# Problems and Issues in Selecting, Harvesting, and Cataloging Web Resources

Joanne Archer and John Schalow
University of Maryland Libraries

Academic research libraries have long seen it to be part of their mission to build coherent collections of scholarly and research resources to support the needs of the institutions in which they are based. The advent of the Internet has made achieving this goal more difficult as materials traditionally published in print have migrated to digital only formats. Not only are these materials being published electronically but more importantly they have become ephemeral in nature, removed from websites as they become outdated or as organizations cease to exist. At the University of Maryland, this move to web publication has particularly impacted special collections in the areas of state documents, historic preservation, and university publications.

Other research libraries have begun to address the disappearance of web-based content by initiating web harvesting projects. These projects utilize a variety of different tools with the ultimate goal of preserving the content of selected websites as the sites evolve over time. Notable projects include the California Digital Libraries Web at Risk project (http://webarchives.cdlib.org/p/about) and the Library of Congress Web Archives (http://lcweb2.loc.gov/diglib/lcwa/html/lcwa-home.html). The importance of these types of effort is demonstrated by the results of pilot study by the Chesapeake Project Legal Information Archive which showed that more than 14% of URLs archived between 2007-2009 had disappeared by the end of the two year time period.

In 2008 the University of Maryland investigated issues related to harvesting web content in conjunction with Columbia University and with the support of the Andrew Mellon Foundation. This project compared harvesting tools and techniques, outlined possible workflows, and considered metadata issues. As a result of this investigation, the Libraries secured a 6-month trial subscription to the Internet Archive's web harvesting tool, Archive-It (http://www.archive-it.org/). Archive-It provides the ability to harvest, catalog, manage and browse digital content while also providing long-term storage for this content. Currently more than 40 University Libraries and 20 State Libraries/Archives use Archive-It to capture and preserve web content.

This paper will survey the different methods of preserving web based content and then focus on presenting the results of the University of Maryland's work. The paper will address issues such as the establishment of selection policies, the integration of access with existing tools and systems, and the creation of sustainable workflows within the library.