Workshop on Information Integration
Philadelphia, PA
October 26-27, 2006

Submission by Jason R. Baron
Date: September 5, 2006

Position Paper:
Through A Lawyer's Lens: Measuring Performance in Conducting Large Scale Searches Against Heterogeneous Data Sets In Satisfaction of Litigation Requirements

Statement of the Problem

Increasingly, civil litigation in the United States involves requests on the part of one or both parties to a lawsuit for responsive documents from vast corporate or institutional repositories comprised of electronic data. Today's "state of the art" document request requires searches to be conducted against a spectrum of electronic applications residing on desktops, networks, and even backup media, for the purpose of obtaining every type of electronic object (and its corresponding metadata) created or received by the party responding to document production. Failure to meet minimal legal standards in performing adequate searches for electronic data has led to both headlines and multi-million dollar, even billion dollar, adverse judgments against underperforming parties in litigation. Yet even where lawyers are attempting to respond to discovery in good faith, it has become almost axiomatic that they cannot solely rely on manual means to cull through terabytes of data for "responsive" documents; thus, the legal community increasingly has found it necessary to turn to automated methods for making this Herculean task more manageable.

While lawyers have been using automated search tools for decades to perform "legal research" against static data sets of case law and other public materials, the new challenge faced by the profession is how best to design search processes and utilize search tools to find responsive information in largely unstructured, heterogeneous environments. Although there is a great deal of hype in the "legal tech" community regarding the efficacy of one or the other new forms of conceptual search techniques, there has been to date very little in the way of academic research aimed at evaluating the efficacy of potentially competing search methods in the context of real-world civil discovery, using known statistical measures such as recall and precision. See Baron (2005) (referencing Blair & Maron study); NIST/TREC Legal Track (2006); Collaborative Expedition Workshop #45 (2005). Correspondingly little in the way of case law exists in which a court has been called upon – to date -- to evaluate the efficacy of a particular chosen search methodology, apart from approving the use of designated "keywords" when searches are performed. See Sedona Conference (2005).

In particular, Government agencies are frequently drawn into lawsuits brought both on behalf of the United States (*e.g., United States v. Philip Morris*, the just-

decided tobacco lawsuit in which over 1 billion stored records were involved), and more frequently, as named party defendants in particular lawsuits (*e.g., Armstrong v. Executive Office of the President*, involving White House email). Although the litigation-driven need for optimizing searches against electronic record databases will be felt acutely by the National Archives and Records Administration over the next several years (especially when 100 million+ emails are expected to be received at the next Presidential transition), a wide spectrum of federal agencies will experience increasing litigation risk in failing to manage their growing electronic data repositories in all forms.

The workshop is especially timely in light of changes going into effect on December 1, 2006, to the way in which all lawyers in the United States litigating in the federal court system will need to approach the issue of how to preserve, access, and retrieve "electronically stored information" (ESI) -- a new term of art introduced into the federal rules – relevant to particular causes of action. These rules will require early, heightened interaction between opposing parties, prior to judicial intervention in a case, including on the issue of how each party's ESI will be searched to meet the demands of mutual ongoing or anticipated discovery.

In sum, further research into evaluating what constitutes efficacious means for searching large data collections in response to legal demands is a subject of enormous importance not only to the legal community, but also to the federal sector, especially to the extent that litigation-driven demands impact on an agency's budgeting of its overall resources.

References

Baron, Jason R., "Toward A Federal Benchmarking Standard for Evaluating Information Retrieval Products Used in E-Discovery, " 6 *Sedona Conference Journal* 237-246 (2005) (available on Westlaw, Lexis)

Blair, David C., and M.E. Maron, "An Evaluation of Retrieval Effectiveness For A Full-Text Document-Retrieval System," 1985 *ACM* Vol. 28, #3: 289-299.

Collaborative Expedition Workshop #45, *Advancing Information Sharing, Access, Discovery and Assimilation of Diverse Digital Collections Governed by Heterogeneous Sensitivies*, held Nov. 8, 2005, see http://colab.cim3.net/cgi-bin/wiki.pl?AdvancingInformationSharing_DiverseDigitalCollections_HeterogeneousSensitivities_11_08_05

NIST/TREC Legal Track 2006 home page, see http://trec-legal.umiacs.umd.edu

Sedona Conference, *The Sedona Principles: Best Practices Recommendations & Principles for Addressing Electronic Document Production* (2005 version), Principle 11 at p. 44, see http://www.thesedonaconference.org/content/miscFiles/publications_html.