Exploring assemblages of appraisal in web archives

Ed Summers
University of Maryland

Introduction

How much of the web is archived? It's difficult to say because we don't really know how big the web is. In 2008 Google announced it had **1 trillion URLs** in its index (Alpert and Hajaj 2008). But even Google doesn't know the full extent of the Web. In June 2017 the Internet Archive stated on its website that it has collected **298 billion pages**. Based on this we could estimate that about 28% of the web has been archived. But this estimate is much too high, because the web is sure to have grown in the last 10 years, and the Internet Archive's count includes many snapshots of the same URL over time.

Failing to capture everything should not be surprising to the experienced archivist. Over the years, archival scholars have argued that gaps and silences in the archival record are inevitable (Cook 2011) and sometimes even desirable (Mayer-Schönberger 2009). The central challenge facing the archival community is to better understand our predisposition to privilege dominant cultures and materials, and the biases that are built into our preservation infrastructures, which result in gaps in society's archives.

Even after over 20 years of active web archiving we know surprising little about how archivists appraise and select web content for preservation. If we can't keep it all, how we decide what to keep from the web is certain to shape the historical record.

Materials and Methods

To study appraisal in web archives 29 ethnographic interviews were conducted with individuals involved in the appraisal or selection of web content for preservation. Rather than providing a generalized picture, the research goal was instead to evoke a thick description of how practitioners enact appraisal in their particular work environments.

Interview subjects were selected using purposeful sampling that included practicing web archivists but also sought out extreme or deviant cases such as researchers, managers, local government employees, volunteers, social activists, and business entrepreneurs.

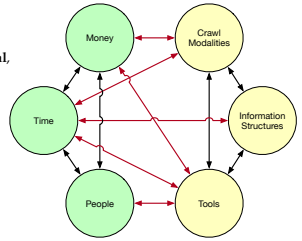
Field notes conducted during these interviews were analyzed using inductive thematic analysis. Analysis began with reading all the field notes together, followed by line by line coding. While coding was done without reference to an explicit theoretical framework, it was guided by an interest in understanding archival appraisal as a sociotechnical and algorithmic system (Botticelli 2000, Kitchin 2016)

Results

Coding and analysis surfaced six interconnected and interdependent themes that fell roughly into two categories, the **social** and the **technical**, which are illustrated here in different colors.

The dense network of lines between the themes represent the dependency relationships that were observed between them in our interview data.

Instead of reducing appraisal in web archives to a simple and rationalized representation the diagram suggests the complexity and inherent messiness of the social and material practices of appraisal in web archives.



Crawl Modalities
The selection strategies designed into tools and chosen by archivists in their work: domains, websites, topics, and events.

Information Structures
Specific formations of web content that archivists interacted with using their tools: hierarchies, networks, streams, and lists.

Tools
Configurations of tools that were used: archiving services, storage, spreadsheets, email, social media, content management systems.

People
Field archivists, managers, technicians, journalists, volunteers, software developers, groups (activists, professional), and institutions.

Time
How long to collect, how often to collect, how quickly web content needed to be gathered, perceptions of change in content.

Money
Grants from foundations and agencies to support collection activities, staffing, subscription fees, relationship between money and storage

Conclusion

The findings highlighted sites of breakdown that are illustrated by the red lines in the thematic diagram. These breakdowns are not deficiencies, but rather examples of infrastructural inversion (Bowker 2000), or sites where the infrastructure of web archiving became legible.

Breakdowns between **People** and **Tools** could be seen in the use of applications outside of web archiving software such as email, spreadsheets and forms to provide missing features for documenting provenance and enabling communication and collaboration in web archiving tools between archivists, technicians, software developers, and researchers.

Breakdowns between **Crawl Modalities** and **Information Structures** were also evident. For example, an inter-institutional collaboration focused on documenting fracking by archivists was problematized when fracking corporations extended across state and national boundaries. Where should the scope of collection begin and end? The archivists improvised communication tools to track their selection and make these boundaries legible and actionable.

There were also breakdowns in **Money** and **Crawl Modalities** where archivists were not able to determine how much it would cost to archive a website, and used "test crawls" to estimate the size of websites by examining how many URLs were left uncrawled using various reports.

Appraisal decisions depend on visualizations of the material archive

Archivists experienced difficulty in determining the dimensions of websites and domains which complicated and entangled **Information Structures** and **Money**.

While our research methodology and findings do not suggest specific implications for design (Dourish 2011) they do highlight rich sites for for repair work and improvisational and participatory design (Jackson 2014).

Literature Cited

Alpert, J., & Hajaj, N. (2008, July). We knew the web was big. *Google Official Blog*. Retrieved from https://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html

Botticelli, P. (2000). Records appraisal in network organizations. *Archivaria*, 1(49), 161-191.

Bowker, G. C., & Star, S. L. (2000). *Sorting things out: Classification and its consequences*. MIT Press.

Cook, T. (2011). We are what we keep; we keep what we are: Archival appraisal past, present and future. *Journal of the Society of Archivists*, 32(2), 173–189.

Dourish, P., & Bell, G. (2011). *Divining a digital future: Mess and mythology in ubiquitous computing*. MIT Press.

Jackson, S. J. (2014). Rethinking repair. In P. Boczkowski & K. Foot (Eds.), *Media technologies: Essays on communication, materiality and society*. MIT Press.

Kitchin, R. (2016). Thinking critically about and researching algorithms. *Information, Communication & Society*, 20(1), 1-16.

Mayer-Schönberger, V. (2009). *Delete: The virtue of forgetting in the digital age*. Princeton University Press.

Acknowledgments

Thank you to Ricky Punzalan for much guidance during the planning and execution of the study, Leah Findlater and Jessica Vitak also helped the selection of research methods. The Maryland Institute for Technology in the Humanities and the Documenting the Now project (funded by the Mellon Foundation) provided support for this research. Many thanks to the generous members of the web archiving community that shared their time, expertise and wisdom.

Noun Project images by Nirbhay, il Capitano, Creative Stall, Setyo Ari Wibowo, Agni, and Shuaib Usman Yusuf.

MITH DocNow

UNIVERSITY OF MARYLAND 18 56

Further Information

To learn more about the study described by this poster please see the paper that was presented this year at the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing in Portland Oregon.

Summers, E., & 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* (pp. 821–834). New York, NY, USA: Association for Computing Machinery.

A Creative Commons licensed pre-print of the article is also available at the arXiv:

https://arxiv.org/abs/1611.02493