Selected Papers of #AoIR2024:
The 25th Annual Conference of the
Association of Internet Researchers
Sheffield, UK / 30 Oct - 2 Nov 2024

WEB ARCHIVING AFTER PLATFORMIZATION: READING ARCHIVED SOCIAL MEDIA ALONG THE GRAIN

Kieran Hegarty
RMIT University

A key part of the emergence of the “everyday Internet” (Goggin, 2004), the World Wide Web is now over three decades old. In digital media and internet research, many have suggested that, since its inception, the web has moved from a relatively open, navigation-based information space to a more centralised environment dominated by several large commercial platforms (Plantin et al., 2018; Lingel, 2020). The concept of “platformization” has been used to describe both key technical-material changes in the way information moves around the web (Helmond, 2015) as well as the reorganisation of cultural practices and markets (Nieborg & Poell, 2018) in the interests of companies who operate large platforms. Yet, the impact of platformization on the records of the web’s past—in web archives—is only starting to be explored (Acker & Kreisberg, 2020; Ben-David, 2020). While this recent work sheds light on the changing power dynamics that condition the formation of archives, there is currently little research exploring how platformization impacts the day-to-day work of building collections, how web archivists navigate the challenges of collecting material from platform environments, and what this means for the content and character—and therefore the future use—of web archives. In light of the fundamental shifts in power dynamics that shape how information is produced, distributed, and circulated, understanding how the “platformization of the web” (Helmond, 2015) has shaped the archived web is, therefore, a critical task.

METHODS

This paper considers the impact of “the rise of the platform as the dominant infrastructural and economic model of the social web” (Helmond, 2015, p. 1) on the content and character of web archives and explores how internet research can read web archives for traces of platformization. To do this, I draw on empirical and historical research at two libraries undertaking web archiving—the National Library of Australia (NLA) and the oldest library in Australia, the State Library of New South Wales (SLNSW). Both libraries have legislated collecting mandates and legal deposit provisions at the national and state level, enabling them to collect a wide range of “works that are available online” (*Copyright Act 1968* (Cth) s 195CA). The NLA was one

Suggested Citation (APA): Hegarty, K. (2024, October). *Web archiving after platformization: Reading archived social media along the grain.* Paper presented at AoIR2024: The 25th Annual Conference of the Association of Internet Researchers. Sheffield, UK: AoIR. Retrieved from http://spir.aoir.org.

of the first libraries in the world to start a web archiving program, commencing their PANDORA web archive in 1996 (now called the Australian Web Archive). The SLNSW collects material relevant to Australia’s most populous state and has developed a specific social media archive since 2012. Throughout 2021, I interviewed staff at the NLA and SLNSW, observed and participated in everyday work practices, and analysed organisational records. Through these data, I gained insight into the impact of platformization on day-to-day efforts to collect and provide access to a record of the web at national and state levels.

WEB CRAWLING AND WORKAROUNDS

Web archiving institutions typically rely on two techniques to collect material hosted on social media platforms: (1) web crawling and (2) collecting web material from a database made available through an application programming interface (API) (Brügger, 2018). Web crawling uses automated software to retrieve and store images, files, and code from a web server that can be reassembled to represent what a webpage looked like at a specific moment. The NLA takes this approach, with the results of crawls made openly available via the Australian Web Archive. In web archives research and practice, content hosted on social media platforms is recognised as a significant challenge for traditional web crawling techniques (Ben-David, 2020; Espley et al., 2014). In response to difficulties collecting social media using web crawling, staff at the NLA engage in workarounds to gather material that falls within the library’s broad collecting remit (e.g. the social media presence of political candidates and parties). These workarounds change as the library tries to keep up with ongoing changes to platform design and policies. As I reviewed the history and current practice of collecting at the NLA, I saw how my participants would respond to changes to platforms by experimenting with different strategies and techniques. The rationale, nature, and timing of platform-side changes were unclear to my participants, but the results were clear—again and again, the harvests simply no longer worked. The result is inconsistent collections, filled with unexpected ruptures left underexplained to both creators and users of web archives.

API-BASED COLLECTING AND PLATFORM CONSTRAINTS

Meanwhile, API-based collecting returns content from a database, which might include profile information, posts, number of “likes” or “shares”, and other metadata based on specific queries or criteria. While the NLA has been reticent to adopt API-based social media collecting to supplement web crawling, the SLNSW has taken this path since 2012 with its Social Media Archive. While APIs enable the sanctioned collection of data from social media platforms, the owners of these services can change the design of APIs and the terms that govern their use, with each change reflecting the priorities and interests of the company (Helmond et al., 2019; van der Vlist et al., 2022). As one of my interviewees involved in the SLNSW Social Media Archive told me in 2021, “we’re at the mercy of the platforms to a certain degree… the ramifications of API changes mean what may have been practical to do five years ago is no longer”. By embarking on API-based social media collecting, the contemporaneous collection of material available online is therefore constrained and enabled by platform strategies governing how social media content can be created, distributed, and used. Through the offer of APIs, platforms simultaneously open up new opportunities for developing library collections while also significantly constraining what data are collected and how these data can be made accessible to library users.

CONCLUSION: READING WEB ARCHIVES ALONG THE ARCHIVAL GRAIN

This exploration of web archiving at the NLA and SLNSW illustrates how the platformization of the web significantly impacts the content and character of web archives. In examining the constraints of web crawling and API-based collecting, I illustrate the power of platforms to establish (and thereby change) the conditions that enable and constrain flows of information hosted on their services into web archives and shape how this information can be used. Because of this, I suggest internet researchers can read web archives "*along* the archival grain" (Stoler, 2002, p. 100) for traces of platformization. This approach, advanced by anthropologist Ann Laura Stoler (2002, p. 87) in her work on colonial archives, focuses less on "mining… the *content…* of archival sources" and more on "their peculiar placement and *form*". Stoler (2002, p. 100) writes, "[w]e need to read [the archive] for its regularities, for its logic of recall, for its densities and distributions, for its consistencies of misinformation, omission, and mistake". Through this approach, web archives after platformization might be read not just as inconsistent, inaccessible, and error-ridden collections of social media but as evidence of contemporary struggles over data governance and platforms' refusal to allow public institutions access to data unless on the terms they stipulate.

REFERENCES

Acker, A., & Kreisberg, A. (2020). Social media data archives in an API-driven world. *Archival Science*, *20*(2), 105–123. https://doi.org/10.1007/s10502-019-09325-9

Ben-David, A. (2020). Counter-archiving Facebook. *European Journal of Communication*, *35*(3), 249–264. https://doi.org/10.1177/0267323120922069

Brügger, N. (2018). *The Archived Web: Doing History in the Digital Age*. MIT Press.

Copyright Act 1968 (Cth). https://www.legislation.gov.au/C1968A00063/

Espley, S., Carpentier, F., Pop, R., & Medjkoune, L. (2014). Collect, Preserve, Access: Applying the Governing Principles of the National Archives UK Government Web Archive to Social Media Content. *Alexandria*, *25*(1–2), 31–50. https://doi.org/10.7227/ALX.0019

Goggin, G. (2004). *Virtual Nation: The Internet in Australia*. UNSW Press.

Helmond, A. (2015). The Platformization of the Web: Making Web Data Platform Ready. *Social Media + Society*, *1*(2), 1–11. https://doi.org/10.1177/2056305115603080

Helmond, A., Nieborg, D. B., & van der Vlist, F. N. (2019). Facebook's evolution: Development of a platform-as-infrastructure. *Internet Histories*, *3*(2), 123–146. https://doi.org/10.1080/24701475.2019.1593667

Lingel, J. (2020). *An Internet for the People: The Politics and Promise of craigslist*. Princeton University Press.

Nieborg, D. B., & Poell, T. (2018). The platformization of cultural production: Theorizing the contingent cultural commodity. *New Media & Society*, *20*(11), 4275–4292. https://doi.org/10.1177/1461444818769694

Plantin, J.-C., Lagoze, C., Edwards, P. N., & Sandvig, C. (2018). Infrastructure studies meet platform studies in the age of Google and Facebook. *New Media & Society*, *20*(1), 293–310. https://doi.org/10.1177/1461444816661553

Stoler, A. L. (2002). Colonial archives and the arts of governance. *Archival Science*, *2*(1–2), 87–109. https://doi.org/10.1007/BF02435632

van der Vlist, F. N., Helmond, A., Burkhardt, M., & Seitz, T. (2022). API Governance: The Case of Facebook’s Evolution. *Social Media + Society*, *8*(2). https://doi.org/10.1177/20563051221086228