Exploring assemblages of appraisal in web archives

Ed Summers
University of Maryland

Research Question

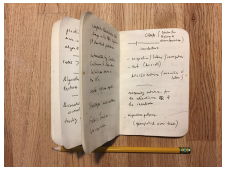
Even after over 20 years of active web archiving we know surprising little about how archivists appraise and select web content for preservation.

Since we can't keep it all, how we decide what to keep from the web is certain to shape the historical record (Cook 2011). In this context, we ask the following research questions:

1. How are archivists deciding what to collect from the web?
2. How do technologies for web archiving figure in their appraisal decisions?
3. Are there opportunities to design more useful systems for the appraisal of content for web archives?

Methodology

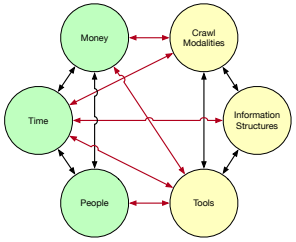
This study conducted a series of semi-structured interviews with 29 individuals involved in the selection of web content. Participants include web archivists as well as researchers, managers, local government employees, volunteers, social activists, and entrepreneurs. The field notes from these interviews 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).

Findings

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 (green and yellow respectively). Appraisal in the context of web archiving is a complex interplay between the following:



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

Information Structures: Specific formations of web content that archivists interacted with during appraisal: 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 examples of infrastructural inversion (Bowker 2000), or sites where the infrastructure of web archiving became legible.

Breakdowns between **People** and **Tools** were seen in the use of external applications such as email, spreadsheets and forms to provide missing communication features for documenting provenance and appraisal decisions.

Breakdowns between **Crawl Modalities**, **Information Structures** and **Tools** were also evident when archivists improvised communication tools to coordinate selection decisions when geopolitical boundaries complicated collection policies.

Breakdowns in **Money**, **Crawl Modalities** and **Information Structures** occurred when archivists could not determine how much it would cost to archive a website, and attempted to estimate the size of websites.

Appraisal decisions depend on visualizations of the material archive

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

Literature Cited

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.

Acknowledgments

Thank you to Ricky Punzalan for much guidance during the planning and execution of the study, Leah Findlater and Jessica Vitak also helped in 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 generous support for this research. Many thanks to the 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

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.