# CRAWLING THE CRAWLERS: SEARCH ENGINE BEHAVIOR AND ITS IMPLICATIONS FOR WEBSITE DESIGN AND PRESERVATION

DR. JOAN A. SMITH
EMORY UNIVERSITY

**Abstract:** We know how search engines crawl websites: they harvest links from pages and iterate through each to develop an"imag" of the world wide web as a whole. During the last decade of search engine growth, websites have attempted to help or even to more directly inuence the crawlers and the subsequent website ranking. For example, conventional wisdom holds that search engines *prefer* sites that are wide rather than deep, and that having a site index will result in more thorough crawling by the *big three crawlers: Google, Yahoo, and MSN*. But how do crawlers actually behave on websites? Does site design really a?ect this behavior? We created a series of 10 websites to monitor search engine behavior when crawling with very large websites (wide and deep), as well as their behavior on websites where resources *disappear*. We analyzed the logs of each of these sites for over a full year to see if the conventional wisdom holds true. GIF animations of Apache log data are used to illustrate the crawling patterns. We found that each search engine exhibited di?erent behavior and crawl persistence, and that site design does appear to a?ect this behavior. We plot the progress of the crawlers through the sites, and their behaviors regarding the various le types. A side benet of search engine activity on a site is the "cached page" which is accessible if the original is unavailable. How long will such pages persist in the cache if the web source page disappears? We examine this issue and the role that the cache and the websites design play in website preservation.

**Emory University Computer Science Seminar**
**Friday, 24 October 2008 at 3:00pm**