Classification of the End-of-Term Archive: Extending Collection Development Practices to Web Archives

Background

The *eotcd* Project builds on a project conducted collaboratively by the Library of Congress, the US Government Printing Office, the Internet Archive, the University of North Texas (UNT) Libraries, and the California Digital Library. That project captured the entirety of the federal government's public Web presence before and after the 2009 change in presidential administrations. The result is the End-of-Term Archive (EOT Archive), an approximately 16 terabyte Web archive of government information that is replicated in repositories at some of the collaborating organizations, including UNT.

As web archives become more available and accessible, many libraries will be collecting materials from these important information repositories. Librarians will need the capability to identify and select materials in accord with collection development policies. Additionally, libraries will need to characterize these materials using common metrics; however, such metrics are not established for Web archives, making it difficult for librarians to communicate the scope and value of these materials to administrators. The *eotcd* project utilizes the EOT Archive to investigate innovative solutions to address these needs.

Subject Matter Experts

Participants in this study are 10 librarians serving as Subject Matter Experts (SMEs) in the area of collection development for government information. Research will be conducted concurrently in two work areas:

1. *EOT Archive Classification*
   Classification of the EOT Archive will involve structural analysis and human analysis. Link analysis and visualization techniques will identify the organizational and relational structure of the EOT Archive. SMEs will map the Archive's URLs to the SuDocs classification system using a classification tool. The resulting SME classification map will serve as the standard against which the effectiveness of the structural analysis will be evaluated. It is hoped that effective machine-classification of Web archive content will enable librarians to discover materials appropriate for their collections.

2. *Web Archive Metrics*
   Identification of metrics for Web archives will be informed by the project's SMEs who will participate in two focus groups to identify and refine the criteria libraries use for acquisition decisions. A tool will be developed to translate these criteria into measurable units appropriate to the EOT Archive. An acquisitions exercise will test the effectiveness of this tool. The resulting metrics will hopefully enable characterization of materials in Web archives in units of measurement familiar to libraries and their administrations.

Status

Visualization of the Archive is underway using various open-source tools. Archive service models and associated metrics have been drafted. Assessment of the need for both access and acquisition service models was explored using a survey of depository libraries. Lastly, the SMEs will soon commence classification of the Archive's contents using a Web-based tool the project developed.

Project Website

The presentation is available on the *eotcd* project website**: http://research.library.unt.edu/eotcd**