A Quantitative Evaluation of Dissemination-Time Preservation Metadata

Joan A. Smith and Michael L. Nelson

Old Dominion University, C.S. Dept, Norfolk VA 23529

Abstract. One of many challenges facing web preservation efforts is the lack of metadata available for web resources. In prior work, we proposed a model that takes advantage of a site's own web server to prepare its resources for preservation. When responding to a request from an archiving repository, the server applies a series of metadata utilities, such as Jhove and Exif, to the requested resource. The output from each utility is included in the HTTP response along with the resource itself. This paper addresses the question of feasibility: Is it in fact practical to use the site's web server as a just-in-time metadata generator, or does the extra processing create an unacceptable deterioration in server responsiveness to quotidian events? Our tests indicate that (a) this approach can work effectively for both the crawler and the server; and that (b) utility selection is an important factor in overall performance.

1 Background

There are many on-going efforts aimed at web preservation. One problem shared by these efforts is the dearth of metadata available directly from websites themselves. For preservation, we need much more metadata than is customarily available from an HTTP request-response event. A common approach to this problem is to crawl the site then have the archivist store the resources for later analysis and repository ingestion. However, we believe that the best time to analyze a file is at the time of the request, when the server itself is more likely to be able to provide preservation-related information. We also believe that automated metadata utilities installed at the originating web server can contribute meaningfully to web preservation.

We demonstrated this as a proof-of-concept in prior work [1, 2], but the question remained whether it is *practical* to use the site's web server as a just-in-time metadata generator. Does performance suffer an unacceptable deterioration? Can an archival request be serviced simultaneously with quotidian web requests? To investigate the feasibility of this approach, we constructed a "typical" website for testing based on an analysis of published web site characteristics. We then subjected this test website to varying request (load) levels and harvested the contents to determine the performance impact of creating preservation metadata at dissemination time, i.e., at the time of the request. We found that for all metadata utilities but one, we could process the results without a significant impact on server performance overall. Our tests indicate that (a) this approach

B. Christensen-Dalsgaard et al. (Eds.): ECDL 2008, LNCS 5173, pp. 346–357, 2008.

can work effectively for both the crawler and the server; and that (b) utility selection is an important factor in overall performance.

2 Related Work

2.1 Characterizing a Typical Website

Typical Website Content. Since the website's resources would be passed through the rigors of various metadata utilities, we wanted our test web to mimic a "typical" website in terms of content and structure. But what, exactly, is a typical website and what does a typical web page contain? An extensive survey of web content was published by Berkeley in 2003 [3]. At that point, surface web composition was roughly 23.2% images, 17.8% HTML, and 13% PHP, with the rest a collection of other formats ranging from PDFs to animations. More recent studies support this rough proportion, noting that most web pages have one or more images embedded in them thus contributing to a higher ratio of images to HTML resources but still supporting the intuitive impression that the web is largely HTML [4, 5, 6].

With regard to website size and content, a 2004 report on the composition of various national domains [4] showed a wide range of average number of pages per site, with a low of 52 (Spain) to a high of 549 (Indochina). That same study also indicated a preponderance of HTML over other document types, with PDF and plain text files accounting for up to 85% of the remainder (these figures do not include image files). Various studies on web content and configuration [5, 6] found that most HTML documents contain less than 300 words, with a per-page average of 281 HTML tags and a 221x221 pixel image (usually GIF or JPEG) that acted as a document header, much like the banner name of a newspaper. A 2004 examination of e-commerce sites at a large server farm [7] found an average object size of 9 KB and a much higher percentage of image use than seen in other studies, which the authors attribute to the nature of e-commerce sites. Other researchers [8, 9] have noted an increasing use of dynamic presentation technologies like Javascript, PHP, and Active Server pages.

Despite the many web studies available, no clear characterization of a "typical" website emerges, except perhaps at the extremes: single-page sites (often at "spam farms") and infinite sites, which use dynamic-generation to create infinite pages, such as a meeting-schedule site with a limitless value for future date. We are therefore left to "guesstimate" the composition of a small departmental or community website in terms of size and types of resources. The general tendency seems to be a small website of a few hundred files, with the HTML pages roughly 5 KB to 25 KB in size; having approximately 3 or more images embedded per HTML page; containing links to various internal resources distributed throughout the site, and a variety of external links on selected pages.

Typical Website Traffic Patterns. Many studies have been done on web traffic patterns, including some at large commercial sites [7, 9]. Data from these studies enable researchers to model request patterns realistically when simulating