Exploratory Analysis of the End of Term Web Archive: Comparing two collections.

Mark Phillips
UNT Libraries
mark.phillips@unt.edu

Dan Chudnov
District Data Labs
dchudnov@districtdatalabs.com

James Jacobs
Stanford University Libraries
jrjacobs@stanford.edu

ABSTRACT

This paper reports on a preliminary exploration into two web archives created from the US Federal government web during the 2008 and 2012 presidential election. These web archives were created as part of a collaborative effort by institutions from across the United States to document the transition of the executive branch of government and changes that might occur as policies and initiatives shift between administrations. This research represents the first work to compare these two web archives and document differences that exist between them.

CCS Concepts

Information systems➡Digital libraries and archives

Keywords

Web archiving; descriptive metrics; End of Term crawl; digital library collections

1. INTRODUCTION

In 2008 and 2012 a group of institutions across the United States worked together to create a collaborative archive of the US Federal government web called the End of Term Presidential Archive (EOT). These collections primarily consist of the .gov and .mil domains. Currently no comparative analysis of these two collections has been performed to answer simple questions such as the following: How much is the US .gov domain – as well as individual agency domains – changing and growing over time? What content and URLs were present in 2008 that were no longer present in 2012? Was there noticeable change in the file types published by agencies during this four-year period? This work presents an analysis of several of these and similar questions by exploring data present in EOT2008 and EOT2012 collections.

2. RESEARCH QUESTIONS

This study compares two web archives of the US Federal web domain collected at the end of the Bush administration (2008) and the end of the first Obama administration (2012). At this time, effort has been made to provide infrastructure to facilitate discovery of these collections by researchers [1]. Additionally the 2008 EOT archive was discussed [5] and studied individually in a series of papers [2] [3] [4] that represent the output of an IMLS funded research project related to that EOT archive. Besides this access related work and the single collection research there has been no work to compare these two web archives in order to better utilize and give access to them. Because of this there are several key questions about these two collections that have not been answered.

- When was content captured during the yearlong EOT crawl process? How do the compare or differ?
- What is the overlap of content between these two collections at the domain and URL level?
- At a macro level, what was present in 2008 that was no longer available in 2012?
- What is the difference between the collection intent as defined by the list of seed URLs compared to what was captured and is included in the collections.

3. METHODS

This exploratory analysis makes use of CDX files generated for the EOT2008 and the EOT2012 collections to answer a series of research questions presented above. In addition to answering these research questions, overall descriptive statistics were generated as a way of understanding the overall size, scope and makeup of these two collections. This analysis also makes use of the seed list of URLs as an indicator of collection intent of the partners of the EOT collaboration. By comparing the seed URLs against what was collected, we hope to garner a better understanding of content that was unknown pre-crawl, yet was collected.

4. NEXT STEPS

This work only focused on research questions that can be answered with relatively high-level metadata extracted from these two web collections. In addition to these questions there are a number of more involved questions that would require analysis of not only metadata about the content crawled but the full crawled content. Another area of study is textual analysis of the content itself for language identification, topical/agency analysis, and to inform future EOT collection decisions. Finally, questions that might be answered with the generation and mining of the network created by the links within the collections is another area for study. These areas, while out of scope of this work, represent future research areas of both analysis of these two collections specifically as well as the ability to offer examples of different ways to compare two web archives collected over time.

5. REFERENCES

[1] End of Term Web Archive – http://eotarchive.cdlib.org

[2] Murray, K., Ko, L., and Phillips, M. 2011. Curation of the End-of-Term Web Archive. *IS&T Archiving Conference (2011)*. http://digital.library.unt.edu/ark:/67531/metadc36301/

[3] Murray, K., Hartman, C. 2012. Curation of the End-of-Term Web Archive. *IS&T Archiving Conference (2012)*. http://digital.library.unt.edu/ark:/67531/metadc93305/

[4] Phillips, M., Murray, K. 2013. Improving access to web archives through innovative analysis of PDF content. IS&T Archiving Conference (2013). http://digital.library.unt.edu/ark:/67531/metadc155622/

[5] Seneca, T., Grotke, A., Hartman, C., and Carpenter, K. (2012) It Takes A Village To Save The Web: The End Of Term Web Archive, *Documents to the People (DttP),* 2012, Chicago: American Library Association