Participatory Web Archiving: Opening the Black Box of SavePageNow

Paper submission for *The Web that Was: Archives, Traces, Reflection (RESAW19)*
University of Amsterdam, the Netherlands, June 19-21, 2019

Keywords: web archives; participatory web archiving; infrastructure studies
Topics: Archives and access; Realtime, time travel and other web temporalities; Software histories

In this paper we report on the preliminary findings and methodological considerations of a pilot project examining usage of the Internet Archive's SavePageNow (SPN) tool between 2013-2018. The Internet Archive (IA) has collected, maintained and provided access to the past Web since its inception in 1996. With the addition of SPN[1] in 2013, the IA enables anyone with a web browser and an Internet connection to add a web resource to the WayBack Machine. The IA is widely considered to be the largest web archive in the world, and according to recent estimates, over 100 URLs per second are added to the archive via the SPN API.[2] By extension, SPN can therefore be characterized as one of the largest public-access, 'participatory web archiving' projects, and as such it offers a window into how web archiving is being enacted at scale outside of traditional archival environments. By improving our understanding of this tool and its influence, this work reveals the ways in which participatory web archiving services generate stakeholders in web archives. In doing so, the paper also contributes to literature concerning the challenges of infrastructure studies and the role of automation in the circulation and production of information online.

Recent work in and on web archives illustrates the critical importance of studying web archival practices (Ogden et al., 2017; Summers and Punzalan, 2017) - framing these archival interventions as a relatively under-examined, yet increasingly embedded component of Web architecture. Accordingly, research examining IA and the Wayback Machine has begun to make visible and contextualise the often invisible infrastructural workings of the world's largest public web archive. Studies have explored the language and geographic distributions of content held within Wayback (AlSum et al., 2014; Thelwall and Vaughan, 2004), and conducted comparative studies of collection holdings held in web archives elsewhere. Further work has characterised the use of Wayback through the examination of IA access logs (AlNoamany et al., 2013) and the circulation of Wayback links on different social media platforms (Zannettou et al., 2018). Other complementary research has identified SPN as one of several mechanisms by which IA are attempting to 'leverage the power and labour of the crowd' in order to diversify the selection of domains and content for the Wayback Machine (Ogden et al., 2017). However, external observations about IA's use of the crowd as *citizen archivists* (Owens, 2013) to augment its archival holdings have been elusive until now.

[1] https://archive.org/web/
[2] https://twitter.com/brewster_kahle/status/994380510011928578

This study builds on existing work on IA to include a pilot exploration of the SavePageNow tool and API as a form of *participatory web archiving* - or web archiving activities that are centred on collaboration and the availability of open nomination and collection tools. In this paper we report on the preliminary findings of this pilot project. First, we discuss the processes undertaken to collect SPN data from the IA, as well as the documentary sources (including press releases and contemporary forum posts) used to frame our decisions surrounding what data to collect. Next, we examine the types of URLs archived via SPN between 2013-2018 over the course of one day per year in an effort to characterise SPN's use and contribution to Wayback. As a proxy for measuring the value of SPN by its various users, the extent to which these URLs are available on the live Web and other web archives is analysed. Following recent calls to study 'archival algorithmic systems' (Summers and Punzalan, 2017), we also examine the role of bots and automation in the production and circulation of the archived Web through an analysis of the various user agents used to submit URLs to SPN. To further contextualise the SPN assemblage, we also report on the results of additional experiments used to trace both automated and human-driven data flows through the infrastructure that facilitates SPN and the Wayback Machine.

To conclude, we discuss both concrete and potential implications of a publicly available tool for participatory web archiving. In doing so we consider the power and presumed democratising effects of SPN for shaping the archival record, as well as the multiple ways in which SPN *creates* and *transforms* users into stakeholders of the Internet Archive. We also collectively reflect on the opportunities and limitations of our approach, incorporating our plans to expand the project in the future to include a mixed-methods approach for contextualising public web archival practices. We also discuss how our findings suggest future directions for research on web archives that draws on critical data studies, infrastructure studies and Science and Technology Studies, more generally.

Bibliography

AlNoamany Y, Alsum A, Weigle MC, et al. (2013) Who and What Links to the Internet Archive. *CoRR* abs/1309.4016. Available at: http://arxiv.org/abs/1309.4016.

AlSum A, Weigle MC, Nelson ML, et al. (2014) Profiling web archive coverage for top-level domain and content language. *International Journal on Digital Libraries* 14(3): 149–166. DOI: 10.1007/s00799-014-0118-y.

Ogden J, Halford S and Carr L (2017) Observing Web Archives: The Case for an Ethnographic Study of Web Archiving. In: *Proceedings of WebSci '17*, Troy, NY USA, 2017. ACM. DOI: https://doi.org/10.1145/3091478.3091506.

Owens T (2013) Digital Cultural Heritage and the Crowd. *Curator: The Museum Journal* 56(1): 121–130. DOI: 10.1111/cura.12012.

Summers E and Punzalan R (2017) Bots, Seeds and People: Web Archives As Infrastructure. In: *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work*

and Social Computing, New York, NY, USA, 2017, pp. 821–834. CSCW '17. ACM. DOI: 10.1145/2998181.2998345.

Thelwall M and Vaughan L (2004) A fair history of the Web? Examining country balance in the Internet Archive. *Library & Information Science Research* 26: 162–176.

Zannettou S, Blackburn J, De Cristofaro E, et al. (2018) Understanding Web Archiving Services and Their (Mis)Use on Social Media. In: January 2018. arxiv. Available at: https://arxiv.org/abs/1801.10396.