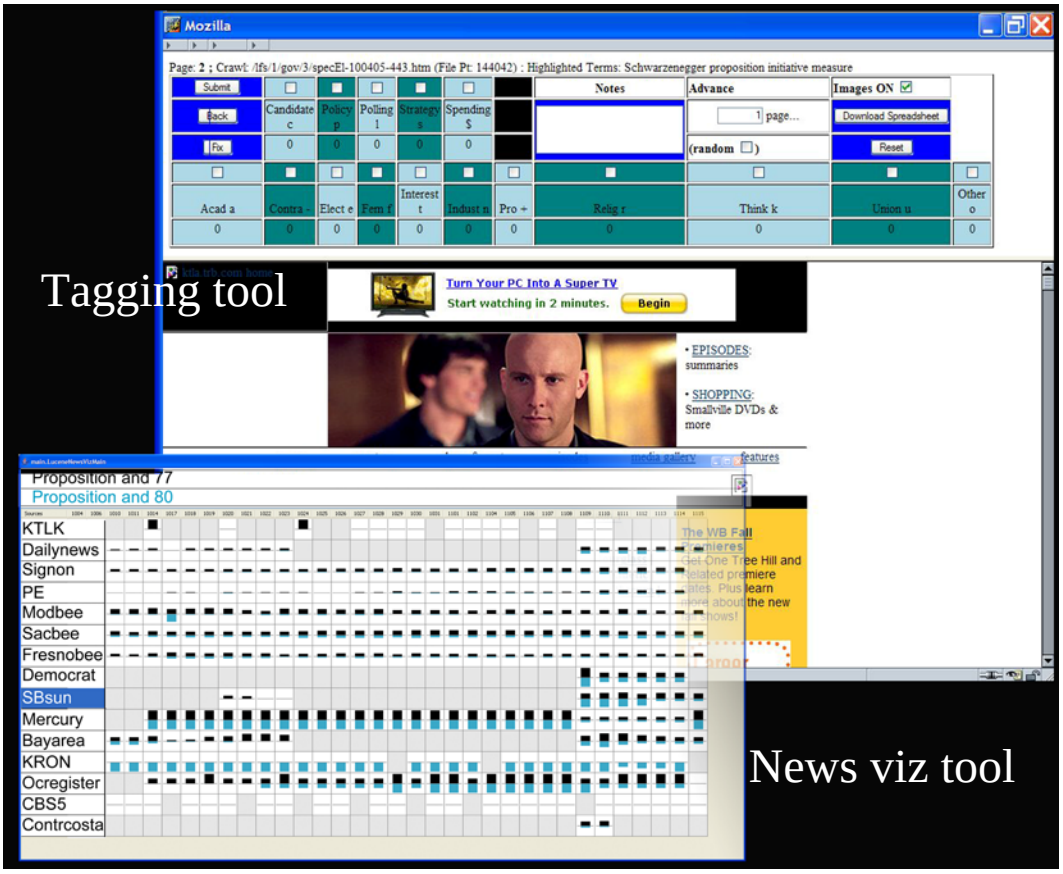
SGER: A Web Sociologists Workbench

Andreas Paepcke and Hector Garcia-Molina
Stanford University
(paepcke;hector)@cs.stanford.edu

The World-Wide Web is increasingly the place where our cultural and political history is recorded. The Web, unfortunately, is ephemeral. Pages come and go; links that work today are dead tomorrow. Efforts are under way to archive this legacy. Our project has with past NSF funding developed such an archiving system, and we are regularly collecting portions of the Web into storage.

With this grant we move beyond archiving, asking the question: what do we do after archiving succeeds? In search for an answer we partnered with political science faculty at Stanford. From them we learned current practice for political and historical analysis over newspapers. Those methods will not scale to archives with millions of Web pages. Newspaper analyses, for example, has humans read articles and manually tag the pages by topic. The tags classify content, such as 'feminism,' 'labor unions,' or 'presidential election.' Statistical analysis is then performed over the tags. Our project builds analysis tools that automate where feasible, and support scientists during remaining manual work.



Our **tagging tool** shows sociologists one archive page at a time. A simple configuration file automatically creates the tagging interface at the top. Its checkboxes allow the scientists to tag very rapidly. They may also randomly sample by checking the 'random' box. The tool automatically creates an Excel spreadsheet for subsequent statistical analysis.

Our **news visualization tool** has social scientists issue keyword queries. The tool shows how newspapers featured stories containing those keywords over time. Each column shows one day of publications. Each row corresponds to a news source.