Web Archiving Environmental Scan

Gail Truman
Truman Technologies, LLC
4096 Piedmont Avenue, Ste. 217
Oakland, CA 94611 USA
+1-510-502-6497
gail@trumantechnologies.com

Andrea Goethals
Harvard Library
90 Mt. Auburn St.
Cambridge, MA 01970 USA
+1-617-495-3724
andrea_goethals@harvard.edu

ABSTRACT
This poster session summarizes the output of a comprehensive Web archiving environmental scan conducted between August and December 2015, with a focus on the preservation-related findings. The scan was commissioned by Harvard Library and made possible by the generous support of the Arcadia Fund.

Keywords
Web Archiving, Preservation, Best Practices

1. INTRODUCTION
Websites are an integral part of contemporary publication and dissemination of information, and as more and more primary source material is published exclusively to the Web, the capture and *preservation* of this ever-growing and ever-changing, dynamic content has become a necessity to support researcher access and institutional needs. Today's research libraries and archives recognize Website archiving ("Web archiving") as an essential component of their collecting practices, and various programs to archive portions of the Web have been developed around the world, within various types and sizes of institutions, including national archives and libraries, government agencies, corporations, non-profits, museums, cultural heritage and academic institutions.

To meet Website acquisition goals, many institutions rely on the expertise of external Web archiving services; others, with in-house staff, have developed their own Web archiving services. Regardless of the approach, the rate at which textual, visual, and audio information is being produced and shared via the Web, combined with the complexity and specialized skills and infrastructure needed for Web archiving processes today – from capture through quality assurance, description, and eventual discovery, to access and analysis by researchers – poses significant resource and technical challenges for all concerned.

Harvard Library sponsored an environmental scan [1] to explore and document current Web archiving programs (and institutions desiring a similar capacity) to identify common concerns, needs, and expectations in the collection and provision of Web archives to users; the provision and maintenance of Web archiving infrastructure and services; and the use of Web archives by researchers. The ultimate goal of the survey was to identify opportunities for future collaborative exploration

This environmental scan is not the first investigation into these areas. Other surveys over recent years have provided valuable information about the landscape of Web archiving activities, such as:

- The National Digital Stewardship Alliance (NDSA)'s Web Archiving in the United States. A 2013 Survey [2]
- NDSA Web Archiving Survey Report, 2012 [3]
- North Carolina State University (NCSU) social media scan, 2015 [4]
- A Survey on Web Archiving Initiatives, Portugal, 2011 [5]
- Use of the New Zealand Web Archive [6]
- Researcher Engagement with Web Archives, 2010 (Dougherty, M) [7]

While there may be overlapping areas covered within these reports and surveys, each examines a particular subtopic or geographical region in relation to Web archiving practices. The NDSA surveys are focused on the USA; the NCSU scan is focused on other areas of social media (such as Twitter) and does not include use cases or details about individual institutions; the Portuguese study examined 42 global Web archiving programs reporting only on the staffing and size (size in terabytes) of each institution's collections; and the Dougherty/JISC study focuses solely on the uses and needs of individual researchers. Other more narrowly focused surveys, such as the IIPC working group surveys, address targeted informational needs.

2. THE SCAN
Through engagement with 23 institutions with Web archiving programs, two service providers and four Web archive researchers, along with independent research, Harvard Library's environmental scan reports on researcher use of – and impediments to working with – Web archives. The collective size of these Web archiving collections is approximately 3.3 PB, with the smallest collection size under one TB and the largest close to 800 TB. The longest-running programs are over 15 years old; the youngest started in 2015. The poster includes the general findings of the scan but emphasizes the findings that are related to preservation.

3. GENERAL FINDINGS
The environmental scan uncovered 22 opportunities for future research and development. At a high level these opportunities fall under four themes: (1) increase communication and collaboration, (2) focus on "smart" technical development, (3) focus on training and skills development, and (4) build local capacity.

4. PRESERVATION FINDINGS
The environmental scan revealed many challenges preserving Web archives - some of them are organizational and some of them are technical. The end result, as one participant put it, is that "Web preservation is at a very immature stage".

The main organizational challenges were knowing whether or not the organization needed to take local responsibility for preservation, being able to trust other organizations to provide preservation services for Web content, lack of funding to pay for the infrastructure demanded by Web archiving, and lack of dedicated staffing with clear roles and responsibilities. Figure 1 shows that more than half of the scan participants report having no dedicated full-time staff for their Web archiving activities.

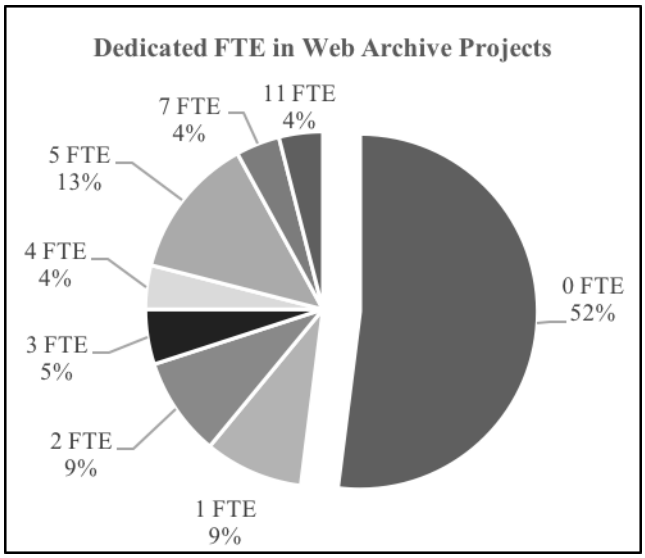

Figure 1. More than half of participants report having no dedicated full-time staff for their Web archive projects.

The technical preservation challenges were largely related to the scale of the Web content being collected and the diversity of the formats captured. Specifically, the main technical preservation challenges were the lack of tools for preparing captured Web content for preservation (see Figure 2); the challenges transferring and storing the large ARC/WARC files; the difficulty capturing certain formats in the first place, particularly social media; the challenges QAing the captures; and the increasing challenges in playing back the Web archives, especially as browsers evolve.

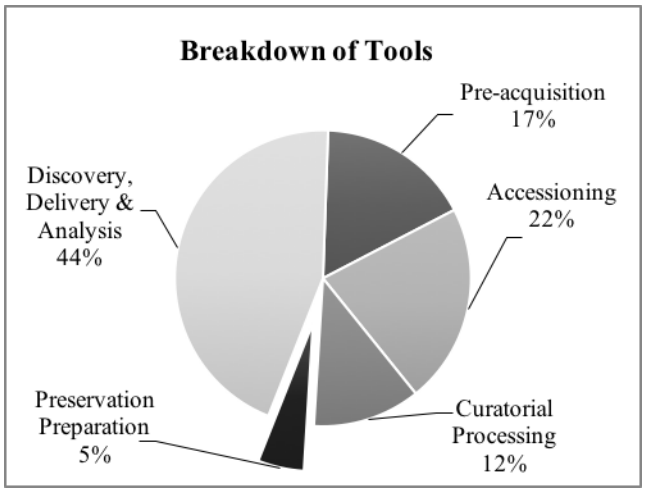

Figure 2. The distribution of tools according to the Web archiving life cycle function.

There were a great number of issues raised relative to the ARC/WARC formats themselves. These difficulties ranged from the complexities of de-duplication, to the difficulty of characterizing the files they wrap, to the difficulties of having to use specialized custom-built tools to process them; and to the problems trying to integrate Web archives with other preserved content.

Art-related Websites frequently break when being archived due to their high levels of dynamic content and interactivity. Preserving that interactivity is currently not possible – and highly desired.

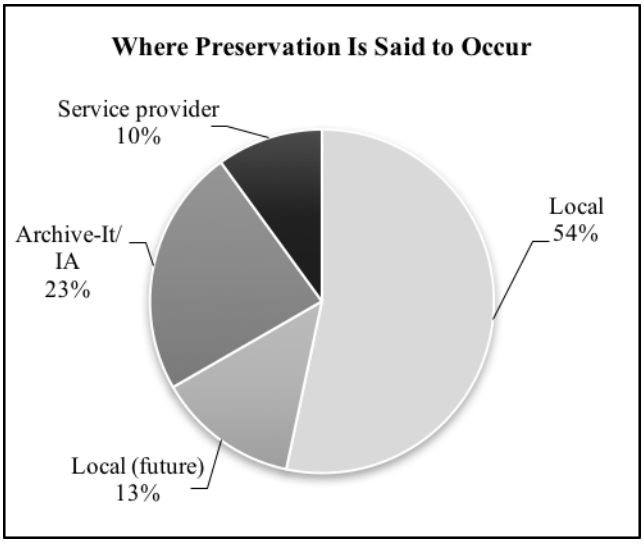

Figure 3. Location of each Web archiving program's preservation copies (now and planned).

5. REFERENCES

[1] Truman, Gail. 2016. *Web Archiving Environmental Scan*. Harvard Library Report. http://nrs.harvard.edu/urn-3:HUL.InstRepos:25658314

[2] National Digital Stewardship Alliance. 2013. *Web Archiving in the United States: A 2013 Survey*. An NDSA Report. http://www.digitalpreservation.gov/ndsa/working_groups/documents/NDSA_USWebArchivingSurvey_2013.pdf

[3] National Digital Stewardship Alliance. 2012. *Web Archiving Survey Report*. An NDSA Report. http://www.digitalpreservation.gov/ndsa/working_groups/documents/ndsa_web_archiving_survey_report_2012.pdf

[4] NCSU Libraries. 2015. Environmental Scan. https://www.lib.ncsu.edu/social-media-archives-toolkit/environment

[5] Gomes, Daniel, Miranda, Joao, and Costa, Miguel. 2011, A Survey on Web Archiving Initiatives. http://sobre.arquivo.pt/about-the-archive/publications-1/documents/a-survey-on-web-archiving-initiatives

[6] National Library of New Zealand. 2015. Use of the NZ Web Archive. http://natlib.govt.nz/librarians/reports-and-research/use-of-the-nz-web-archive

[7] Dougherty, M., Meyer, E.T., Madsen, C., van den Heuvel, C., Thomas, A., Wyatt, S. 2010. Researcher Engagement with Web Archives: State of the Art. London: JISC.