Issues and Controversies in Human* Rights (pp. 1-31). University of Toronto Press.

Tiidenberg, K. (2018). Research ethics, vulnerability, and trust on the internet. In J. Hunsinger, L. Klastrup & M. Allen (Eds.), *Second international handbook of internet research* (pp. 1-15). Springer.

Tsesis, A., (2014). The right to be forgotten and erasure: Privacy, data brokers, and the indefinite retention of data. *Wake Forest Law Review, 49*(2), 433-484.

United Nations. (1948). *Declaration of Human Rights.* Retrieved April 23, 2021 from https://www.un.org/en/about-us/universal-declaration-of-human-rights

Utrecht Data School. (2019). *Data Ethics Decision Aid* (DEDA). Utrecht University. Retrieved April 23, 2021 from https://dataschool.nl/en/deda/

Van Dijck, J., Poell, T., & de Waal, M. (2018). *The platform society: Public values in a connective world.* Oxford Scholarship Online.