University of North Texas Libraries

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

ABSTRACT. 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. Organization of the content in Web archives using established schemes is a promising solution to enable the extension of collection development practices to this new class of materials. Additionally, libraries will need to characterize these materials using common metrics; however, such metrics do not exist, making it difficult for librarians to communicate the scope and value of these materials to administrators.

This project will address these two collection development needs relative to government information, which is included in many library collections and for which a well-established classification scheme exists, the Superintendent of Documents (SuDocs) Classification Numbering System. In a unique approach to organizing Web archives, this project proposes to classify the materials in the 2008-2009 End-of-Term (EOT) Web in accord with the SuDocs system. This important archive includes the entirety of the federal government's public Web presence before and after the 2009 change in presidential administrations. This proposed classification will enable librarians to select materials consistent with their collection development policies. Additionally, in response to the deficit of established metrics for materials in Web archives, this project proposes to identify metrics that fill the gap. The results will enable librarians to translate measurable units for selected materials in Web archives to units more familiar to libraries and their administrations.

The University of North Texas is partnering with the Internet Archive in this forward-thinking project that will investigate innovative solutions to two research questions:

1. How effective is the organization of large-scale unstructured Web archives using a pre-defined classification system, the SuDocs classification numbering system, as evaluated by government information librarians?
2. What measurable units for the materials in Web archives best support management acquisition decisions in libraries?

Participants in this study will be 10 librarians who will serve as Subject Matter Experts (SMEs) in the area of collection development for government information. Tools built for the project will use open source platforms and will be publically available. Research will be conducted concurrently in two work areas: EOT Archive Classification and Web Archive Metrics.

Classification of the 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 seed URLs to the SuDocs classification system using a Web based application the project will develop. The resulting classification map, the *SME Map*, will serve as the standard against which the effectiveness of the structural analysis method will be evaluated.

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 EOT Archive. An acquisitions exercise will test the effectiveness of this tool. The final evaluation will report the findings and metrics.

The findings from this research will set the stage for testing this approach with other government information Web archives at state and international levels, where information is typically classified using established schemes. After Web archive content is classified, it will become much more feasible to apply subject analysis to the content and build information retrieval systems that allow librarians to identify and select materials for their collections. With the capability to characterize these materials using common metrics, librarians will be able to include this increasingly critical class of materials in their collections.