Selected Papers of #AoIR2017:
The 18th Annual Conference of the
Association of Internet Researchers
Tartu, Estonia / 18-21 October 2017

MAINTAINING THE WEB: WEB ARCHIVING, LABOUR AND THE INTERNET ARCHIVE

Jessica Ogden
University of Southampton

Susan Halford
University of Southampton

Les Carr
University of Southampton

Introduction

Web archives – including social media archives – have become a critical resource for accessing historical snapshots of the Web. Beyond the increased promotion of web archives as a source for scholarly research, a growing number of examples point towards the use of web archives as tools for political accountability (e.g. Politwoops[1]); to provide temporary access points during times of restricted web access (e.g. during the 2013 US government shutdown[2]); and to reconstruct deleted domains (e.g. Ben-David's (2016) work on the former Yugoslav top-level domain[3]) – to name a few. Growing concerns and public debates over the trustworthiness of online media have positioned both web archiving and web archives as necessary and legitimate sources in the face of an ever-shifting 'ephemeral Web,' political unrest and algorithmically-generated access to web-based information.

In practice, web archiving initiatives have spanned from the large-scale activities of national libraries and archives, the Internet Archive and the work of networked communities such as Archive Team, to the individual efforts of scholars creating web archives for their own purposes. However, the processes undertaken to produce these vast and longitudinal resources are not well understood, and are often projected as uncontroversial technical solutions to the failings of web architecture. The ways in which practices and tools shape the nature of collection and access strategies – e.g. the

[1] https://www.politwoops.eu/ (Accessed: 27/09/2017)
[2] http://blog.archive.org/2013/10/02/governmentblackout/ (Accessed: 27/09/2017)

Suggested Citation (APA): Ogden, J., Halford, S. & Carr, L. (2017, October 18-21). *Maintaining the Web: Web Archiving, Labour and the Internet Archive*. Paper presented at AoIR 2017: The 18th Annual Conference of the Association of Internet Researchers. Tartu, Estonia: AoIR. Retrieved from http://spir.aoir.org.

timing, frequency and length of collection – are under-documented, along with the motivations and meaning-making activities that drive archival practices. These aspects of collection and access present problems related to the use of web archives and the ways in which web archives are interpreted and understood as archival sources. Whereas an abundance of technical research exists around developing tools and improving the efficiency and quality of web archival captures, little research exists around the interactive nature and structuring effects of human, algorithmic and automated agents in decisions around how, what and when to archive. This research contends that these (often) undocumented activities and practices are critical for interpreting the affordances of web archives and the types of claims made possible by their use. Here web archives are positioned as contingent constructions that inherently rely on the work of web archivists (as both networked human and non-human agents) to transform and preserve the Web(s) into archived snapshots. Drawing on Downey's (2014) work on the materiality of 'information labour,' here the concept of *web archival labour* is explored to encompass these practices and highlight the ways in which web archivists shape and maintain the preserved Web.

Methodology and Findings

This research uses an ethnographic approach to explore the mechanisms and circumstances surrounding the collection and maintenance of web archives at multiple sites of production. This paper presents the preliminary analysis of a study carried out at the Internet Archive ('the Archive'), a private, non-profit digital library that has been archiving the Web since 1996. A combination of non/participant observation, documentary sources and interviews were conducted over the course of four weeks in collaboration with web archivists, engineers and managers at the Archive. In the case of observations, ethnographic records were made of non/participation activities with the aim of providing the basis for 'thick descriptions' of practice. Observation pro-forma were not used however, details surrounding the actors (participants), artefacts/objects (technologies, tools) and activities were recorded in an effort to produce an ethnographic account. Interviews were used in combination with observations as a mechanism for clarifying existing ethnographic records and focusing subsequent observation activities. The interviews took a semi-structured approach using a combination of descriptive, structural and contrasting questions in direct response to the answers provided by informants within the context of the interview. All data was transcribed which provided the basis for domain and thematic analyses, to produce a model of significant components of web archival practices at the Archive.

Contrary to popular narratives surrounding the Archive that have emphasized abundance over more selective approaches to archiving, the data points towards a complex system of internal 'hybrid crawling' strategies for prioritising which web assets to collect. In recent years the Archive has begun to leverage their extensive existing archives for understanding networked linking behavior in an effort to balance the breadth and depth of crawling activities, whilst discovering new sources (e.g. websites linked from Twitter and Wikipedia) for identifying websites to crawl based on measures of popularity, 'novelty' and those sites that are in danger of going offline. The team has devised multiple mechanisms for identifying different types of 'undesirable domains,' including rule-based link pattern-matching and the development of 'gamified' tools for the manual curation of pornography and ‘domain squatter’ services. Collectively, these efforts can be seen as *knowledge work*, or what Downey (2014) calls the ‘high value labour' that goes into the production of information. These activities, seen in combination with other knowledge practices around the prioritisation, development and maintenance of technologies for web archiving by Archive staff all have ramifications for how web resources are actively transformed into archived, offline-online versions of their former selves. And as the global Wayback Machine currently provides access to 273 billion webpages from over 361 million websites – often inaccessible elsewhere – these editorial decisions have a tangible impact on not only the fidelity of archived captures, but indeed whether or not certain parts of the Web are preserved at all.

By observing the assemblages of practice and labour that drive web archiving (in this case at the Internet Archive), this research underscores some of the explicit and implicit values placed on practice by practitioners. A focus on labour acknowledges the constraints of automation towards a wider recognition of the ‘maintenance work’ required to sustain the complex sociotechnical relationships (Arnold 2016) between the creation, access and use of web archives. Web archival maintenance work (as an aspect of labour) can then be conceptualised as a place where the values of archivists are again ordered and re-articulated (e.g. through tool development and bug fixing, quality assurance work or the continuous selection and capture of particular ‘high priority’ domains). Here, by foregrounding the structure and agency of web archivists – as individuals, organisations, collectives, algorithms and code – a discussion of labour enables an exploration of the power relations that underpin widespread assumptions about what is collected, who is responsible for collecting and maintaining the ‘collective memory’ of the Web(s), for whom and for what purpose. This opens the door for further work examining the digital geographies and emergent inequalities of web archival labour, as an inherently time and place-based set of activities that enable and constrain participation, tool development and uptake, selection of domains and more. Going forward, this research indicates the importance of understanding web archiving both for the wider field of Internet researchers and social scientists interested in the politics of web-based information.

References

Arnold, Hillel. 2016. ‘Critical Work: Archivists as Maintainers’. August. http://hillelarnold.com/blog/2016/08/critical-work/.

Ben-David, Anat. 2016. ‘What Does the Web Remember of Its Deleted Past? An Archival Reconstruction of the Former Yugoslav Top-Level Domain’. *New Media & Society* 18 (7): 1103–19. doi:10.1177/1461444816643790.

Downey, Gregory J. 2014. ‘Making Media Work: Time, Space, Identity, and Labor in the Analysis of Information and Communication Infrastructures’. In *Media Technologies: Essays on Communication, Materiality, and Society*, edited by Tarleton Gillespie, Pablo J. Boczkowski, and Kirsten A. Foot, 141–65. Cambridge, Massachusetts; London, England: MIT Press.